ComSIS
Computer Science and Information Systems
Volume 21, Number 3, June 2024

Contents

Editorial
Guest Editorial: Adaptive Smart Areas and Intelligent Agents
Guest Editorial: Role of Agents in Traffic and Transportation
Guest Editorial: Innovations in Intelligent Systems and Applications

Papers

685 Low-Code Development Using Requirements and Knowledge Representation Models
Kamil Rybiński, Michał Śmiałek
725 A Practical and UC-Secure Decentralized Key Management and Authentication Scheme Based on Blockchain for VNDN
Xian Guo, Sarah Almadhehagi, Tao Feng, Di Zhang, Yongbo Jiang, Junli Fang
759 SSEPC Cloud: Carbon Footprint Aware Power Efficient Virtual Machine Placement in Cloud Milieu
Bivasa Ranjan Parida, Amiya Kumar Rath, Bibudhendu Pati, Chhabi Rani Panigrahi, Hitesh Mohapatra, Tien-HsiungWeng, Rajkumar Buyya
781 A Novel Approach to Source Code Assembling in the Field of Algorithmic Complexity
Đorde Pešić, Milena Vujošević Janičić, Marko Mišić, Jelica Protić
807 Medical Record Information Storage Scheme based on Blockchain and Attribute Role-Based Access Control
Aoao Bian, Dezhi Han, Mingming Cui, Dun Li
831 CTA-Net: A Gaze Estimation network based on Dual Feature Aggregation and Attention Cross Fusion
Chenxing Xia, Zhanpeng Tao, Wei Wang, Wenjun Zhao, Bin Ge, Xiuju Gao, Kuan-Ching Li, Yan Zhang
851 The Security and Privacy Challenges Toward Cybersecurity of 6G Networks: A Comprehensive Review
Yanlu Li, Yufeng Xiao, Wei Liang, Jiahong Cai, Ronglin Zhang, Kuan-Ching Li, Muhammad Khurram Khan
899 Reaching Quality and Efficiency with a Parameter-Efficient Controllable Sentence Simplification Approach
Antonio Menta, Ana Garcia-Serrano
923 Visit Planner: A Personalized Mobile Trip Design Application based on a Hybrid Recommendation Model
Harris Papadakis, Costas Panagiotakis, Paraskevi Fragopoulou, Georgios Chalkiadakis, Errikos Streviniotis, Ioannis-Panagiotis Ziogas, Michail Koutsmanis, Panagiotis Bariamis
947 ECW-EGNet: Exploring Cross-ModalWeighting and Edge-Guided Decoder Network for RGB-D Salient Object Detection
Chenxing Xia, Feng Yang, Songsong Duan, Xiuju Gao, Bin Ge, Kuan-Ching Li, Xianjin Fang, Yan Zhang, Ke Yang
971 MakeCode Arcade Platform for Game-Based Learning
Patrik Voštinár
989 News Recommendation Model Based on Encoder Graph Neural Network and Bat Optimization in Online Social Multimedia Art Education
Jing Yu, Lu Zhao, Shoulin Yin, Mirjana Ivanović
1013 Agile Culture Clash: Unveiling Challenges in Cultivating an Agile Mindset in Organizations
Michael Neumann, Thorben Kuchel, Philipp Diebold, Eva-Maria Schön

Special section: Papers selected from European Conference on Advances in Databases and Information Systems

1033 Graph Rewriting Primitives for Semantic Graph Databases Sanitization
Adrien Boiret, Cédric Eichler, Benjamin Nguyen, Sara Taki
1055 A Comprehensive Review of the Data and Knowledge Graphs Approaches in Bioinformatics
Ylenia Galluzzo
1077 Operational Semantics in a Domain-Specific Robot Control Language: a Pedagogical Use Case
William Steingartner, Valerie Novitzká

Special section: Computer Science and Information Systems Applied to Management, Innovation and Sustainability

1097 Natural Language Processing: An Overview of Models, Transformers and Applied Practices
Santiago Canchila, Carlos Meneses-Eraso, Javier Casanoves-Boix, Pascual Cortés-Pellicer, Fernando Castelló-Sirvent
1147 MCClusteringSM: An approach for the Multicriteria Clustering Problem based on a Credibility Similarity Measure
Lugo Medrano Cesar, Gastelum Chavira Diego Alonso, Valdez Lafarga Octavio, Velarde Cervantes Jose Luis
1179 Analysis of the Countries' business attraction with the ELECTRE-III method
Tanya Samantha Garcia-Gastelum, Pavel Anselmo Álvarez, Ernesto León-Castro, Cristhian Ramón Uzeta-Obregon
1203 The incidence of some critical cognitive factors that prevent change of order in university students: uncertainty analysis
José M. Brotons-Martínez, Rubén Chavez Rivera, Jesús Ricardo Ramos-Sánchez

ISSN: 2406-1018 (Online)

Computer Science and Information Systems

Vol 21, No 3, June 2024

ComSIS

Computer Science and Information Systems

Published by ComSIS Consortium

Volume 21, Number 3
June 2024

ComSIS is an international journal published by the ComSIS Consortium

ComSIS Consortium:
University of Belgrade:
Faculty of Organizational Science, Belgrade, Serbia
Faculty of Mathematics, Belgrade, Serbia
School of Electrical Engineering, Belgrade, Serbia
Serbian Academy of Science and Art:
Mathematical Institute, Belgrade, Serbia
Union University:
School of Computing, Belgrade, Serbia
University of Novi Sad:
Faculty of Sciences, Novi Sad, Serbia
Faculty of Technical Sciences, Novi Sad, Serbia
Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
University of Niš:
Faculty of Electronic Engineering, Niš, Serbia
University of Montenegro:
Faculty of Economics, Podgorica, Montenegro

EDITORIAL BOARD:
Editor-in-Chief: Mirjana Ivanović, University of Novi Sad
Vice Editor-in-Chief: Boris Delibašić, University of Belgrade
Managing Editors:
Vladimir Kurbalija, University of Novi Sad
Miloš Radovanović, University of Novi Sad

Editorial Assistants:
Jovana Vidaković, University of Novi Sad
Ivan Pribela, University of Novi Sad
Davorka Radaković, University of Novi Sad
Slavica Kordić, University of Novi Sad
Srđan Škrbić, University of Novi Sad

Editorial Board:
A. Badica, *University of Craiova, Romania*
C. Badica, *University of Craiova, Romania*
M. Bajec, *University of Ljubljana, Slovenia*
L. Bellatreche, *ISAE-ENSMA, France*
I. Berković, *University of Novi Sad, Serbia*
D. Bojić, *University of Belgrade, Serbia*
Z. Bosnic, *University of Ljubljana, Slovenia*
D. Brđanin, *University of Banja Luka, Bosnia and Hercegovina*
R. Chbeir, *University Pau and Pays Adour, France*
M-Y. Chen, *National Cheng Kung University, Tainan, Taiwan*
C. Chesñevar, *Universidad Nacional del Sur, Bahía Blanca, Argentina*
W. Dai, *Fudan University Shanghai, China*
P. Delias, *International Hellenic University, Kavala University, Greece*
B. Delibašić, *University of Belgrade, Serbia*
G. Devedžić, *University of Kragujevac, Serbia*
J. Eder, *Alpen-Adria-Universität Klagenfurt, Austria*
Y. Fan, *Communication University of China*
V. Filipović, *University of Belgrade, Serbia*
T. Galinac Grbac, *Juraj Dobrila University of Pula, Croatia*
H. Gao, *Shanghai University, China*
M. Gušev, *Ss. Cyril and Methodius University Skopje, North Macedonia'*
D. Han, *Shanghai Maritime University, China*
M. Heričko, *University of Maribor, Slovenia*
M. Holbl, University of Maribor, Slovenia
L. Jain, *University of Canberra, Australia*
D. Janković, *University of Niš, Serbia*
J. Janousek, *Czech Technical University, Czech Republic*
G. Jezic, *University of Zagreb, Croatia*
G. Kardas, *Ege University International Computer Institute, Izmir, Turkey*
Lj. Kašćelan, *University of Montenegro, Montenegro*
P. Kefalas, *City College, Thessaloniki, Greece*
M-K. Khan, *King Saud University, Saudi Arabia*
S-W. Kim, *Hanyang University , Seoul, Korea*
M. Kirikova, *Riga Technical University, Latvia*
A. Klašnja Milićević, *University of Novi Sad, Serbia*
J. Kratica, *Institute of Mathematics SANU, Serbia*
K-C. Li, *Providence University, Taiwan*
M. Lujak, *University Rey Juan Carlos, Madrid, Spain*
JM. Machado, *School of Engineering, University of Minho, Portugal*
Z. Maamar, *Zayed University, UAE*
Y. Manolopoulos, *Aristotle University of Thessaloniki, Greece*
M. Mernik, *University of Maribor, Slovenia*
B. Milašinović, *University of Zagreb, Croatia*
A. Mishev, *Ss. Cyril and Methodius University Skopje, North Macedonia*
N. Mitić, *University of Belgrade, Serbia*
N-T. Nguyen, *Wroclaw University of Science and Technology, Poland*
P Novais, *University of Minho, Portugal*
B. Novikov, *St Petersburg University, Russia*
M. Paprzicky, *Polish Academy of Sciences, Poland*
P. Peris-Lopez, *University Carlos III of Madrid, Spain*
J. Protić, *University of Belgrade, Serbia*
M. Racković, *University of Novi Sad, Serbia*
M. Radovanović, *University of Novi Sad, Serbia*
P. Rajković, *University of Nis, Serbia*
O. Romero, *Universitat Politècnica de Catalunya, Barcelona, Spain*
C, Savaglio, *ICAR-CNR, Italy*
H. Shen, *Sun Yat-sen University, China*
J. Sierra, *Universidad Complutense de Madrid, Spain*
B. Stantic, *Griffith University, Australia*
H. Tian, *Griffith University, Australia*
N. Tomašev, *Google, London*
G. Trajčevski, *Northwestern University, Illinois, USA*
G. Velinov, *Ss. Cyril and Methodius University Skopje, North Macedonia*
L. Wang, *Nanyang Technological University, Singapore*
F. Xia, *Dalian University of Technology, China*
S. Xinogalos, *University of Macedonia, Thessaloniki, Greece*
S. Yin, *Software College, Shenyang Normal University, China*
K. Zdravkova, *Ss. Cyril and Methodius University Skopje, North Macedonia*
J. Zdravković, *Stockholm University, Sweden*

ComSIS Editorial Office:
University of Novi Sad, Faculty of Sciences,
Department of Mathematics and Informatics
Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia
Phone: +381 21 458 888; **Fax:** +381 21 6350 458
www.comsis.org; Email: comsis@uns.ac.rs

Volume 21, Number 3, 2024
Novi Sad

Computer Science and Information Systems

ISSN: 2406-1018 (Online)

The ComSIS journal is sponsored by:

Ministry of Education, Science and Technological Development of the Republic of Serbia
http://www.mpn.gov.rs/



Computer Science and Information Systems

AIMS AND SCOPE

Computer Science and Information Systems (ComSIS) is an international refereed journal, published in Serbia. The objective of ComSIS is to communicate important research and development results in the areas of computer science, software engineering, and information systems.

We publish original papers of lasting value covering both theoretical foundations of computer science and commercial, industrial, or educational aspects that provide new insights into design and implementation of software and information systems. In addition to wide-scope regular issues, ComSIS also includes special issues covering specific topics in all areas of computer science and information systems.

ComSIS publishes invited and regular papers in English. Papers that pass a strict reviewing procedure are accepted for publishing. ComSIS is published semiannually.

Indexing Information

ComSIS is covered or selected for coverage in the following:

- Science Citation Index (also known as SciSearch) and Journal Citation Reports / Science Edition by Thomson Reuters, with 2023 two-year impact factor 1.2,
- Computer Science Bibliography, University of Trier (DBLP),
- EMBASE (Elsevier),
- Scopus (Elsevier),
- Summon (Serials Solutions),
- EBSCO bibliographic databases,
- IET bibliographic database Inspec,
- FIZ Karlsruhe bibliographic database io-port,
- Index of Information Systems Journals (Deakin University, Australia),
- Directory of Open Access Journals (DOAJ),
- Google Scholar,
- Journal Bibliometric Report of the Center for Evaluation in Education and Science (CEON/CEES) in cooperation with the National Library of Serbia, for the Serbian Ministry of Education and Science,
- Serbian Citation Index (SCIndeks),
- doiSerbia.

Information for Contributors

The Editors will be pleased to receive contributions from all parts of the world. An electronic version (LaTeX), or three hard-copies of the manuscript written in English, intended for publication and prepared as described in "Manuscript Requirements" (which may be downloaded from http://www.comsis.org), along with a cover letter containing the corresponding author's details should be sent to official journal e-mail.

Criteria for Acceptance

Criteria for acceptance will be appropriateness to the field of Journal, as described in the Aims and Scope, taking into account the merit of the content and presentation. The number of pages of submitted articles is limited to 20 (using the appropriate LaTeX template).

Manuscripts will be refereed in the manner customary with scientific journals before being accepted for publication.

Copyright and Use Agreement

All authors are requested to sign the "Transfer of Copyright" agreement before the paper may be published. The copyright transfer covers the exclusive rights to reproduce and distribute the paper, including reprints, photographic reproductions, microform, electronic form, or any other reproductions of similar nature and translations. Authors are responsible for obtaining from the copyright holder permission to reproduce the paper or any part of it, for which copyright exists.

Computer Science and Information Systems

Volume 21, Number 3, June 2024

CONTENTS

Editorial

Guest Editorial: Papers selected from European Conference on Advances in Databases and Information Systems

Guest Editorial: Computer Science and Information Systems Applied to Management, Innovation and Sustainability

Papers

685 **Low-Code Development Using Requirements and Knowledge Representation Models**
Kamil Rybiński, Michał Śmiałek

725 **A Practical and UC-Secure Decentralized Key Management and Authentication Scheme Based on Blockchain for VNDN**
Xian Guo, Sarah Almadhehagi, Tao Feng, Di Zhang, Yongbo Jiang, Junli Fang

759 **SSEPC Cloud: Carbon Footprint Aware Power Efficient Virtual Machine Placement in Cloud Milieu**
Bivasa Ranjan Parida, Amiya Kumar Rath, Bibudhendu Pati, Chhabi Rani Panigrahi, Hitesh Mohapatra, Tien-HsiungWeng, Rajkumar Buyya

781 **A Novel Approach to Source Code Assembling in the Field of Algorithmic Complexity**
Đorđe Pešić, Milena Vujošević Janičić, Marko Mišić, Jelica Protić

807 **Medical Record Information Storage Scheme based on Blockchain and Attribute Role-Based Access Control**
Aoao Bian, Dezhi Han, Mingming Cui, Dun Li

831 **CTA-Net: A Gaze Estimation network based on Dual Feature Aggregation and Attention Cross Fusion**
Chenxing Xia, Zhanpeng Tao, Wei Wang, Wenjun Zhao, Bin Ge, Xiuju Gao, Kuan-Ching Li, Yan Zhang

851 **The Security and Privacy Challenges Toward Cybersecurity of 6G Networks: A Comprehensive Review**
Yanlu Li, Yufeng Xiao, Wei Liang, Jiahong Cai, Ronglin Zhang, Kuan-Ching Li, Muhammad Khurram Khan

899 **Reaching Quality and Efficiency with a Parameter-Efficient Controllable Sentence Simplification Approach**
Antonio Menta, Ana Garcia-Serrano

923 **Visit Planner: A Personalized Mobile Trip Design Application based on a Hybrid Recommendation Model**
Harris Papadakis, Costas Panagiotakis, Paraskevi Fragopoulou, Georgios Chalkiadakis, Errikos Streviniotis, Ioannis-Panagiotis Ziogas, Michail Koutsmanis, Panagiotis Bariamis

947 **ECW-EGNet: Exploring Cross-ModalWeighting and Edge-Guided Decoder Network for RGB-D Salient Object Detection**
Chenxing Xia, Feng Yang, Songsong Duan, Xiuju Gao, Bin Ge, Kuan-Ching Li, Xianjin Fang, Yan Zhang, Ke Yang

971 **MakeCode Arcade Platform for Game-Based Learning**
Patrik Voštinár

989 **News Recommendation Model Based on Encoder Graph Neural Network and Bat Optimization in Online Social Multimedia Art Education**
Jing Yu, Lu Zhao, Shoulin Yin, Mirjana Ivanović

1013 **Agile Culture Clash: Unveiling Challenges in Cultivating an Agile Mindset in Organizations**
Michael Neumann, Thorben Kuchel, Philipp Diebold, Eva-Maria Schön

Special section: Papers selected from European Conference on Advances in Databases and Information Systems

1033 **Graph Rewriting Primitives for Semantic Graph Databases Sanitization**
Adrien Boiret, Cédric Eichler, Benjamin Nguyen, Sara Taki

1055 **A Comprehensive Review of the Data and Knowledge Graphs Approaches in Bioinformatics**
Ylenia Galluzzo

1077 **Operational Semantics in a Domain-Specific Robot Control Language: a Pedagogical Use Case**
William Steingartner, Valerie Novitzká

Special section: Computer Science and Information Systems Applied to Management, Innovation and Sustainability

1097 **Natural Language Processing: An Overview of Models, Transformers and Applied Practices**
Santiago Canchila, Carlos Meneses-Eraso, Javier Casanoves-Boix, Pascual Cortés-Pellicer, Fernando Castelló-Sirvent

1147 **MCClusteringSM: An approach for the Multicriteria Clustering Problem based on a Credibility Similarity Measure**
Lugo Medrano Cesar, Gastelum Chavira Diego Alonso, Valdez Lafarga Octavio, Velarde Cervantes Jose Luis

1179 Analysis of the Countries' business attraction with the ELECTRE-III method
Tanya Samantha Garcia-Gastelum, Pavel Anselmo Álvarez, Ernesto León-Castro, Cristhian Ramón Uzeta-Obregon

1203 The incidence of some critical cognitive factors that prevent change of order in university students: uncertainty analysis
José M. Brotons-Martínez, Rubén Chavez Rivera, Jesús Ricardo Ramos-Sánchez

Computer Science and Information Systems 21(3): i–iii https://doi.org/10.2298/CSIS240300iI

Editorial

Mirjana Ivanović, Miloš Radovanović, and Vladimir Kurbalija

University of Novi Sad, Faculty of Sciences
Novi Sad, Serbia
{mira,radacha,kurba}@dmi.uns.ac.rs

The third issue of Volume 21 of Computer Science and Information Systems features 13 regular articles and two special sections: "Papers selected from European Conference on Advances in Databases and Information Systems (ADBIS)" (3 articles) and "Computer Science and Information Systems Applied to Management, Innovation and Sustainability" (3 articles). As customary, we are thankful for the hard work and enthusiasm of our authors, reviewers, and guest editors, without whom the current issue and the publication of the journal itself would not be possible.

We are happy to announce the updated impact factors of our journal for 2023: the new two-year IF 1.2, and the five-year IF 1.1.

The first regular article, "Low-Code Development Using Requirements and Knowledge Representation Models" by Kamil Rybiński and Michał Śmiałek, presents a visual language RSL-DL that can be used to represent non-standard domain logic (i.e., data processing) operations in a visual form, to complement existing visual code tools for visually describing the user interface and standard application logic. RSL-DL allows to capture domain knowledge with complex domain rules aligned with requirements models, while the RSL-DL environment includes an inference engine that enables processing queries to domain models and selecting appropriate invocations to generated code.

In the second regular article, "A Practical and UC-Secure Decentralized Key Management and Authentication Scheme Based on Blockchain for VNDN," Xian Guo et al. introduce an efficient decentralized key management solution based on blockchain for vehicular named data networking (VNDN), which tackles the vulnerability of having a centralized certificate authority by completely eliminating it. Also, based on this key management scheme, a lightweight mutual authentication scheme and a key agreement protocol for vehicle-to-everything (V2X) are proposed.

"SSEPC Cloud: Carbon Footprint Aware Power Efficient Virtual Machine Placement in Cloud Milieu," by Bivasa Ranjan Parida et al. propose a hybrid metaheuristic technique combining the salp swarm optimization and emperor penguins colony algorithm (SSEPC) to place virtual machines in the most suitable regions, availability zones, datacenters, and servers in a cloud environment, while optimizing consumption of energy and carbon emissions, as well as average response time and service level agreement (SLA) violations.

Đorđe Pešić et al., in "A Novel Approach to Source Code Assembling in the Field of Algorithmic Complexity," propose a novel template- and rule-based approach and a corresponding software system for assembling synthetic source code segments of defined time complexity for the purposes of teaching computational complexity analysis. Based on the developed grammar, the system can produce source code segments with a broad scope of different time complexities while guaranteeing the complexity of the generated segment.

The article "Medical Record Information Storage Scheme based on Blockchain and Attribute Role-Based Access Control," authored by Aoao Bian et al., proposes a solution for secure storage and sharing of medical record information based on blockchain, combining role-based access control (RBAC) with attribute-based access control (ABAC). Advanced encryption standard and feature-aware stateful routing technologies are applied to further enhance the security and storage efficiency of the scheme.

In their article entitled "CTA-Net: A Gaze Estimation Network Based on Dual Feature Aggregation and Attention Cross Fusion," Chenxing Xia et al. propose a parallel CNNs-Transformer aggregation network (CTA-Net) for gaze estimation, which fully leverages the advantages of the Transformer model in modeling global context while the convolutional neural networks (CNNs) model is beneficial for retaining local details.

"The Security and Privacy Challenges Toward Cybersecurity of 6G Networks: A Comprehensive Review," by Yanlu Li et al. analyzes the security issues of 6G networks and presents some possible solutions. After discussing recent developments in mobile communication technology (research and performance motivation behind 6G networks, as well as key technologies), security threats in 6G networks are analyzed concerning architecture, major visions and related applications, followed by solutions to security issues in applying key technologies for 6G networks.

Antonio Menta and Ana Garcia-Serrano, in their article "Reaching Quality and Efficiency with a Parameter-Efficient Controllable Sentence Simplification Approach," tackle the resource hungriness of current approaches to automatic text simplification (ATS) based on large language models (LLMs) by proposing an accurate solution when powerful resources are not available, using the transfer learning capabilities across different domains with a set of linguistic features employing a reduced size pre-trained language model.

In "Visit Planner: A Personalized Mobile Trip Design Application based on a Hybrid Recommendation Model," Harris Papadakis et al. present Visit Planner (ViP), a mobile application prototype that provides a solution to the challenging tourist trip design problem. ViP follows a holistic approach offering personalized recommendations for Points of Interest (POIs) based on preferences either explicitly collected by the application, or inferred by the users' ongoing interaction with the system.

In their article "ECW-EGNet: Exploring Cross-Modal Weighting and Edge-Guided Decoder Network for RGB-D Salient Object Detection," Chenxing Xia et al. propose an Exploring Cross-Modal Weighting and Edge-Guided Decoder Network (ECW-EGNet) for RGB-D salient object detection (SOD), which includes three prominent components: a Cross-Modality Weighting Fusion (CMWF) module, a Bi-directional Scale-Correlation Convolution (BSCC) module, and an Edge-Guided (EG) decoder.

"MakeCode Arcade Platform for Game-Based Learning," by Patrik Voštinár presents research into whether the MakeCode Arcade environment is suitable for teaching programming through games as a game-based learning platform. Conclusions from a survey taken by 315 students are drawn, and experiences in using the tool in different settings are described: teaching in extracurricular activity at university level, teaching normal lessons in primary schools, use as a motivational tool for distance workshops, and also in workshops for teachers.

The article "News Recommendation Model Based on Encoder Graph Neural Network and Bat Optimization in Online Social Multimedia Art Education," by Jing Yu et al. pro-

pose a novel news recommendation model based on encoder graph neural network (GNN) and Bat optimization in online social networks, where Bat optimization used to improve the effect of news clustering, the concept of metadata is introduced into the GNN, and the ontology of learning resources based on knowledge points is established to realize the correlation between news resources.

Finally, “Agile Culture Clash: Unveiling Challenges in Cultivating an Agile Mindset in Organizations,” by Michael Neumann et al. identifies challenges that arise from the interplay between agile culture and organizational culture, using a mixed-method research approach: gathering qualitative data among from a network of agile practitioners, and obtaining quantitative data by means of a questionnaire study with 92 participants. Seven key challenges were identified, referring to technical agility (doing agile) and cultural agility (being agile).

Computer Science and Information Systems 21(3): v–v https://doi.org/10.2298/CSIS240300vV

Guest Editorial: Papers selected from European Conference on Advances in Databases and Information Systems

Genoveva Vargas-Solar, Barbara Catania, Ester Zumpano, Kjetil Nørvåg, Silvia Chiusano, Tania Cerquitelli, and Robert Wrembel

This special guest section publishes three extended versions of selected papers accepted in workshops held in conjunction with ADBIS 2022: European Conference on Advances in Databases and Information Systems (http://adbis.eu/). Workshops covered diverse topics, such as intelligent data processing, large-scale knowledge graph analysis, modern data engineering, advanced data systems management, and semantic web design for cultural heritage.

The selected papers were chosen considering their technical quality assessed by the international program committee that evaluated workshop papers. They went through a strict evaluation of four reviewers, specialists in the topics of the papers. The following papers were included into this special guest section of ComSIS.

1. "Operators for Semantic Graph Databases as Graph Rewriting", by Adrien Boiret, Cédric Eichler, and Benjamin Nguyen: This paper introduces basic graph rewriting operators for constructing sanitization mechanisms. The contribution is a formal framework for sanitizing semantic graph databases, enhancing privacy protection while facilitating data sharing and publication.

2. "A Review: Biological Insights on Knowledge Graphs", by Ylenia Galluzzo: This paper surveys knowledge graph applications in biology and clinical contexts, focusing on recent progress in biomedical knowledge representation. The paper suggests future research avenues to enhance precision medicine in biomedicine.

3. "Abstract machine for operational semantics of domain-specific language", by William Steingartner, Róbert Baraník, and Valerie Novitzká: This paper delves into structural operational semantics for a domain-specific language focusing on robot control. It presents a method for abstract implementation on an abstract machine, contributing to semantic methods research in software engineering. An application is introduced to visualize program execution steps, aiding formal semantics teaching for domain-specific languages.

The Guest Editors extend their sincere gratitude to everyone who contributed to the success of this special issue. We thank the authors for sharing their scientific insights with the community. Authors invested effort in producing the versions that are published in ComSIS, considering reviewers' suggestions and requests for consolidating the work presented in the papers.

We are also profoundly thankful to Prof. Mirjana Ivanović, the editor-in-chief of Computer Science and Information Systems (ComSIS), for her invaluable guidance and support throughout the process. We are also grateful with Prof. Sonja Ristić of the ComSIS consortium for her help in the evaluation process of the selected papers.

A special acknowledgment is due to the reviewers for their constructive feedback significantly enhanced the quality of the papers accepted for this special issue.

Computer Science and Information Systems 21(3): vii–viii https://doi.org/10.2298/CSIS240300viiL

Guest Editorial: Computer Science and Information Systems Applied to Management, Innovation and Sustainability

Ernesto Leon-Castro[1], Boris Delibašić[2], and Ivan Luković[3]

[1] Universidad Católica de la Santísima Concepción, Chile
[2] University of Belgrade – Faculty of Organizational Sciences, Serbia
[3] University of Belgrade – Faculty of Organizational Sciences, Serbia

The special section Computer Science and Information Systems Applied to Management, Innovation and Sustainability is a selection of the best papers presented at the International Conference on Innovation and Sustainability (ICONIS) 2022. The main idea of this special section is to show different approaches to information systems in different areas and applications.

An information system (IS) collects, processes, stores, analyses, and disseminates information for a specific purpose and has become very relevant in all organizations; they are used in the operation and management of nearly all actions (Adeoti-Adekeye, 1997; Rainer et al., 2020), because of that, multiple methodologies have been developed. The main purpose of this IS is to support the business goal and organizational capabilities and help to make better decisions (Pearlson et al., 2024).

The special section consists of four papers.

The first one, "Natural Language Processing (NLP): A survey on models and applied practices" (authors: Santiago Canchila Corredor, Fernando Castello-Sirvent, Carlos Meneses Eraso, Javier Casanoves Boix), presents how NLP has evolved constantly because of its impact on day-to-day life, commercial life, and entrepreneurial endeavors. The papers aim to present an overview of the different NLP models. Also, guidance on what the future of NLP may look like is given.

The second paper, "MCClusteringSM: An approach for the Multicriteria Clustering Problem based on a Credibility Similarity Measure" (authors: Lugo Medrano Cesar, Gastelum Chavira Diego Alonso, Valdez Lafarga Octavio, Velarde Cervantes Jose Luis), presents a new algorithm to improve the multicriteria clustering problem by exploiting a fuzzy outranking relation through a direct algorithm. The proposed methodology was used in a credibility similarity measure to evaluate information, and different comparisons with other methods were made.

The third paper is "Analysis of the Countries' business attraction with a fuzzy outranking method" (authors: Tanya Samantha Garcia-Gastelum, Pavel Anselmo Alvarez, Ernesto Leon-Castro, and Cristhian Ramon Uzeta-Obregon); since attracting investment is one of the countries' goals to improve their competitiveness, investors must consider different elements before deciding. Based on that, this paper presents the ELECTRE-III methodology to rank the 190 countries considered in the Doing Business 2020 survey. A ranking indicating the best countries to invest in is presented among the main results.

Finally, the fourth article is about the incidence of some critical cognitive factors that prevent change of order in university students: uncertainty analysis (authors: Jose M. Brotons-Martınez, Ruben Chavez Rivera, and Jesus Ricardo Ramos-Sanchez). In this

paper, an analysis of the importance of emotional intelligence in the academic performance of undergraduate students, the paper aims to identify the critical cognitive factors that prevent change of order and process them through fuzzy cognitive maps. The Bar On test was applied to 160 regular students and 63 lagging students in three universities, two from Mexico and one from Spain. Among the main results, the authors found that ten items present the greatest distance between the groups and need improvement to improve student performance.

With these four articles, readers will get an idea of how different IS can help carry out organizational tasks, what new trends are, and how to analyze information to find new realities within organizational problems. Consider scenarios that, at first glance, are not considered and help make decisions that are appropriate to the decisions.

Computer Science and Information Systems 21(3):685–724 https://doi.org/10.2298/CSIS230102024R

Low-Code Development Using Requirements and Knowledge Representation Models*

Kamil Rybiński[1] and Michał Śmiałek[1]

Warsaw University of Technology
pl. Politechniki 1, Warsaw, Poland
{kamil.rybinski, michal.smialek}@pw.edu.pl

Abstract. Low-Code Development is a new software development paradigm which is typically used to generate web applications from high-level visual notations. These notations allow to express the User Interface and Application Logic (user-system interactions) in a way which is understandable by the end-users. Using certain model-driven approaches, the low-code environments allow for generating the front-end layer and basic CRUD operations in the back-end. Yet still, non-standard domain logic (data processing) operations still necessitate the use of traditional programming. In this article we present a visual language, called RSL-DL that can be used to represent such non-standard domain logic in a visual form. It allows to capture domain knowledge with complex domain rules aligned with requirements models. The language synthesises and extends approaches found in knowledge representation (ontologies) and software modelling language engineering. The development environment of RSL-DL enables fully automatic generation of domain logic code by reasoning over and reusing domain knowledge. The environment includes a dedicated model editor and a transformation engine. The language's abstract syntax is defined using a meta-model expressed in MOF. Its semantics is expressed with several translational rules that map RSL-DL models onto typical programming language constructs. The article presents a list of these rules in an informal way, and then introduces their formalisation using a graphical transformation notation. The RSL-DL environment includes an inference engine that enables processing queries to domain models and selecting appropriate invocations to generated code. It was also initially validated through studies that involve understandability, operability and complexity assessment. Based on these results, we conclude that declarative knowledge representations can be successfully used to produce imperative back-end code with non-trivial logic.

Keywords: low-code, model-driven web engineering, knowledge representation, ontologies, code generation.

1. Introduction

The term "Low-Code Software Development" (LCSD) has emerged as a new approach to application development where only a limited amount of coding is required. Since its emergence around nine years ago [47], the term was used in the industry to label cloud-based development platforms that use visual notations to reduce the need for traditional

* Extended version of the paper "Beyond Low-Code Development: Marrying Requirements Models and Knowledge Representations" published at the FedCSIS'22 conference

programming. Research on low-code approaches is currently yet sparse. However, just recently, it has been observed that LCSD can be seen as a subdomain [12] or as overlapping [50] with the Model-Driven Software Development (MDSD, MDD), where it concentrates on automatic generation of data-rich web/mobile applications from visual specifications (models).

A typical LCSD platform offers a visual language that enables the definition of user interface elements and application logic (navigation through the UI, handling user interactions). The language typically includes a visual notation to define the data model of the problem domain under consideration. These elements allow to generate and deploy automatically, much of a web/mobile application code. However, some non-standard data processing necessitates manual coding of appropriate APIs and services. In this article, we present an attempt to go beyond this and further reduce (or even eliminate) the need for traditional coding in software development.

The ultimate target of our research is a true general-purpose no-code development platform [67]. By this, we mean an environment that extends the low-code approach with ways to specify (business) domain logic and generate even complex data processing services. These services would integrate with the code generated from the low-code parts and thus enable creating a software system without programming in a traditional sense. What is essential, such a no-code approach would not be limited to a specific problem domain area.

When searching for candidate technology paths to fulfil the above vision, we can make certain observations. Specifically, low-code approaches use high-level, domain-agnostic abstractions familiar to domain experts (or software users). We can note that such abstractions can serve the same function as detailed software requirements in a typical software development project. Thus, we can expect to use specific requirements artefacts as the actual low-code source specifications. This way, the abstractions used in LCSD would be compatible with the abstractions used in traditional software processes.

So far, no research explicitly addresses the use of requirements artefacts in low-code approaches. However, we can refer to relevant research on Model-Driven Web Engineering (MDWE) that can be seen as a predecessor and a synonym for LCSD. A survey by Valderas and Pelechano [62] shows that most MDWE methods use some form of "classical" use case models [23] to specify functional requirements for web/mobile applications. They are usually supported by constrained language scenarios and/or activity diagrams to define detailed navigation through the user interface. Moreover, several approaches use some form of a data dictionary to define the problem domain. Based on this, certain automated transformations to other artefacts were proposed [28]. Nevertheless, most model-driven approaches cannot generate code directly from requirements.

One of the sparse systems that offer such capabilities is the ReDSeeDS[1] platform [57] which can be seen as a representative of requirements-based MDD. The system was created before the emergence of the low-code movement, but it certainly fulfils the definition of LCSD. It uses precisely specified requirements models: use cases, scenarios in constrained language, and visual domain vocabularies. Similarly to low-code platforms, these artefacts are represented with a visual language called RSL (Requirements Specification Language) [25, 39]. Specifications expressed in RSL allow for the generation of fully functional UI and application logic code for data-rich web applications. However,

[1] https://github.com/smialekm/redseeds

based on RSL alone, one cannot generate data processing (domain logic) code beyond simple CRUD, and data persistence operations [68]. Thus, similarly to other low-code platforms, more complex data processing has to be coded manually.

Considering the above, we raise the question of representing more complex domain logic at the level of abstraction used by low-code and requirements-based MDD approaches. Inspiration for responding to this question can be drawn from research on ontologies and knowledge representations, as they seem to share common goals. These approaches enable the representation of domain knowledge, independently of any technology and any particular problem domain. Ontologies offer means to create a shared, reusable vocabulary composed of concepts, rules and relationships between these concepts [16]. They constitute a common ground for communicating knowledge by domain experts and processing this knowledge by machines. For this purpose, they represent knowledge in a declarative way and are well adapted to conduct various reasoning tasks.

In this research, we investigate how the features of ontology-based domain logic representations can be applied in the context of LCSD. We present an extension to the RSL mentioned above, which we call RSL-DL (RSL Domain Logic). The new language draws several of its constructs from ontology-based knowledge representation approaches [53]. It then extends and combines them with MDD technologies to provide full code generation capabilities. What is important, RSL-DL allows for the generation of fully operational code directly from general domain rules (descriptions of reality) as required by specific requirements models (use cases and their logic). Hence, such generated back-end code is fully compatible with the front-end code generated from RSL specifications. Through this, we demonstrate an entirely visual extension to a low-code language (RSL) that has the potential of eliminating the need to code (in a traditional sense) complex data processing services.

The results presented in this article extend the results presented at the FedCSIS conference [51]. The extension includes a significant expansion of the presentation of the research context in the introduction, related work and RSL sections. We have also significantly extended the language examples and included the full list of translational rules with additional examples. Finally, we have added an additional validation approach which assesses complexity of RSL-DL code versus traditional code.

The structure of this article consists of nine sections. Section 2 describes the research context of our work, including overview of research on the low-code paradigm, visual languages and ontologies. Section 3 gives an overview of RSL which was the starting point for our research. Sections 4 and 5 define RSL-DL – its syntax and semantics. Section 6 includes the most important details of the language's transformation engine implementation. Sections 7 and 8 provide evaluation of the language through two case studies and three validation analyses. Section 9 concludes with a summary and discussion on future work. In addition, we also provide a Supplement (to be found in our github page[2]) that contains all the details that were omitted in the article for brevity.

2. Motivation and related work

The term "low-code" was probably first used in a Forrester report [47] only in 2014. Despite this relatively recent emergence, predictions for LCSD are promising – in 2019,

[2] https://github.com/smialekm/redseeds/tree/main/RSL-DL

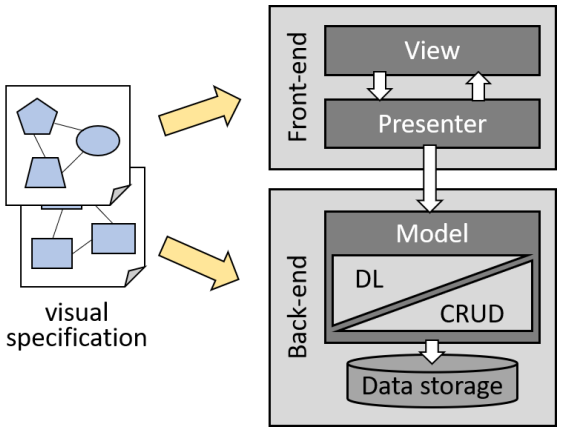


Fig. 1. Application generation in LCSD/MDWE

Gartner predicted that 65% of software in 2024 will be built using some form of low-code environment [63]. Still, there is not much data on the effectiveness of low-code approaches compared to more traditional ones. However, recent research by Bexiga et al. [7] and Trigo et al. [61] present very promising conclusions based on industrial experience. It is argued that this can be caused by a much higher involvement of design experts (e.g. UI/UX designers without programming skills) in software development. Yet, this does not eliminate the need for professional software developers' involvement and is not limited to simple applications produced by "citizen developers" [48]. Current studies report numerous industry-grade low-code platforms used extensively by professionals [52].

Sparse research results on LCSD can be supported by previous and current research on MDWE [29, 30] which can be seen as a predecessor and now a synonym for LCSD. A study by Wakil and Jawani [66] shows that research on MDWE is already quite broad and mature. Several studies show very significant productivity gains [33, 43, 14] and better maintainability [34]. It includes various controlled experiments and comparisons of complexity between high-level visual models and equivalent code. LCSD and MDWE approaches are typically based on some form of visual notation (language). Such notations offer high-level representations of the flows of interaction between application users and the system under development. A prominent example in the MDWE domain is the Interaction Flow Modeling Language (IFML) [8]. Other examples – in the LCSD domain – are the Business Process Technology (BPT) language [21] and the Mendix notation [20].

The information scope of these LCSD/MDWE languages allows the associated development environments to generate significant portions of the target system's logic and the user interface. Figure 1 shows a typical approach to code generation found in such systems. It presents one of the popular architectural patterns, namely the Model-View-Presenter (MVP) [45]. The View (presentation) and Presenter (application logic) layers form the front-end of a system and can usually be fully generated from an LCSD/MDWE visual specification. The role of a code generation engine is to "inject" technology-specific elements in addition to the interaction flow and UI layouts already contained in the source visual specifications.

Still, LCSD/MDWE systems have limited capabilities regarding the generation of the domain/business logic code, or more broadly – the system's back-end (see again Figure 1). Current LCSD/MDWE languages can support generation of code for elementary CRUD (Create-Read-Update-Delete) operations [5, 49]. Generation of code for more complex data processing (general Domain Logic - DL) is limited by the information scope of the visual language constructs. In this research, we propose new constructs that significantly extend capabilities to generate complex domain logic code. What is important, these new constructs are domain-agnostic, as contrasted with various domain-specific and often very formalised notations (see, e.g. work by Hinchey et al. [22] and Brito et al. [9]).

Our research is in line with the work by Atkinson et al. [4] that shows significant similarities between ontologies and models. The authors argue that the concept of ontology constitutes a subset of the concept of model. Also, Henderson-Sellers [19] points out that a combination of models and meta-models with domain ontologies is helpful in representing vocabularies for specific problem domains. He argues that modelling languages such as UML can describe domain knowledge, but they need particular extensions to provide adequate reasoning support. Our approach goes in this specific direction, as it extends RSL, which is also an extension to UML. In summary, the above discussions give good motivation for our work, where a modelling language that combines ontology constructs is applied to generate code directly from requirements.

Marrying ontologies with models allows applying model-driven techniques and especially model transformations [58]. Appropriate works include more general discussions on introducing ontology-aware transformations [3, 42] and comprehensive formalised propositions on meta-models for ontology languages [44, 15]. An example of such a language is CoCoViLa by Haav, and Ojamma [18]. Our current work can be compared or even contrasted with such approaches, as it introduces a common meta-model for a semantically rich language that can express any problem domain. In this context, an essential feature of our approach is its extensive reliance on inference mechanisms, especially for generating data processing code. It is somewhat similar to business rule engines [13] that use notations like JBoss Rules [11] or SBVR [40]. However, instead of interpreting them during runtime, it generates code fully integrated with the rest of the system. Similarly, our solution can be compared with the approaches that enable code generation directly from ontologies (cf. Ontology-Driven Software Development [2, 17]). Stevenson and Dibson [59] propose a tooling framework for generating Java code from OWL specifications. Another example is work by Völkel and Sure [64] in which Java-based APIs are generated directly from ontologies expressed in RDF Schema. Especially relevant in the context of the low-code paradigm is the work of Krouwel et al. [32] who generate low-code Mendix applications from enterprise ontologies. Finally, our approach can be compared to that of Jbara et al. [24], and Vuorimaa et al. [65]. Both those solutions utilize declarative languages and have reported to have a positive impact on usability, especially for inexperienced ("citizen") developers and domain experts. However, our approach strives to provide more advanced code generation capabilities.

3. RSL: low-code at the requirements level

The necessary background to our research on RSL-DL is the Requirements Specification Language and its tooling environment (ReDSeeDS). As we strive to achieve compatibility

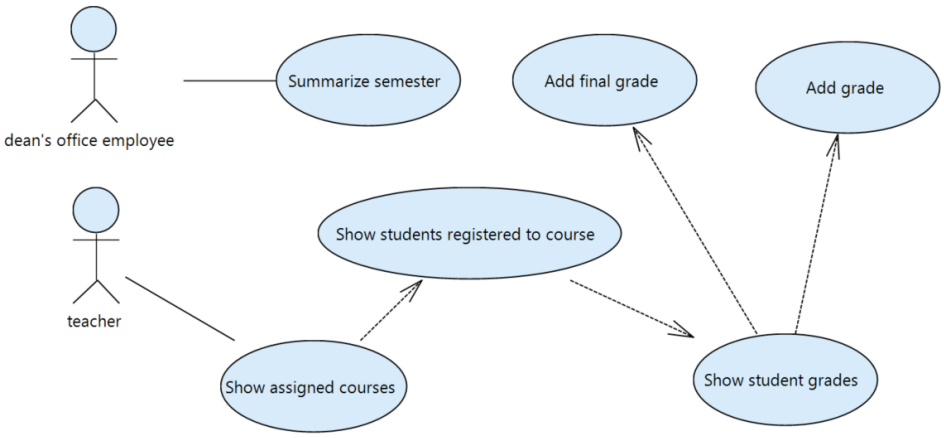


Fig. 2. Example use case model

between these two languages, some aspects of RSL-DL were strongly influenced by RSL. Here, we will present selected RSL constructs that will be most relevant in this context.

Figure 2 presents a simple use case model of a course management system that will serve as a frame for illustrating RSL. An important element worth explaining are the relationships between use cases. RSL has abandoned the ambiguous "extend" and "include" relationships of UML [6]. Instead, it has introduced the "invoke" relationship with procedure call semantics (see, e.g. [56], sec. 2.4.1). A detailed discussion of these relationships is out of the scope of this article. However, we will use an important feature of "invocation" which is parameter passing between use cases.

Figure 3 shows example scenarios of two of the use cases from Figure 2. Note that the "Add final grade" use case can be invoked from other use cases (here: "Show student grades"), which should pass two objects as parameters ("student" and "course"). These two parameters are defined in the "precondition" section of the respective scenario. The other part defines the application logic and contains a sequence of simple subject-verb-object sentences of various types. The most important of them from the point of view of this article are the "Query" type sentences. These sentences define actions of the system that are performed on certain data elements.

According to RSL semantics rules [55], we can transform scenarios into code. Figure 4 presents fragments of an application logic class generated from the presented scenario. The class contains methods for handling user interactions, as specified by the "Select" sentences in the scenarios. For instance, the sentence "Teacher selects save final grade" is translated into the "saveFinalGradeTriggered" method. Contents of these methods reflect consecutive sentences in the scenarios. "Query" and CRUD type sentences are translated into calls to back-end service operation. For instance, the sentence "System computes weighted average grade data" is transformed into a call to the "computeWeightedAverageGradeData" service. UI presentation sentences are translated into calls to the View layer. For instance, the sentence "System shows semester summarized message" is transformed into a call to the "showSemesterSummarizedMessage" method. A more detailed

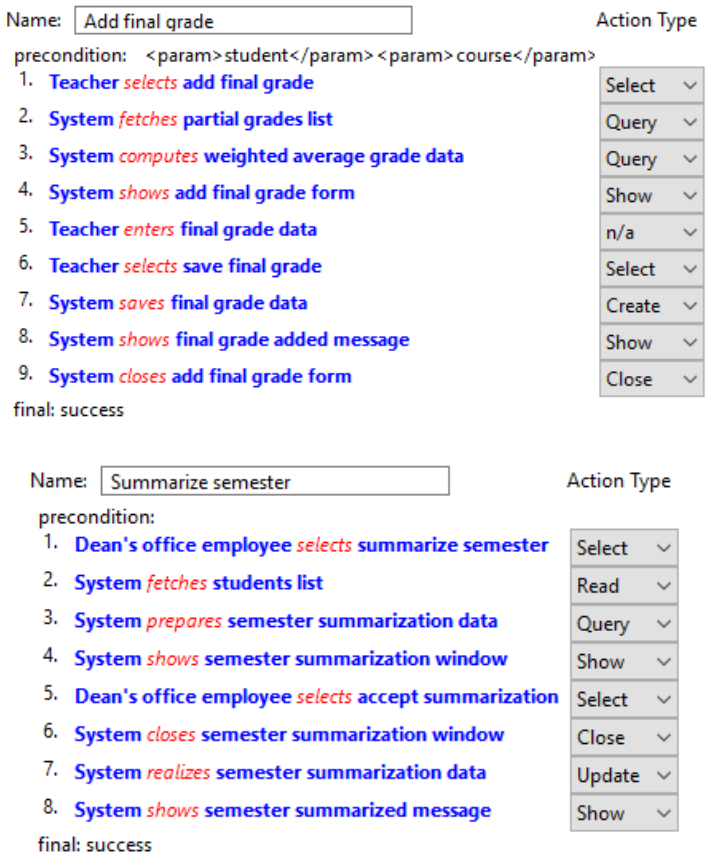


Fig. 3. Example scenarios

discussion of the rules and generated code, including code of the View layer, is presented elsewhere [56].

The relevant parts of the back-end service code generated from the RSL scenario is presented in Figure 5. It contains an interface implementation with empty methods. The operation parameters are determined from scenario sentences before the appropriate calls. In further sections, we will present the syntax and semantics of RSL-DL that will fill the currently empty method bodies. Note that other methods (not relevant to further presentation of RSL-DL) were omitted in Figure 5.

4. RSL-DL syntax

The syntax of RSL-DL aims to represent all information important from the point of view of code generation. It includes detailed dependencies between individual domain elements and proper definitions of the elements themselves. Figure 6 presents an elementary example of concrete syntactic elements of the language. It contains definitions of four

```
public class SummarizeSemesterPresenter extends
                AbstractUseCasePresenter {
    // ...
    public void summarizeSemesterTriggered(){
        studentListDTO =
            service.readStudentList();
        semesterSummarizationDataDTO =
            service.preparesSemesterSummarizationData
                (studentListDTO);
        view.showShowSemesterSummarizationWindow(this);
        pageOpened();
    }

    public void saveFinalGradeTriggered(){
        view.closeSemesterSummarizationWindow();
        pageClosed();
        service.realizesSemesterSummarizationData
            (semesterSummarizationDataDTO);
        view.showSemesterSummarizedMessage();
    }
    // ...
}
```

Fig. 4. Presenter code generated for the use case scenarios

"Identity" type entity notions (student, course, partial grade, weighted grade), describing concrete objects in the specific problem domain. Notions can also have conditions, and in our example, we can see one kind of condition: "inheritance". Thus, the condition for the "partial grade" notion is that it must follow all the rules for the "weighted grade" notion. In addition to entity notions, we can define property notions, like "grade weight" in Figure 6. This kind of notions define concrete atomic values and can be used as attributes of other notions, which can be indicated by "attribute links" (lines with a diamond shape).

Relationships in RSL-DL (see "grading" in Figure 6) are represented by hexagons and can link many notions. To some extent, this syntax resembles that of UML's n-ary associations. In our example, the relationship is of type "Data Based Reference" which is a basic type that reflects the situation where references between objects are contained in their data (e.g. in their attributes). In our current example, "student" and "course" contribute to the relationship "grading" that results with a "partial grade". Arrow directions distinguish between types of notion participations in the specific relationship. Since a given student can have many partial grades in a course, then the particular participation is marked as "multiple".

The abstract syntax for the above-presented core language elements is presented in Figure 7 using the MOF notation [41]. The meta-class "DLNotion" represents notions and the meta-class "DLRelationship" represents dependencies between them. Concrete participations of notions in relationships are represented by the meta-class "DLRelationshipParticipation". The meta-model contains two types of such participations – standard and auxiliary. Standard ones correspond to the main subjects of relationships and are denoted with solid arrows in concrete syntax. Auxiliary ones point to elements that define relationship contexts and are denoted with dashed lines. For example, one could use a relationship context to indicate which object should be used when computing values based on that object's attributes. In this case, the attributes participate through standard participations, and their "parent" participates through auxiliary participation.

```
public class ServiceImpl implements IService {
    //...
    PartialGradeListDTO readPartialGradeList
        (Long inputStudentID, Long inputCourseID)
    {}

    WeightedAverageGradeDataDTO
        computeWeightedAverageGradeData
            (List<PartialGradeListItemDTO> partialGradeListDTO,
                    Long inputStudentID, Long inputCourseID)
    {}

    List<SemesterSummarizationDataItemDTO>
        preparesSemesterSummarizationData
            (List<StudentListItemDTO> studentListDTO)
    {}

    void realizesSemesterSummarizationData
        (List<SemesterSummarizationDataItemDTO>
            semesterSummarizationDataDTO)
    {}
    //...
}
```

Fig. 5. Back-end access code generated for the two example scenarios

Besides notions, there is a special kind of relationship participants – primitives ("DL-Primitive" meta-class). These elements define general concepts that do not have concrete instances. Examples of such primitives in RSL-DL are "current date", "number Pi" and "Planck constant".

As indicated above, notions can have types. The first one ("identity") was explained in the example above. The "template" type indicates templates that can be used to simplify defining other notions. These two types correspond approximately to concrete and abstract classes of e.g. UML. Two other types define notions whose representatives' (objects') roles can change during their lifetime. It is inspired by ontology-based inference engines with their capabilities to "discover" object types or change them dynamically. The "inferred role" type indicates roles that can be inferred, e.g. from various status attributes

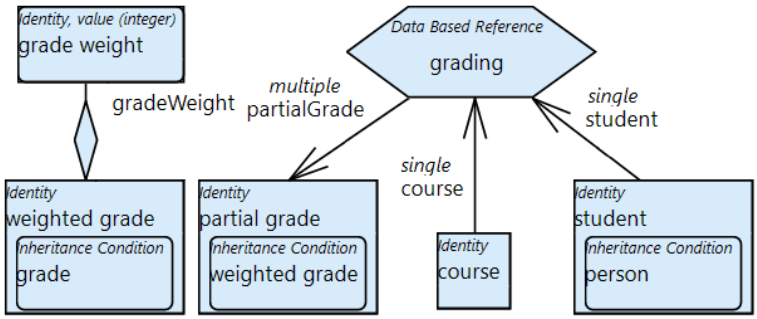


Fig. 6. RSL-DL concrete syntax example

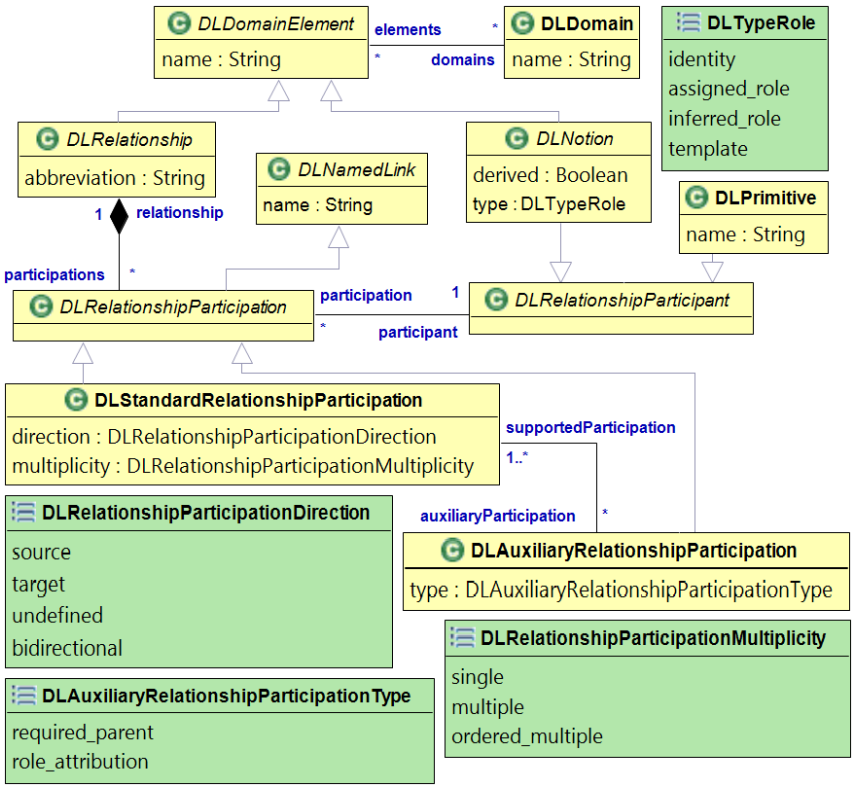


Fig. 7. Meta-model fragments for core RSL-DL elements

of an object. The "assigned role" type indicates roles that can be explicitly changed during the lifetime of an object.

More details related to the syntax for notions are shown in Figure 8. The "DLProperty" meta-class is used to denote notions with concrete atomic values or value sets. The "DLEntity" meta-class denotes more complex notions that cannot be reduced to single values. The "DLAttributeLink" meta-class allows indicating attribute dependencies between notions. Such links can be marked as "derived", which means that their values need to be inferred from other notions. An important type of notion features are conditions ("DLCondition" meta-class). Their role is to further detail notion characteristics. Apart from the previously described "inheritance condition", two additional condition types exist. The "identity condition" type defines conditions that have to be fulfilled for a given notion's object to make sense. The "validity conditions" type defines conditions that denote the correctness of a given notion's object. In general, there can exist objects that meet appropriate identity conditions but do not meet validity conditions and thus are treated as invalid but belonging to the given notion. We should note that conditions do not include graphical links to other model elements. It is due to their potential complexity and interweaving. Thus, for instance, inheritance was not represented using a simple arrow as in UML.

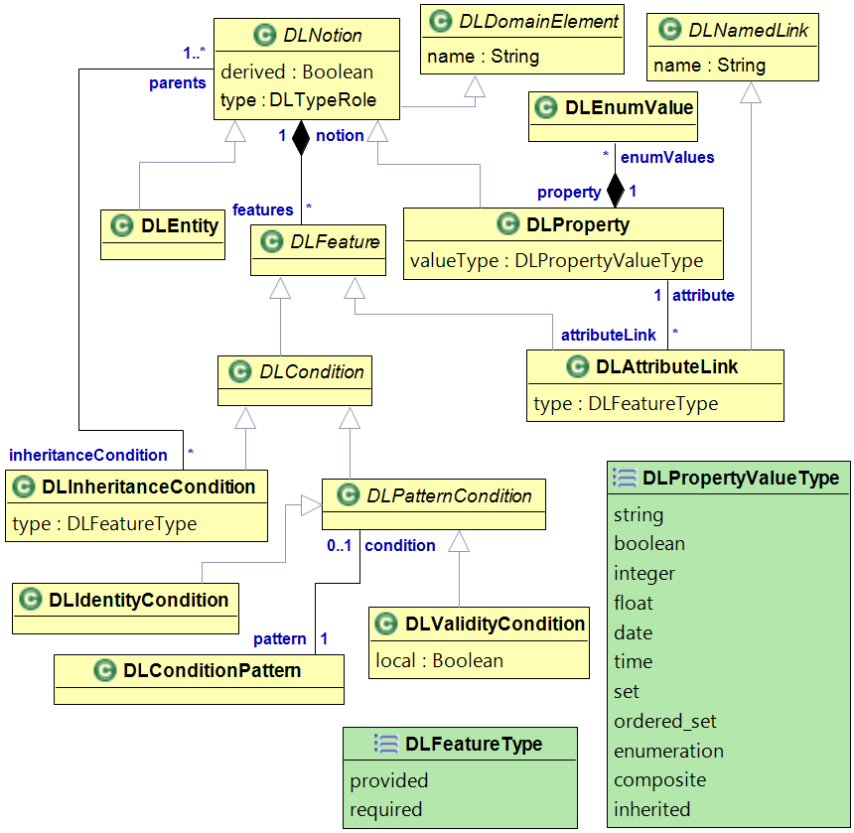


Fig. 8. Meta-model fragments for notions

Figure 9 presents the hierarchy of relationships found in RSL-DL. From the conceptual point of view, two main types of dependencies exist between notions in the problem domain that are significant for code generation. It is reflected in RSL-DL through dividing relationships into two categories: transitions ("DLTransition") and references ("DLReference"). Transitions describe how to obtain notion objects based on other notion objects. References describe specific roles played by objects in relation to other objects. Both relationship categories are further divided based on how they are defined. "Transitions" can be described using simple rules ("DLPatternBasedTransition") or algorithms consisting of many steps ("DLAlgorithmicTransition"). "References" can be described using rules that define certain conditions ("DLPatternBasedReference") or take the form of the previously described data-based references ("DLDataBasedReference"). This division of references is inspired by the division into fact and rule spaces found in ontologies. In practice, of all these relationships, the algorithmic transitions are not preferred as they are not fully declarative and thus arguably less usable [24, 65].

Figures 8 and 9 contain two additional elements: "DLTransitionPattern" and "DLConditionPattern". Each instance of these two meta-classes contains a string with a textual condition formula, a specific condition type and optionally – the condition's subject link

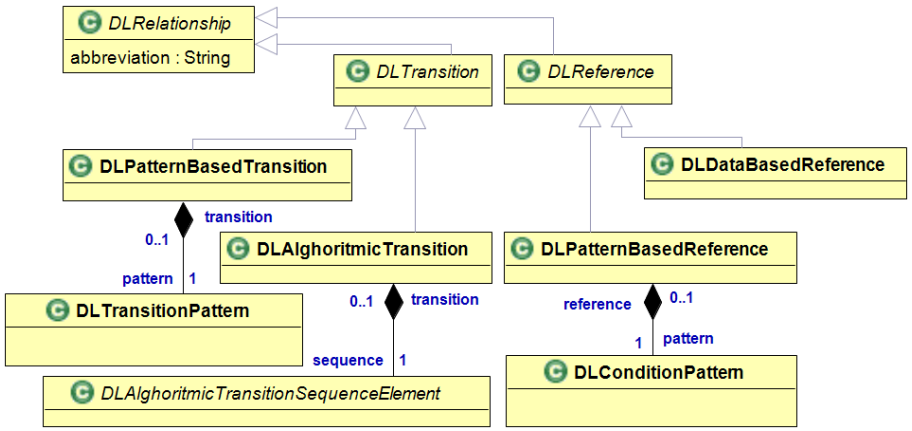


Fig. 9. Meta-model details for relationships

(see the detailed metamodel in the Supplement). The syntax of the condition is expressed in a language based on the notation used in the Symja library [31]. Some examples of several types of these patterns are given in section 7 in descriptions of the generated code.

5. Translational semantics rules

Of the many approaches to define semantics for RSL-DL (treated as a programming language [54]) we choose the translational method, which is more in line with the model-driven paradigm. This approach defines rules that translate specific patterns of RSL-DL constructs into fragments of Java code. Each rule has an informal textual description and is formalised as a procedure in the MOLA graphical model transformation language [26].

Full specification of semantics for RSL-DL consists of 16 translational rules (all the details can be accessed in the Supplement). The first ten rules define the generation of the target Java class structure, including their fields and method signatures. These rules depend only on the structure of notions and relationships between them, found in a particular RSL-DL model. The following two rules additionally use an inference engine and are used to generate method bodies. Rules 13-16 further add to the generation of method bodies. They use a symbolic computation library to transform Symja-based formulas found in pattern condition expressions into the contents of method bodies. The results are used directly or as part of a loop or a condition depending on the pattern type. We summarise the responsibilities of specific rules in the following list.

- Rule 1 – generate classes from notions, except for those (“simple”) properties that do not contain other properties as attributes.
- Rule 2 – add inheritance dependencies between classes based on inheritance conditions.
- Rule 3 – generate a properly typed value field within a class generated from a “complex” property (a property that contains other properties as attributes).
- Rule 4 – generate fields derived from attribute links within notion-based classes.

- Rule 5 – generate appropriate constructors for notion-based classes, taking into account inheritance conditions and attribute links.
- Rule 6 – generate validation methods in notion-based classes, derived from validity conditions contained within the respective notions.
- Rule 7 – generate eligibility-check methods in notion-based classes, derived from identity conditions within the respective notions, taking into account inheritance conditions.
- Rule 8 – generate RUD (Read-Update-Delete) methods in notion-based classes.
- Rule 9 – generate classes from relationships and static methods within these classes; the generated methods retrieve objects of one type (cf. target relationship participant) derived from objects of other types (cf. source relationship participants).
- Rule 10 – generate additional methods in the above-generated classes; the methods check for the existence of particular relationships between objects potentially participating in a given relationship.
- Rule 11 – generate methods that respond to queries formulated within the application logic code; this rule allows to integrate RSL-DL-based code with code generated from RSL.
- Rule 12 – generate methods that compute values for attribute links marked as derived.
- Rule 13 – generate code that checks conditions based on condition patterns; this code is used in various other parts of the generated code.
- Rule 14 – generate code that filters objects based on condition patterns; used as in Rule 13.
- Rule 15 – generate code derived from transition patterns; used as in Rule 13.
- Rule 16 – generate auxiliary code that accesses full sets of objects of a given type or values corresponding to primitives; used as in Rule 13.

In summary, each "non-trivial" notion in the source RSL-DL model produces two Java classes. One class represents (in simplified terms) a data transfer object (DTO) corresponding to the given notion. The other class is a utility class that holds various data handling methods. These classes are appropriately amended with CRUD and condition-related operations. The "DTO" classes are also organised in an appropriate inheritance hierarchy. Additionally, supportive classes are created for all the relationships in the model. These classes contain methods that return objects participating in relevant relationships.

As introduced above, all the rules are formalised using MOLA procedures. MOLA uses a declarative-imperative visual syntax presented in Figures 10, 12 and 13. Its imperative flow definition is based on a notation resembling activity diagrams in UML. Arrows denote control flow. Iteration "actions" are denoted with thick black frames. Rule "actions" constitute the declarative part of the language. Each rule contains a query on objects expressed through a diagram resembling a UML object diagram combined with a MOF meta-model diagram. Black solid lines denote queried objects, while red dashed lines denote created objects. More details, including the MOLA handbook [1] and a tutorial can be found on the MOLA website[3].

For brevity, we will limit our presentation of rule formalisation to two selected rules – 10 and 11. Figure 10 presents formalisation for the first of them. As the above list indicates, this particular rule defines the creation of methods (see: "check : Operation") that

[3] http://mola.mii.lu.lv/

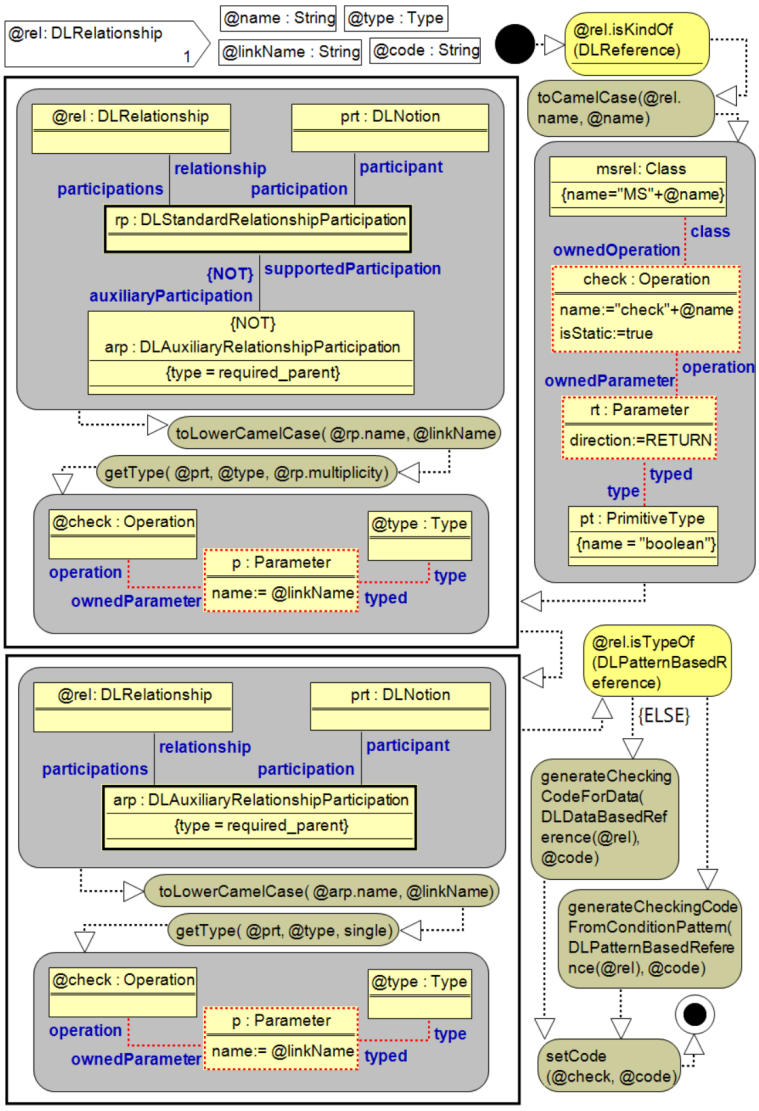


Fig. 10. Formalization of transformation rule 10

check the participation of objects in relationships. The generated method code returns a logical value (see: "rt : Parameter") and has parameters (see: "p : Parameter") generated through iterating over all the participants assigned to the relationship at hand. Since auxiliary participations need to be processed somewhat differently, the formalisation contains two such iterations. Additionally, appropriate procedures generating method code based on patterns ("generateCheckingCodeForData", "generateCheckingCodeFromCondition-Pattern") are invoked at the end. These procedures constitute formalisations of other rules (see Rules 13–16).

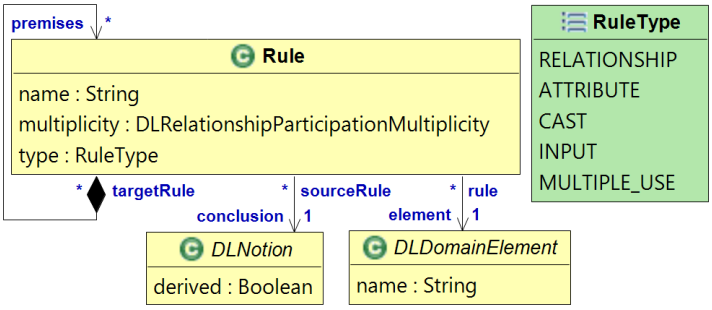


Fig. 11. Inference rule meta-model

To implement Rule 11, we need to use a dedicated inference engine, implemented as part of this work. The engine processes queries derived from "Query" type scenario sentences (see again Figure 3). For each such query, it produces a sequence of inference rules, where each of the rules is based on domain elements defined within an RSL-DL model. The appropriate sequences can be represented with a meta-model shown in a simplified form in Figure 11. The meta-model uses a structure of nested "Rule" meta-classes to reflect appropriate sequences of inference invocations needed to solve specific problems. Each "Rule" points to a domain element ("element") that is the basis for generating a specific method according to one of the previous rules. The "type : RuleType" meta-attribute defines the concrete type of such generation. This "Rule" meta-class also points to a "conclusion" that constitutes a specific notion reflecting objects being the inference results. Furthermore, the given rule's "premises" constitute other rules, preceding this rule in the rule sequence. Base premises that reflect query parameters are represented as additional "artificial" rules.

The algorithm that generates the respective sequence of method invocations from the inference rule structure is presented in Figure 12. It starts from invoking itself ("generateCodeFromSolution") recursively for all the "premises" of the current rule and joining code generated from these premises. If the current element is used as a "premise" in other rules, a proper variable is declared based on the object corresponding to the rule's "conclusion". Otherwise, a "return" statement is generated. In both cases, the way to obtain the assigned or the returned value depends on the type of the rule. In most cases, such a value is obtained by invoking an appropriate "get" method. This method retrieves an object that corresponds to the "conclusion" and is contained in the class derived from the rule's "element". Besides, the "get" method's call accepts parameters that correspond to the rule's premises.

The actual formalisation of Rule 11 that uses the above algorithm is presented in Figure 13. It defines a procedure for creating a method, where the method's contents are generated with the algorithm. The procedure starts by creating a method with the name based on the "conclusion" of the final rule and having the prefix "get". This method is placed in the class derived from the domain element pointed to by this final rule. The return type of this method again corresponds to the rule's "conclusion", and the method's parameters are based on the "premises" within the whole rule sequence.

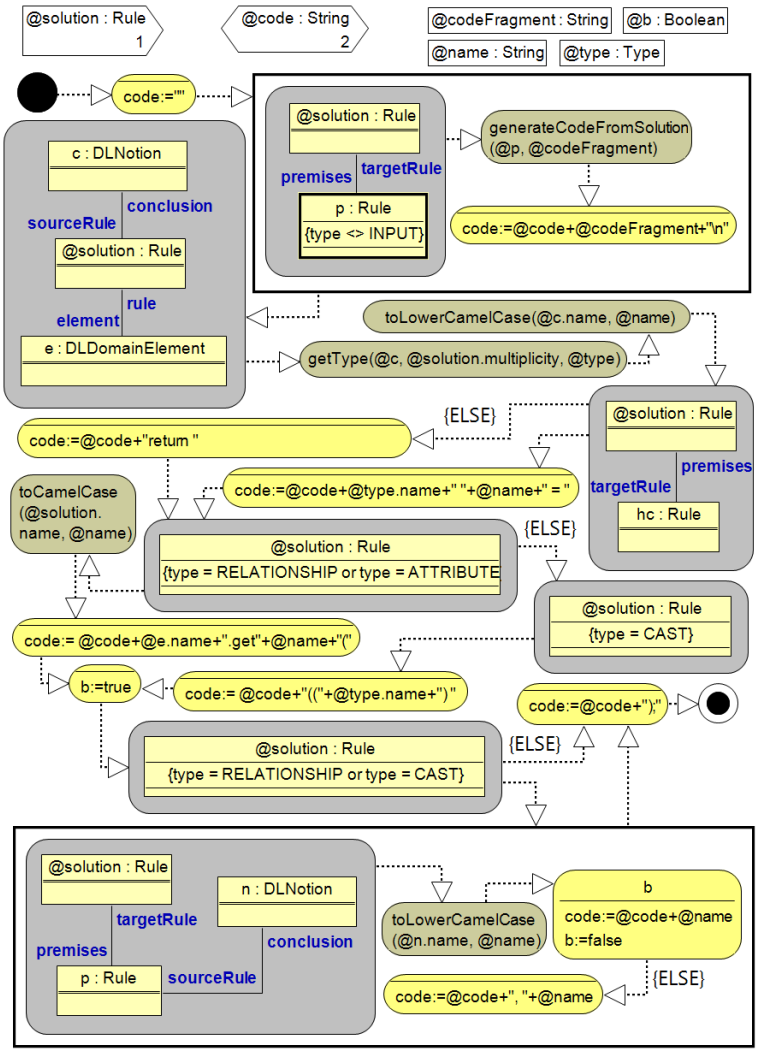


Fig. 12. Algorithm for generating the method body in rule 11

6. Implementation

The presented syntax and semantics of RSL-DL were implemented as an extension to the ReDSeeDS tool. We have constructed a dedicated RSL-DL editor using typical approaches to the construction of visual languages defined with metamodels. This includes the usage of Eclipse Modeling Framework (EMF) with Graphical Modeling Framework (GMF, currently superseded by the Sirius framework) for the implementation of the model repository and the graphical editor. This was then appended with a code generation component.

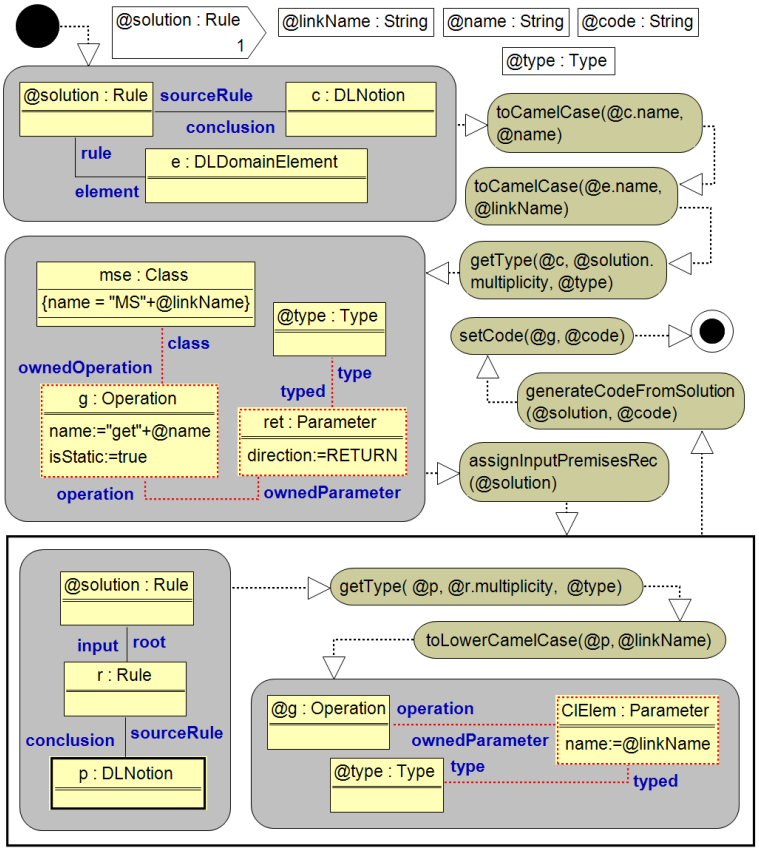


Fig. 13. Formalization of transformation rule 11

Figure 14 presents an overview of the code generation process. The inputs to the process are the RSL and RSL-DL models. The RSL model is used to generate front-end code, as explained in Section 3 [56] ("RSL-to-code transformation"). Scenario sentences from this model (typed as "Query") are also used to determine queries that form input to the inference engine. The inference engine takes these queries and applies them to the RSL-DL model. This results in a set of rules that will be used to generate domain code ("Rules necessary to implement domain operations"). These rules are organised into a tree structure that reflects full inference paths with references to appropriate domain elements from the RSL-DL model (see Figure 11).

The next step is to use a symbolic computation library to transform formulas contained in the rule tree. The goal is to obtain the formulas in a form suitable for the target logic (according to the queries). These transformed rules and the original RSL-DL model form inputs to the final code generation engine ("RSL-DL-to-code transformation"). This engine uses the rules presented in section 5 to produce the back-end code. This is joined with the front-end code that results in the final code for the system.

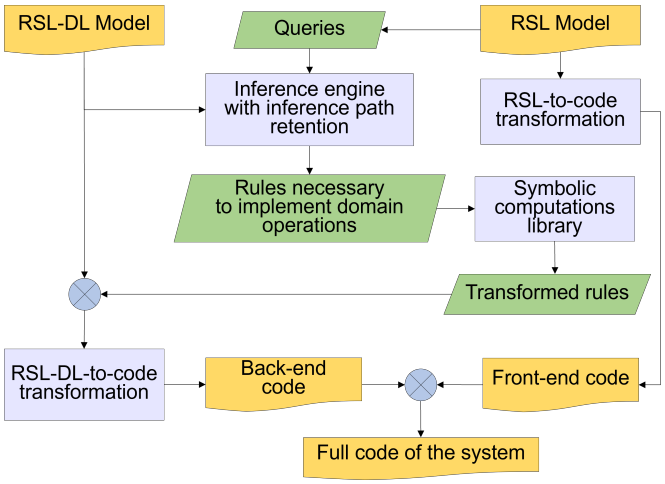


Fig. 14. Overview of the RSL-DL code generation process

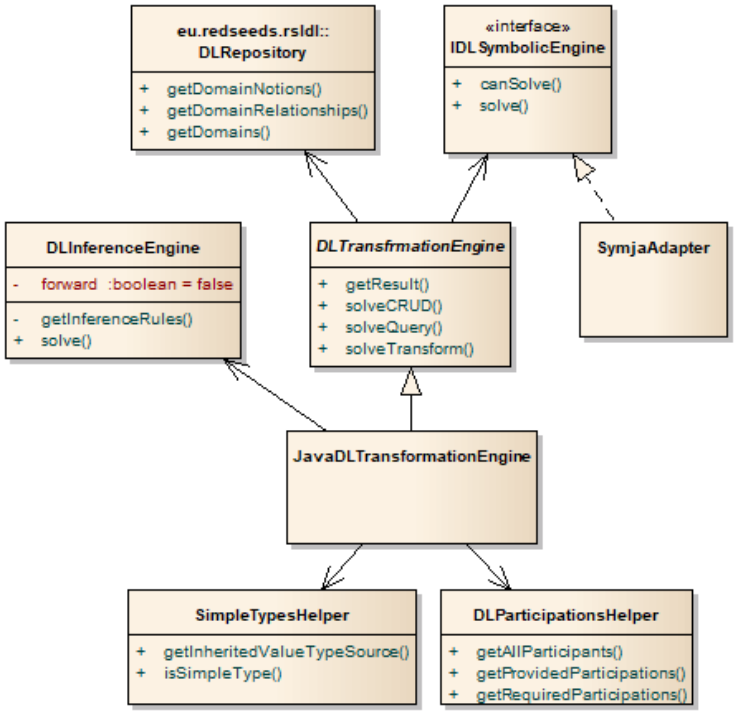


Fig. 15. Overview of the RSL-DL code generation module

The above described code generation component was implemented in Java as a plug-in to the ReDSeeDS tool. Its overall structure is presented in Figure 15 (we omit all the

technical details related to creating Eclipse plug-ins). The main element used to implement RSL-DL code generation is the "DLTransformationEngine" class. It defines the general scheme that the transformation should follow and serves as a basis for implementing transformation classes for specific target languages. As an example, we have implemented the "JavaDLTransformationEngine" class that generates Java code. Such code generators have access to model elements trough the "DLRepository" class, which provides instances of model elements based on EMF repository classes.

Access to symbolic computations is provided through the "IDLSymbolicEngine" interface. This interface contains operations for converting formulas contained in model elements (and inference rules). It is implemented using the SymJa symbolic computation library [31]. Transformation implementations should make use of three more provided classes: "DLInferenceEngine", "SimpleTypesHelper" and "DLParticipationsHelper". The "DLInferenceEngine" class implements the inference mechanisms described above. In response to queries, it provides the inference rule trees. The "SimpleTypesHelper" class helps in determining whether particular concepts should be represented in code as generated classes or built-in types. The "DLParticipationsHelper" class allows for more convenient handling of relationship participations during code generation. Full source code of the implementation can be accessed from the ReDSeeDS repository.

7. Case study

This section will present selected fragments of two more extensive case studies that illustrate several important uses of RSL-DL. The first case study refers to the functional requirements specification presented in Section 3. The second case study introduces a new problem domain (banking) and gives more insight into the inference engine and domain knowledge reuse.

7.1. Course Management System case

The RSL part of this case study was presented in Section 3. Thus, here we will concentrate only on presenting examples of the various concrete RSL-DL language constructs, generated domain logic code, and references to application logic code from Section 3.

Figure 16a involves constructs for checking specific conditions. The model contains elements that describe information related to checking whether the given student is eligible to get a registration for the next semester. It consists of three basic notions – "student", "course" and "final grade". All of them are connected by the "final grading" relationship, which is a data-based reference. It indicates that information about concrete dependencies between representatives of these notions is stored in some data objects. Finally, we define the "student to accept" relationship that is used to define conditions about students' eligibility to be registered for the next semester. The concrete condition embedded in this relationship (not shown here) requires that all the final grades for the student's courses have a value of at least 3 (minimum passing level). The relationship has only one participant – the student, that participates as its target.

Figure 16b involves constructs for finding elements consistent with specific conditions. It defines the "academic year" as having the beginning and the end date. It also shows the "current academic year" relationship that enables the filtering of academic years

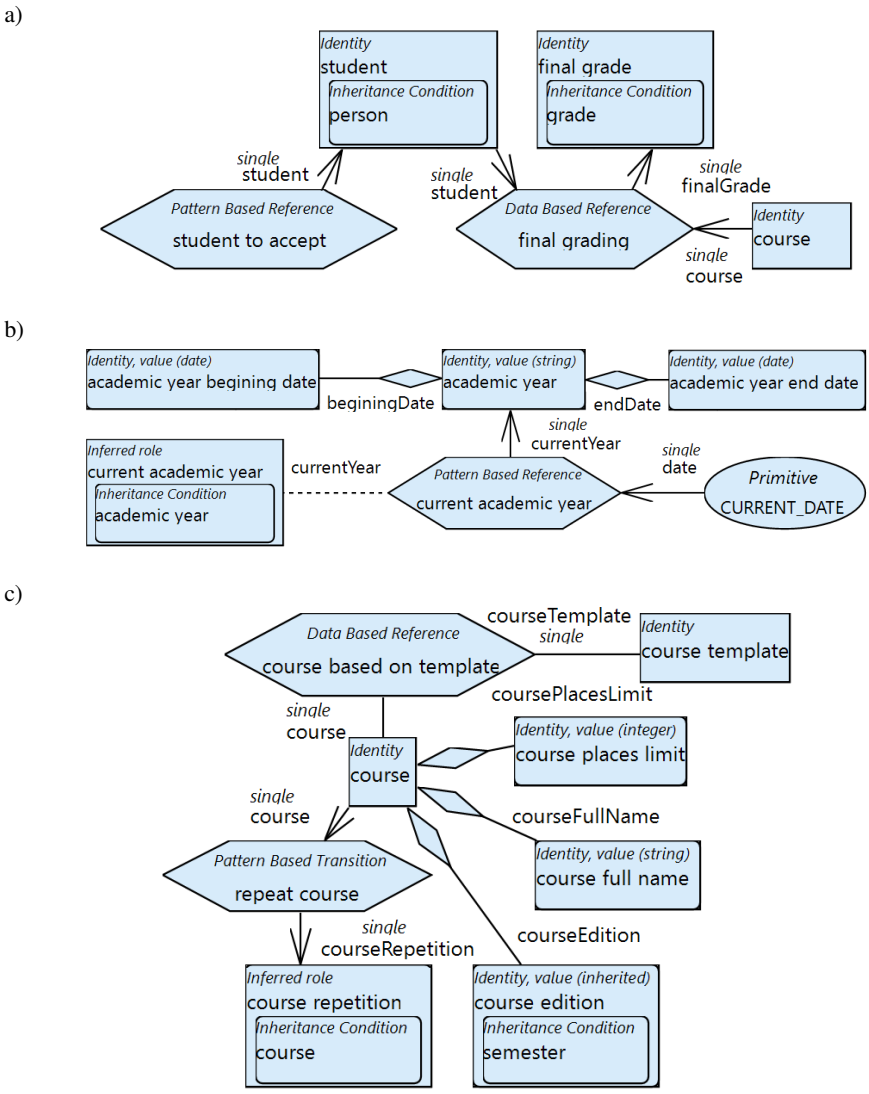


Fig. 16. RSL-DL model defining eligibility of students to be registered for the next semester (a), knowledge about academic years (b), and knowledge about course editions (c)

based on some condition. This condition (not shown in the figure) states that the current date should be between the beginning and the end date of the requested academic year. Additionally, the model distinguishes the "current academic year" notion. This notion is linked with the above relationship through an auxiliary relationship participation denoted with a dashed line. This means that whenever an inference is made based on the above relationship, to obtain an "academic year", then this year can be treated as the "current academic year". It allows for more convenient usage of conditions relating to academic years in other parts of the model. It is also worth noting that the multiplicity of the partic-

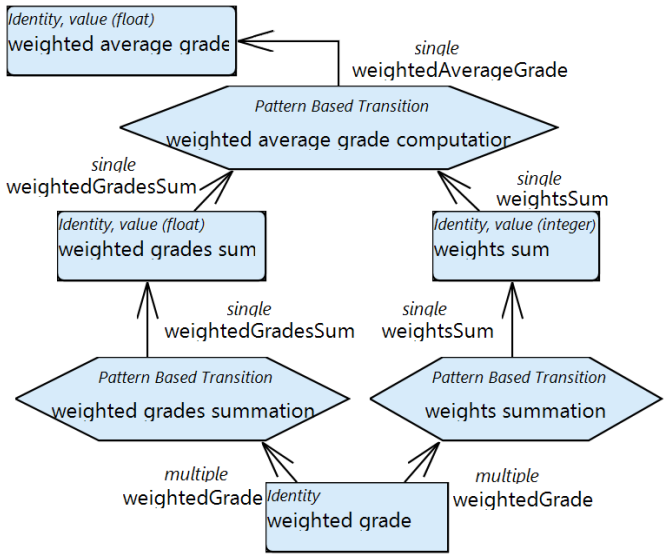


Fig. 17. RSL-DL model containing knowledge about computing of weighted average grades

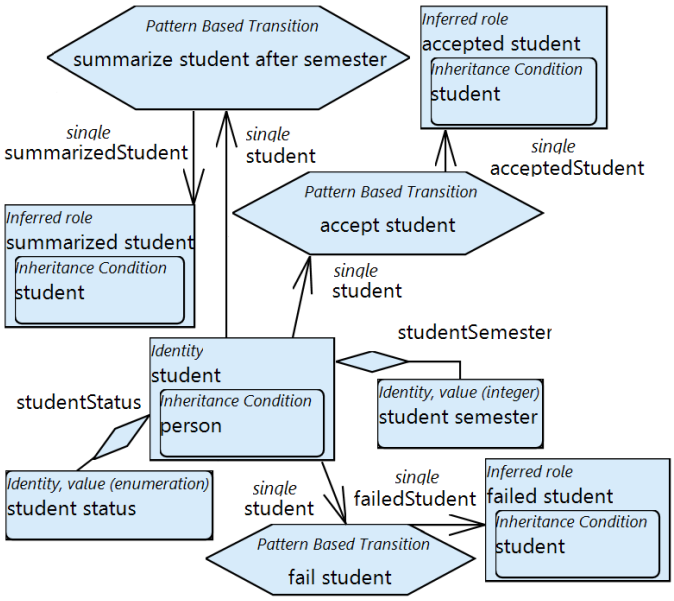


Fig. 18. RSL-DL model containing rules for promoting students to the next level of studies

ipation pointing to the academic year is denoted as single. Thus, there should be only one current academic year for the given "CURRENT_DATE". Considering that the "CURRENT_DATE" is specified as a primitive, there should be only one current academic year at any given moment.

```
public class MSStudentToAccept {
    public static boolean checkStudentToAccept(
        IMStudent student){
            for (IMFinalGrade $_iter:
                student.getFinalGrades())
                    if (!$_iter.getGradeValue()>=3)
                        return false;
            return true;
        }
    public static List<IMStudent> getStudents(){
            List<IMStudent> students
                = MSStudent.getStudents();
            List<IMStudent> result
                = new ArrayList<IMStudent>();
            for (IMCourse $_iter:students)
                if (checkStudentToAccept($_iter))
                    result.add($_iter);
            return result;
        }
}
```

Fig. 19. Code for checking students' registration eligibility

Figure 16c involves constructs resulting in creating complex objects. It focuses on the "repeat course" relationship that allows creating another edition of a course for the current semester, based on one of the previous editions. The "previous" edition can be simply any course represented in the model as the participation leading from the "course" notion. The "new" course is represented as the "course repetition" notion, which also inherits from the "course" notion. The figure also shows additional information about courses. For example, each course is based on a "course template" through the "course based on template" relationship. Note that the direction of relationship participations indicate possible direction of inference and can be used to optimise the inference process.

Figure 17 involves constructs for computing values. It contains the "weighted average grade computation" relationship that contains a transition pattern with an equation that computes the weighted average grade. This equation requires two other values represented by the notions "weighted grades sum" and "weights sum". Therefore, the model contains also transitions that allow for the computation of these two values.

The final part of the presented model fragment involves constructs for modifying complex objects and is shown in Figure 18. The situation here is somewhat similar to the previous one. The whole modification is handled by the "summarize student after semester" transition, which requires two other transitions: "accept student", and "fail student". In contrast to the previous case, these transitions will be used interchangeably, depending on the fulfilment of a condition. This condition refers to the "student to accept" relationship presented in Figure 16a. If the student meets the "student to accept" relationship, the "accept student" transition is invoked, while in the opposite case, the "fail student" transition is invoked.

The next step in our case study is to generate code. The above specification in RSL-DL was formulated within a dedicated RSL-DL tool. The tool also includes a code generation engine that implements all the rules introduced in the previous section. Here we will

```
public class MSCurrentAcademicYear {

    public static boolean checkCurrentAcademicYear
        (IMAcademicYear academicYear){
            LocalDate date = LocalDate.now();
            return academicYear.getBeginingDate().isBefore(date)
                && academicYear.getEndDate().isAfter(date);
        }

    public static IMCurrentAcademicYear getCurrentYear()
            List<IMAcademicYear> academicYears
                = MSAcademicYear.getAcademicYears();
            for (IMAcademicYear $_iter:academicYears)
                if (checkCurrentAcademicYear($_iter))
                    return new MCurrentAcademicYear($_iter);
        }

}
```

Fig. 20. Code for determining the current academic year

present some key fragments of code generated from the above-presented excerpts of the dean office model.

Figure 19 shows code generated on the basis of the model from Figure 16. The actual class ("MSStudentToAccept") is generated per rule no. 9 (generate classes and static methods from relationships). It is a supportive class that corresponds to the "student to accept" relationship and contains only static methods. The "checkStudentToAccept" method was generated based on rule no. 10 (generate existence checking methods). It checks for the eligibility of a given student to be accepted for the next semester. The student is passed as the parameter of this method. It can be noted that the type of this parameter ("IMStudent") is the class corresponding to the "student" notion, generated according to rule no. 1 (generate classes from notions).

The method returns a logical value reflecting the result of the eligibility check. The actual check is based on the contents of the condition pattern in the "student to accept" relationship (not shown in Figure 16). This pattern is defined through three values: 1) formula "gradeValue($)>=3", 2) type "universal quantification", and 3) subject link "finalGrading(student)" that relates to the "final grading" relationship. Note that the "$" sign denotes the target of the "final grading" relationship, which is the "final grade" notion. The above formula was transformed into the appropriate "if" statement in Figure 19. The "for" loop is generated based on rule no. 13 (generate condition checking code from condition patterns), considering the above pattern type. This way, the eligibility check is done for all the grades of a particular student.

The second method of this class ("getStudents") applies the above eligibility check to all the students. This method was generated based on rule no. 14 (generate object filtering code from condition patterns), and it filters out all the subjects (here: students) that fulfil an appropriate condition pattern (here: student eligibility check). We should also note that the code for obtaining the list of all students was generated using rule no. 16 (generate auxiliary code).

Figure 20 shows code generated on the basis of the model from Figure 16b. As in the previous code fragment, the generated supportive class contains two methods. The

```
public class MSRepeatCourse {

    public static List<IMCourseRepetition>
        getCourseRepetitions(List<IMCourse> courses){
            List<IMCourseRepetition> result
                = new ArrayList<IMCourseRepetition>();
            for (IMCourse $_iter:courses)
                result.add(getCourseRepetition($_iter));
            return result;
        }

    public static IMCourseRepetition
        getCourseRepetition(IMCourse course){
            IMCourse courseRepetition = new MCourse();
            courseRepetition.setCourseEdition(
                    MSNextSemester.getNext(
                            course.getCourseEdition()));
            courseRepetition.setCoursePlacesLimit(
                    course.getCoursePlacesLimit());
            courseRepetition.setCourseTemplate(
                    course.getCourseTemplate());
            return new MCourseRepetition(courseRepetition);
        }

}
```

Fig. 21. Code for creating new course editions

first one ("checkCurrentAcademicYear") checks if the given academic year is the current one. It starts with code for obtaining the current date (transformation of the primitive "CURRENT_DATE") according to rule no. 16 (generate auxiliary code). Following this, the method code returns a logical value computed based on the condition pattern in the "current academic year" relationship. The pattern formula (not shown in Figure 16b) has the value of "beginingDate(currentYear) < date && endDate(currentYear) > date" and is typed as "simple". The second method in Figure 20 ("getCurrentYear") uses the first to filter out academic years and return the current one. This method returns just a single object because the appropriate target participation is of the "single" type. It also casts the found year to the class corresponding to the "current academic year" notion ("MCurrentAcademicYear"). It is done based on rule no. 9 (generate classes and static methods from relationships), which determines the correct types within method signatures. In this case, the appropriate casting is done based on the "currentYear" auxiliary participation ("current academic year" notion participating in the "current academic year" relationship).

Figure 21 shows code generated on the basis of the model from Figure 16c. This time, the generated code creates objects representing new course editions based on previous ones. Nevertheless, again, the presented class ("MSRepeatCourse") contains two methods generated according to rule no. 9 (generate classes and static methods from relationships). The first invokes the second one over a set of appropriate objects (list of courses). Contents of the second method ("getCourseRepetition") are generated from rule no. 15 (generate code from transition patterns). It is translated from the transition pattern contained in the "repeat course" relationship. This pattern's formula is "courseEdition(courseRepetition) == nextSemester (courseEdition(course)); coursePlacesLimit(courseRepetition) == coursePlacesLimit(course); template(courseRepetition) == template(course)". It simply states

```
public class MSWeightedAverageGradeComputation {
    public static double
        getWeightedAverageGrade(double weightedGradesSum,
                int weightsSum){
            return weightedGradesSum/weightsSum;
        }
    public static double
        getWeightedAverageGrade(List<IMPartialGrade> partialGrades){
            List<IMWeightedGrade> weightedGrades
                = new List<IMWeightedGrade>(partialGrades);
            double weightedGradesSum
                = MSWeightedGradesSummation
                    .getWeightedGradesSum(weightedGrades);
            int weightsSum = MSWeightsSummation
                    .getWeightsSum(weightedGrades);
            return getWeightedAverageGrade(weightedGradesSum,
                    weightsSum);
        }
}
```

Fig. 22. Code for computing weighted average grades

that appropriate attributes of the repeated course should be the same as those in the base course or appropriately modified (here: the edition is incremented). This kind of transition pattern ("indirect") contains a set of "simple" transition pattern formulas that together allow for obtaining an object of the given type.

Figure 22 shows code generated on the basis of the model from Figure 17, using data structured according to Figure 6. This time, the situation in somewhat different to that in the previous code fragments. This part of the code results from answering a query that asks to compute the "weighted average grade" for the given set of grades. It contains two overloaded methods ("getWeightedAverageGrade"). The second one (with the "List" parameter) is generated according to rule no. 11 (generate methods for queries). It accepts a list of partial grades and produces a specific average value. The method contains a sequence of method calls that reflect the sequence of inference rules returned by the inference engine (see Figure 12). The first one of the overloaded methods is called from the second one. Its signature was generated according to rule no. 9 (generate classes and static methods from relationships). Its body was based on translating a transition pattern with the "simple" formula "weightedGradesSum/weightsSum" according to rule no. 15 (generate code from transition patterns).

Figure 23 shows code generated on the basis of the model from Figure 18. This time we can see only one method ("getSummarizedStudent") generated according to rule no. 9 (generate classes and static methods from relationships). The method's body is generated on the basis of a "mapping" transition pattern with the formula "studentToAccept(student); acceptStudent(student); failStudent(student)". Formulas for this type of transition patterns are composed of three sections: a "simple" condition pattern formula and two "simple" transition pattern formulas (the first one used when the condition is true, and the second one otherwise). Here, the condition pattern refers to the "student to accept" relationship shown in Figure 16. Thus, the current method calls the "checkStudentToAccept" method shown in Figure 19. Depending on its result, it calls one of two methods resulting from transforming the "accept student" and "fail student" relationships.

```
public class MSSummarizeStudentAfterSemester {
    public static List<IMSummarizedStudent> getSummarizedStudents
        (List<IMStudent> students) {
        List<IMSummarizedStudent> result =
                new ArrayList<IMSummarizedStudent>();
        for (IMStudent $_iter:students)
            result.add(getSummarizedStudent($_iter));
        return result;
    }
    public static IMSummarizedStudent getSummarizedStudent
        (IMStudent student) {
        if (MSStudentToAccept.checkStudentToAccept(student))
            return new MSummarizedsStudent(MSAcceptStudent.
                    getAcceptedStudent(student));
        return new MSummarizedsStudent(MSFailStudent.
                getFailedStudent(student));
    }
}
```

Fig. 23. Code for determining student's eligibility for promotion to the next semester

```
public class ServiceImpl implements IService {
    // ...
    List<SemesterSummarizationDataItemDTO>
        preparesSemesterSummarizationData
            (List<StudentListItemDTO> studentListDTO)
    {
        return MSSummarizeStudentAfterSemester.
            getSummarizedStudents(studentListDTO);
    }
    // ...
}
```

Fig. 24. Additional back-end access code

The other method in this class invokes the above described one over a set of appropriate objects (list of students).

Finally, relevant code fragments as presented above, can be now applied to fill-in appropriate empty methods of the back-end service class (see Figure 5). In our example this pertains to the method that corresponds to a non-CRUD operation. Appropriate additional code is presented in Figure 24. It contains a simple call to the operation presented in Figure 23. It is worth noting that such operations are optimised to use only required parameters.

7.2. Banking case study

In this study we consider a fragment of a banking application covering some operations regarding loans and deposits. Figure 25 contains four use cases that constitute the context for the example models and code. We will concentrate on the functionality of two use cases where their scenarios are presented in Figure 26. The first use case consists of two main parts. The first part allows the user to enter all the necessary data. In the second part,

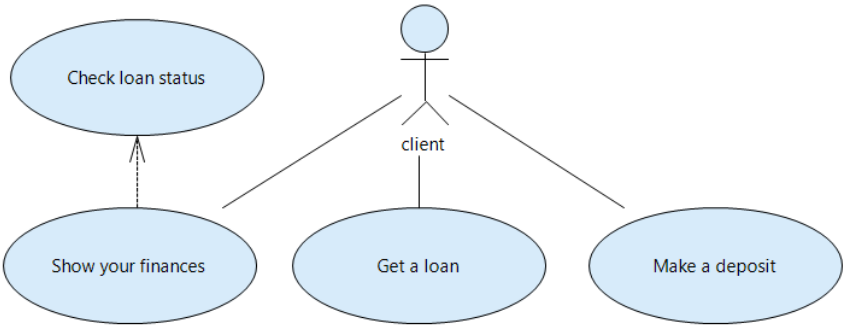


Fig. 25. Banking use case model fragment

the system shows information summarising the deposit for the user to confirm. Note that this gives two occasions to define domain logic. The first of them is to validate the deposit creation data entered by the user (sentence 5). The second one computes data to be shown in the deposit summary window (sentence 6).

Note that the result of the validation action (sentence 5) can lead to two alternate paths that depend on specific conditions ("=>cond" sentences). The first path is expressed with sentences 6-10. The second path is expressed by sentence 5.1.1 and the "rejoin" sentence. This forms application logic that is appropriately generated by the RSL-to-code transformation. The second use case contains simple logic to show loan status. Despite such simple scenario, the domain logic that computes such status can be quite complex.

Figure 27 shows the domain logic for validation, to be used in the first use case. The model contains elements that determine the possibility to create deposits based on given deposit parameters. Deposit parameters are concepts related to "deposit creation data", which is an RSL notion occurring in scenario sentences from Figure 26. Note that we have omitted this RSL model as quite trivial. The formula specifying data correctness is included in the validity condition assigned to the "deposit parameters" concept. It states that the funds on the selected source account should be equal or greater to the declared deposit amount. The model also presents the "source account" pattern-based reference which is used in the above formula. This is accompanied by all the attributes of the involved concepts ("deposit duration", "source account number", "account balance", "account number"). It is worth noting that many of these properties are specialised versions of more general terms used in representing more general dependencies further in the model.

Figure 28 shows three relationships related to various aspects of deposits – mainly general rules for calculating interests. Part a) defines that the "deposit compounding frequency" and the "annual nominal interest rate" depends on the deposit duration. This dependency is represented in the form of a data-based relationship. Part b) defines a formula for computing the final amount of a given deposit or loan which is obtained by using compound interest rules. Part c) defines the relation between the original sum of a given deposit or loan, its final amount and the total interest. The two latter formulas are represented as a pattern-based transitions.

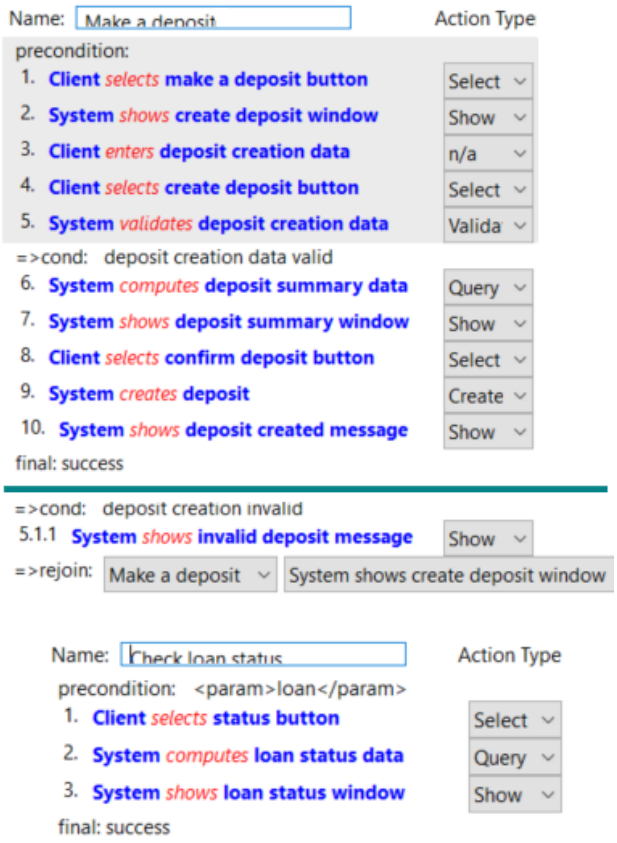


Fig. 26. Example scenarios for the banking case

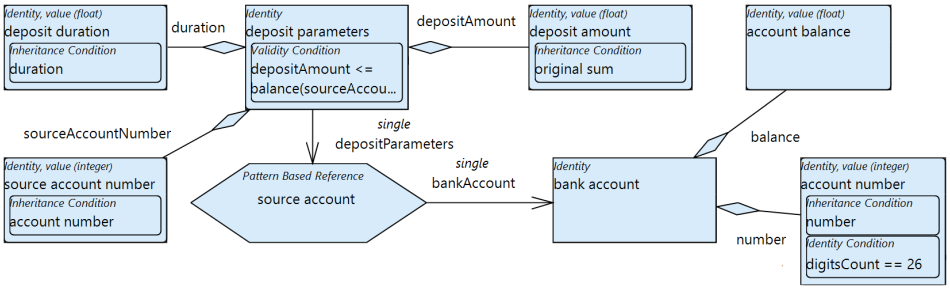


Fig. 27. RSL-DL model defining deposit creation data validation

Figure 29 shows two relationships related to various aspects of loans. Part a) defines the formula for calculating instalments of a loan. Part b) defines the formula for calculating the amount of debt left for a given loan if a certain number of its instalments were paid. Both of these cases are represented in the form of pattern-based transitions.

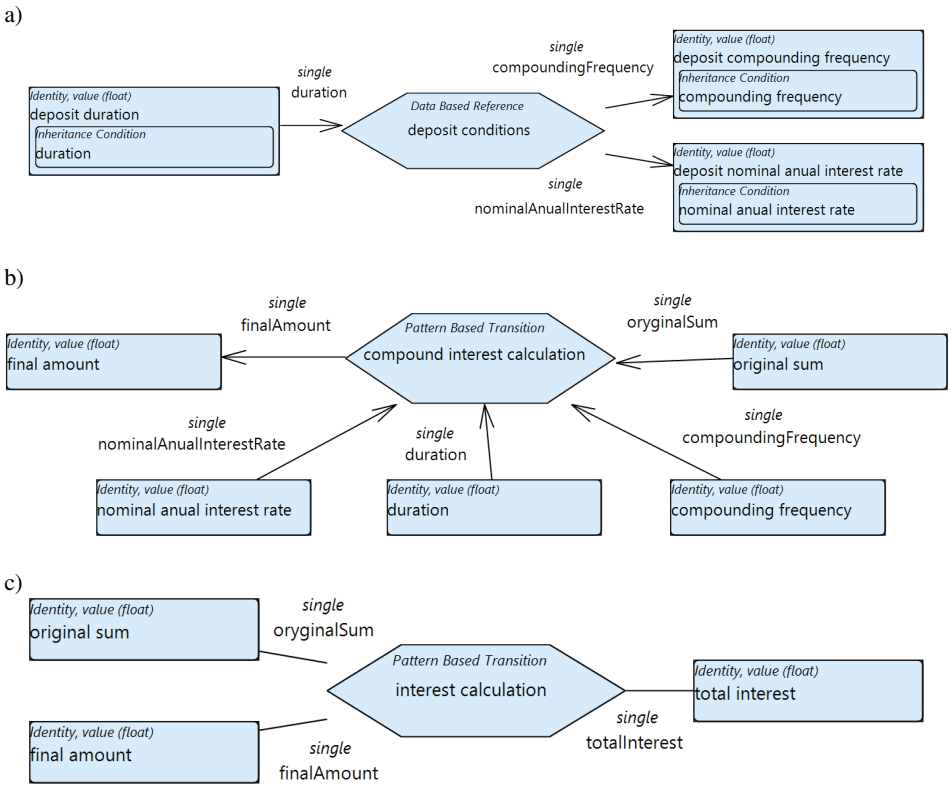


Fig. 28. RSL-DL model defining deposit conditions (a), compound interest calculation rules (b), and simple interest calculation rules (c)

Figure 30 shows code generated on the basis of the model from Figure 27. This code is a rather standard class with its attributes that were generated according to rules no. 1 (generate classes for notions) and no. 4 (generate attributes for attribute links) respectively. Note that the getters, setters, and constructors were omitted for brevity. In addition, we have the validation method, generated according to rule no. 6 (generate validation methods). The method checks validity of deposit parameters according to the validation condition formula from Figure 27. It uses the "MSSourceAccount" class generated from the "source account" reference according to rules no. 9 (generate classes and static methods from relationships) and 14 (generate object filtering code from condition patterns). This way, it is possible to obtain the bank account and its balance required by the formula.

Figure 31 shows code generated from the models in Figure 28. The goal is to obtain the "deposit summary" based on the provided "deposit parameters". Note that the representation of "deposit summary" is not shown in the model for brevity. It is a concept related to the "deposit summary data" RSL notion found in the scenarios. It has five attributes: "deposit amount", "deposit duration", "deposit nominal annual interest rate", "deposit compounding frequency" and "total interest", where the last one is most com-

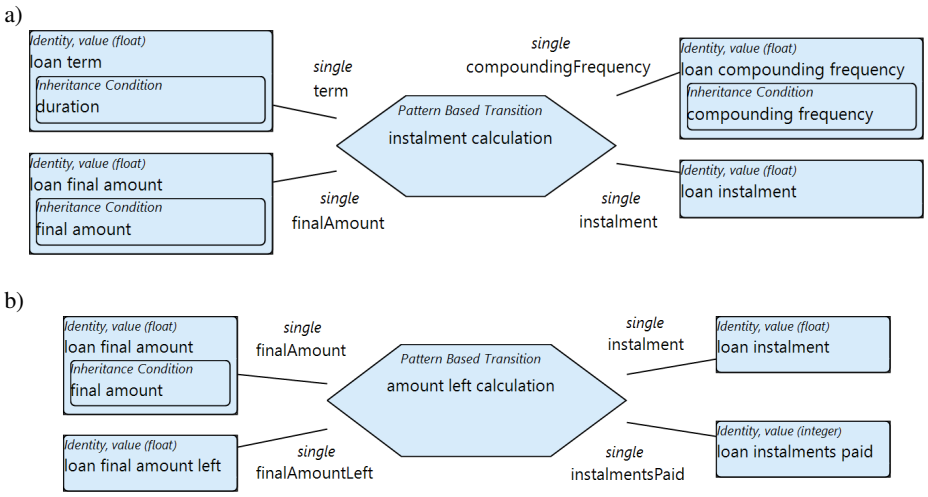


Fig. 29. RSL-DL model defining loan instalment calculation (a), and loan amount left calculation (b)

plex to obtain. This code was generated according to rule no. 11 (generate methods for queries). Its structure corresponds to the structure of the rule tree generated by the inference engine to allow for computing of these values.

The "getDepositSummary" method starts by obtaining the deposit sum and its duration, which are easily available as attributes of the "MDepositParameters" class (cf. Figures 27 and 30). Then, the "nominal annual interest rate" and the "compounding frequency" are obtained using appropriate methods of the "MSDepositCondition" class. These methods were generated according to rule no. 9 (generate classes and static methods from relationships) from the "deposit conditions" data-based reference (see Figure 28a). Finally, appropriate methods of the "MSCompoundInterestCalculation" and the "MSInterestCalculation" class are used to obtain the final amount and the total interest for a

```
public MDepositParameters implements IMDepositParameters {

    private double depositAmount;
    private double depositDuration;
    private int sourceAccountNumber;

    (...)

    public boolean validate(){
        return depositAmount <= MSSourceAccount
            .getBankAccount(this).getBalance();
    }

}
```

Fig. 30. Code for deposit creation data validation

```
class MSDepositSummary {

    public static IMDepositSummary getDepositSummary
            (IMDepositParameters depositParameters){
        double depositAmount = depositParameters.getDepositAmount();
        double depositDuration = depositParameters.getDuration();
        double depositNominalAnualInterestRate =
            MSDepositConditions.getNominalAnualInterestRate(depositDuration);
        double depositCompoundingFrequency =
            MSDepositConditions.getCompoundingFrequency(depositDuration);
        double finalAmount = MSCompoundInterestCalculation.getFinalAmount
            (depositAmount, depositNominalAnualInterestRate,
                    depositCompoundingFrequency, depositDuration);
        double totalInterest = MSInterestCalculation.getTotalInterest
            (finalAmount, depositAmount);
        return new MDepositSummary(depositAmount, depositDuration,
                depositNominalAnualInterestRate, depositCompoundingFrequency,
                totalInterest);
    }

}
```

Fig. 31. Code for deposit summary computation

```
class MSCompoundInterestCalculation {

    public static double getFinalAmount(double originalSum,
            double nominalAnualInterestRate,
            double compoundingFrequency, double duration){
        return originalSum*Math.pow(1+nominalAnualInterestRate/
                compoundingFrequency,
                compoundingFrequency*duration);
    }

}
```

Fig. 32. Code for compound interest calculation

deposit. Both classes are generated according to rules no. 9 and 15 (generate code from transition patterns) from the ”compound interest calculation” and the ”interest calculation” pattern-based transitions (Figure 28b and c). The last step is to return the “deposit summary” object, created from the previously obtained values by using the constructor generated according to rule no. 5 (generate constructors for notion-based classes).

The details of the “MSCompoundInterestCalculation” class used in the above code are shown in Figure 32. The actual code of its only method is derived from the formula “$originalSum * (1 + nominalAnualInterestRate/compoundingFrequency)\hat{}$ $(compoundingFrequency * duration)$” included in the respective pattern-based transition (see Figure 28b).

Figure 33 shows code generated from the model in Figure 29. Its goal is to obtain the “loan status” based on a provided object representing a loan. Note that the the representation of “loan status” is not shown in the model for brevity. It is a concept related to the “loan status data” RSL notion found in the scenarios. It specialises from the “loan” concept and has five additional attributes: “loan nominal annual interest rate”, “loan com-

```
class MSLoanStatus {

    public static IMLoanStatus getLoanStatus(IMLoan loan){
        double originalSum = loan.getLoanAmount();
        double loanDuration = loan.getDuration();
        double loanNominalAnualInterestRate =
            MSLoanConditions.getNominalAnualInterestRate(loanDuration);
        double loanCompoundingFrequency =
            MSLoanConditions.getCompoundingFrequency(loanDuration);
        double finalAmount = MSCompoundInterestCalculation.getFinalAmount
            (originalSum, loanNominalAnualInterestRate,
                    loanCompoundingFrequency, loanDuration);
        double totalInterest = MSInterestCalculation.getTotalInterest
                    (finalAmount, originalSum);
        double loanInstalment = MSInstalmentCalculation.getInstalment
            (finalAmount, loanCompoundingFrequency, loanDuration);
        int loanInstalmentsPaid = loan.getInstalmentsPaid();
        double loanFinalAmountLeft = MSAmountLeftCalculation.getFinalAmountLeft
            (finalAmount, loanInstalments, loanInstalmentsPaid);
        return new MLoanStatus(loan, loanNominalAnualInterestRate,
                loanCompoundingFrequency, loanFinalAmountLeft, loanInstalment,
                totalInterest);
    }

}
```

Fig. 33. Code for loan status computation

pounding frequency", "loan final amount left", "loan instalment" and "total interest". Most of these values overlap with these from the previous code fragment, thus majority of this code is very similar (applied to a loan instead of a deposit). There is however the need to compute two new values – the "loan instalment" and the "loan final amount". Both of these values are obtained by using the class and its methods generated according to rules no. 9 and 15 from the "instalment calculation" and the "amount left calculation" pattern-based transitions (see Figure 29). To obtain the second of these values, the number of loan instalments already paid is also needed, so its value is obtained beforehand from the appropriate attribute of the "MLoan" class.

8. Language validation and discussion

In the previous section, we presented the most interesting parts of code automatically generated from the RSL-DL domain specifications of two cases. This code constituted the back-end of the respective systems and was compatible with the front-end code generated from the RSL specifications. It has to be stressed that the generated code covered complete back-end logic, and almost no manual updates were necessary. The only significant update necessary in the current generator version was adding a database access code. We did not cover this element due to that it was already developed previously, including an appropriate automated transformation for the ReDSeeDS system [68]. All other issues are minor (e.g. handling of import statements) and should be easily solvable through further improvements to the transformation code.

It should be noted that currently, the two languages - RSL and RSL-DL are only partially integrated. In particular, RSL contains its own domain modelling language that is

focused on the front-end perspective. The presented transformations (RSL-to-code and RSL-DL-to-code) are compatible with respect to the structural composition of the generated code. However, these transformations have to be invoked separately and then manually integrated by adding small fragments of interfacing code. To obtain a truly no-code environment, we would first need to fully integrate the domain modelling part of RSL with the syntax and semantics of RSL-DL. Furthermore, we would need to construct a unified transformation that generates fully integrated front-end and back-end code. This should also include the generation of the database structure and data persistency code.

Proper validation of a full no-code system based on RSL and RSL-DL is thus future work. This would involve using prospective ReDSeeDS 2.0 in real-world projects where specifications in RSL+RSL-DL would remove the need to write traditional code completely. However, even partial validation of RSL-DL alone can give interesting insights on the usefulness and improvements that would need to be made to create such a system.

To initially validate RSL-DL, we have used primarily two different approaches. Their aim was to assess certain aspects of the language's usability: understandability and operability. The first approach was to determine language comprehension by its first-time users. It consisted in testing language proficiency, following a brief introduction to the language. The second approach was to determine efficiency of language usage by the users with various experience levels. In both cases, we have used a specially developed RSL-DL editor, used in conjunction with the ReDSeeDS environment. Additionally, we have also conducted a brief complexity analysis of RSL-DL language constructs in comparison to more traditional ones.

8.1. Validation of understandability

The first validation study was conducted with a group of post-graduate computer science students attending the "Model-Driven Software Development" course at the Warsaw University of Technology. The course curriculum included classes on the design and usage of various Domain-Specific Languages. The study was thus well aligned with the aim to acquaint the students with this topic.

The setup of the study was as follows. First, the students attended two lab sessions (four class hours) where they were presented with the RSL and the ReDSeeDS tool. Note that prior to this, the students had no experience with Software Language Engineering but have attended a parallel lecture where they were introduced with the fundamentals of metamodelling. Next, the students were presented with a brief, one-hour introduction to the language. Then, they have spent two hours solving simple exercises using the aforementioned RSL-DL editor. After this, the students were presented with correct solutions to the exercises. Finally, the students were asked to answer 12 questions in an online questionnaire. All of the questions were single-choice, and referred to specific RSL-DL diagrams. Each question had four possible answers. The first eight questions were related to the understanding of the language syntax, the next three related to language usage, and the final one checked more nuanced usage of the language related to its declarative nature. The students were given one class hour (45 minutes) to finish the questionnaire, but most of them have finished in less than 20 minutes.

The results of the study consist in 42 replies to the questionnaire. The average of correct answers in the whole questionnaire was 69%. For syntax understanding (the first eight questions), it was 75%, for usage understanding (the next three questions), it was

60%, and for the last question it was 40%. The above results show that even after a brief introduction, most of issues could be understood correctly, even by non-professionals (students). An important result of this study is the notion of which parts of the questionnaire caused most problems for the participants. Detailed results, together with the question contents are provided in the Supplement.

The relatively low result in the case of the last question can be explained by its advanced nature, going beyond the explanations given to the students. Thus, the 40% can be seen as an unexpectedly good result. It is also worth noting that relatively low percentages of correct answers were associated with questions about differentiation between inferred roles and assigned roles (questions no. 3, 5 and 9). These results were similar to the case of the last question. Further research is needed to determine if that was caused by insufficient explanations (this aspect was under-represented in the exercises) or inherent difficulties caused by the language design. In summary, the overall results of this study indicate that the language is comprehensible even after a very short introduction. However, a more thorough validation with statistical analysis is needed to confirm this, and can be seen as future work.

8.2. Validation of operability

The second validation study was conducted with a group of three software developers with different programming skills. The first person is one of the language authors and thus has very good knowledge of RSL-DL. At the same time, he is an experienced Java programmer. The second person is a Ph.D. student with wide general computer science knowledge and average Java programming experience. The third person is an undergraduate student with more narrow CS knowledge but with relatively high experience in Java programming. The students were not involved in the development of RSL-DL and had no previous knowledge of it.

The study consisted in comparison of coding efficiency and was based on solving specific problems. The setup of the study was as follows. First, the study participants were presented with a 1.5 hour long introduction of RSL-DL and its editor. This included the presentation of three problems: calculation of square mean error (no. 1), calculation of definite integrals (no. 2), and calculation of VAT for product lists (no. 3). The problem formulations involved appropriate formulas and are presented in detail in the Supplement. Next, the participants were supplied with artefacts generated from appropriate RSL specifications (use cases with scenarios) by the ReDSeeDS system. These consisted of pre-initialised RSL-DL models (just the notions) and code skeletons (Data Transfer Objects and method signatures) in Java.

The goal of the participants was to fill-in the provided artefacts to complete domain logic functionality. To prevent from negative bias, the participants were asked to solve the problems using RSL-DL first, and only then to solve them in Java. The participants were also asked to measure time spent on all the tasks. The results of these measurements are given in Table 1.

Comparison of times for the three study participants can be treated as rather anecdotal evidence but they give some insight on the productivity of developing domain logic (back-end) code with RSL-DL. As it can be noticed, productivity of RSL-DL development vs. Java development is significantly higher for an experienced RSL-DL user. Also, a less experienced Java programmer (the Ph.D. student) had certain productivity gains.

Table 1. Results of the operability study

Participant	**Task**	**RSL-DL time** (min.)	**Java time** (min.)
Author	No. 1	3:00	5:00
Ph.D. Student	No. 1	9:00	12:00
Undergrad. Student	No. 1	13:35	4:50
Author	No. 2	1:45	3:40
Ph.D. Student	No. 2	4:00	10:00
Undergrad. Student	No. 2	16:45	5:30
Author	No. 3	4:30	6:00
Ph.D. Student	No. 3	15:00	15:00
Undergrad. Student	No. 3	12:25	6:20

On the other hand, a very experienced Java programmer (the undergraduate student) had performed much better using a traditional programming language. Thus, it can be argued that as general knowledge of developers and their RSL-DL skills raise - productivity gains tend to be significant. It can also be argued that RSL-DL has the potential for extending productivity gains for less experienced programmers. Still, this argumentation has to be acknowledged through a more thorough experimentation with a larger participant scope. This can be seen as future work.

8.3. Analysis of complexity

An important and obvious aspect of RSL-DL design as a low-code language was to reduce complexity of code. To determine characteristics of RSL-DL in this aspect, we need to compare the complexity of RSL-DL models (including model queries) and equivalent 3GL back-end code. We can find several metrics of software (textual code) complexity in literature [27]. Moreover, relatively sparse complexity metrics exist for visual models (mainly formulated in UML) [36]. However, we could not find a generally accepted objective metric for comparing the complexity of textual code and visual models (cf. work by Masmali and Badreddin [37]). Considering this, to assess the potential reduction of complexity, we have applied a simple metric that compares quantities of syntactic constructs.

Our comparative procedure followed the following steps. First, we have manually developed appropriate domain logic code for the specified use cases and scenarios. This involved manual implementation of services, like IService in Figures 5 and 24. Second, after implementing the transformation from RSL-DL to code, we have specified the domain using RSL-DL notation. Finally, after generating the code, we compared it to that developed manually to ensure that they are functionally equivalent.

Quantitative comparison of syntactic constructs shows at least a 30% reduction in complexity for RSL-DL models. As an example, we can consider relevant model and code fragments for computing “weighted average grade” (compare Figures 6, 17 and 22). The model consists of 22 syntactic elements (3 relations, 7 notions, 7 participations, 2 attribute links and 3 formulas) to which we need to add 1 query. The equivalent code consists of 35 syntactic elements (5 classes, 2 fields, 4 methods, 2 constructors, 4 getter methods, 2 loops, and 16 other instructions). It has to be stressed that the above numbers pertain to

manually developed code and not that which was automatically generated. However, differences were minimal and mainly concerned with optimisation (manual code was more optimised and more compact).

The above comparison can be treated as indicative and necessitates more detailed experimentation and analysis, which we see as future work. Regardless of this, significant benefits can also be achieved due to the reusability capabilities of RSL-DL. Knowledge representations expressed in RSL-DL are declarative and independent of any particular context of usage. Thus they can be queried similarly to how ontologies can be queried for inferring various facts. However, RSL-DL queries result not in facts but ready imperative code adapted to the given usage (application logic). Therefore, the only cost of obtaining code adapted to new functional specification (e.g. new use cases and their scenarios) is the cost of formulating new queries.

9. Summary and future work

This work attempted at going beyond typical low-code approaches. We have applied declarative knowledge representations to define non-trivial logic that can be used to produce imperative back-end code in a standard programming language. This generated code can be interfaced with front-end code produced from low-code specifications that use formalised requirements models (use cases, scenarios). It thus can be noted that our approach leads to shifting most of the programming activities to the level typically used in requirements engineering. Programming becomes equivalent to specifying precise requirements models that define various aspects of the system and its problem domain. An RSL-DL conceptual model can be treated – in fact – as a high-level program that can be executed immediately after compiling it into eg. Java and then – executable code. Moreover, RSL-DL models can be seen as "ontologies as code" [35] and a step towards a "fifth generation language" as postulated by Thalheim and Jaakkola [60].

We see two main areas where our approach can benefit software development: reduction of complexity and increased reuse. The first area is in line with the general goals of the low-code movement – to offer means for reducing accidental (technological) complexity in favour of concentrating on the essential (e.g. domain) complexity (see early insights on this by Brooks [10]). A thorough comparison of complexity between RSL-DL and traditional programming languages, and analysis of reusability can be seen as interesting areas of future work.

Another area for future work is the analysis of RSL-DL usability as a low-code language. Generally, it can be expected that better usability is assured through declarative characteristics of the language (see appropriate comparative analyses [24, 65]). This is in line with our initial studies presented in the previous section. However, to fully support this claim, more extensive experimentation should be conducted.

Other areas which we plan to investigate in the future include better integration with requirements processing mechanisms. One aspect of this is the application of natural language processing. Here, valuable insights can be drawn from the concept of naturalistic programming [46]. This concept postulates the use of natural language elements to design programming languages that are more expressive from the programmers' point of view. Another interesting approach in this area is that of Mefteh et al. [38]. In this approach, natural language scenarios are transformed into constrained language models expressed in

RSL. On the other hand, we also plan to integrate our approach with existing approaches to generate CRUD operations and database schemas directly from requirements models expressed in RSL [68].

Finally, we would like to address the fact that both RSL and RSL-DL were created as distinct new languages. Still, they follow typical notations found in requirements engineering, domain modelling and knowledge representation. It thus can be argued that an existing general-purpose language (e.g. a UML profile) could be used to express the same semantics. Both approaches – creating a new language vs. using a general-purpose language – can be seen as equivalent. However, creating a language from scratch allowed us to go beyond the "beaten path" and create visual syntax that would be hard to express using standard notations.

References

1. The MOLA Language Reference Manual Version 2.0 final (2007)
2. Aßmann, U., Ebert, J., Pan, J.Z., Staab, S., Zhao, Y.: Ontology-Driven Software Development. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
3. Aßmann, U., Zschaler, S., Wagner, G.: Ontologies, Meta-models, and the Model-Driven Paradigm, pp. 249–273. Springer (2006)
4. Atkinson, C., Gutheil, M., Kiko, K.: On the relationship of ontologies and models. In: Proc. 2nd Workshop on Meta-Modelling, WoMM 2006. pp. 47–60 (2006)
5. Basso, F.P., Pillat, R.M., Oliveira, T.C., Roos-Frantz, F., Frantz, R.Z.: Automated design of multi-layered web information systems. Journal of Systems and Software 117, 612–637 (2016)
6. van den Berg, K.G., Simons, A.J.H.: Control flow semantics of use cases in UML. Information and Software Technology 41(10), 651–659 (1999)
7. Bexiga, M., Garbatov, S., Seco, J.a.C.: Closing the gap between designers and developers in a low code ecosystem. In: 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings. pp. 1–10. MODELS '20, Association for Computing Machinery, New York, NY, USA (2020)
8. Brambilla, M., Fraternali, P.: Interaction flow modeling language: Model-driven UI engineering of web and mobile apps with IFML. Morgan Kaufmann (2014)
9. Brito, I.S., Barros, J.P., Gomes, L.: From requirements to code (Re2Code) – a model-based approach for controller implementation. In: 2016 IEEE 14th International Conference on Industrial Informatics (INDIN). pp. 1224–1230 (2016)
10. Brooks, F.P.: No silver bullet: Essence and accidents of software engineering. IEEE Computer 20(4), 10–19 (April 1987)
11. Browne, P.: JBoss Drools business rules. Packt Publishing Ltd (2009)
12. Cabot, J.: Positioning of the low-code movement within the field of model-driven engineering. In: Proceedings of the 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings. MODELS '20 (2020)
13. Chisholm, M.: How to build a business rules engine: extending application functionality through metadata engineering. Morgan Kaufmann (2004)
14. Fatolahi, A., Some, S.S., et al.: Assessing a model-driven web-application engineering approach. Journal of Software Engineering and Applications 7(05), 10p (2014)
15. Gašević, D., Djurić, D., Devedžić, V.: Model Driven Engineering and Ontology Development. Springer (2009)
16. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition 5(2), 199–220 (1993)
17. Haav, H.M.: A comparative study of approaches of ontology driven software development. Informatica 29(3), 439–466 (2018)

18. Haav, H.M., Ojamaa, A.: Semi-automated integration of domain ontologies to DSL metamodels. International Journal of Intelligent Information and Database Systems 10(1/2), 94–116 (2017)
19. Henderson-Sellers, B.: Bridging metamodels and ontologies in software engineering. Journal of Systems and Software 84(2), 301–313 (2011)
20. Henkel, M., Stirna, J.: Pondering on the key functionality of model driven development tools: The case of Mendix. In: International Conference on Business Informatics Research. pp. 146–160. Springer (2010)
21. Henriques, H., Lourenço, H., Amaral, V., Goulão, M.: Improving the developer experience with a low-code process modelling language. In: 21th ACM/IEEE International Conference on Model Driven Engineering Languages and Systems. pp. 200–210 (2018)
22. Hinchey, M.G., Rash, J.L., Rouff, C.A.: Requirements to design to code: Towards a fully formal approach to automatic code generation. Tech. rep., NASA (2005)
23. Jacobson, I., Christerson, M., Jonsson, P., Övergaard, G.: Object-Oriented Software Engineering: A Use Case Driven Approach. Addison-Wesley (1992)
24. Jbara, A., Bibliowicz, A., Wengrowicz, N., Levi, N., Dori, D.: Toward integrating systems engineering with software engineering through object-process programming. International Journal of Information Technology pp. 1–35 (2020)
25. Kaindl, H., Smialek, M., Wagner, P., et al.: Requirements specification language definition. Project Deliverable D2.4.2, ReDSeeDS Project (2009)
26. Kalnins, A., Barzdins, J., Celms, E.: Model transformation language MOLA. Lecture Notes in Computer Science 3599, 62–76 (2004)
27. Khan, A.A., Mahmood, A., Amralla, S.M., Mirza, T.H.: Comparison of software complexity metrics. International Journal of Computing and Network Technology 4(1), 19–26 (2016)
28. Koch, N., Kozuruba, S.: Requirements models as first class entities in model-driven web engineering. In: Current Trends in Web Engineering. pp. 158–169. Springer (2012)
29. Koch, N., Kraus, A.: Towards a common metamodel for the development of web applications. In: Web Engineering. pp. 497–506. Springer (2003)
30. Koch, N., Meliá-Beigbeder, S., Moreno-Vergara, N., Pelechano-Ferragud, V., Sánchez-Figueroa, F., Vara-Mesa, J.: Model-driven web engineering. Upgrade-Novática Journal (English and Spanish) 2, 40–45 (2008)
31. Kramer, A.: Symja library-Java symbolic math system (2010), https://github.com/axkr/symja_android_library, last accessed June 2023
32. Krouwel, M.R., Land, M.O.t., Proper, H.A.: Generating low-code applications from enterprise ontology. In: Barn, B.S., Sandkuhl, K. (eds.) The Practice of Enterprise Modeling. pp. 18–32. Springer International Publishing, Cham (2022)
33. Martínez, Y., Cachero, C., Meliá, S.: Evaluating the impact of a model-driven web engineering approach on the productivity and the satisfaction of software development teams. In: Web Engineering. pp. 223–237. Springer (2012)
34. Martínez, Y., Cachero, C., Meliá, S.: Empirical study on the maintainability of web applications: Model-driven engineering vs code-centric. Empirical Software Engineering 19(6), 1887–1920 (2014)
35. Martins, B.F.: The OntoOO-method: An ontology-driven conceptual modeling approach for evolving the oo-method. In: Advances in Conceptual Modeling. pp. 247–254 (2019)
36. Masmali, O., Badreddin, O.: Comprehensive model-driven complexity metrics for software systems. In: 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C). pp. 674–675 (2020)
37. Masmali, O., Badreddin, O.: Towards a model-based fuzzy software quality metrics. In: MODELSWARD. pp. 139–148 (2020)
38. Mefteh, M., Bouassida, N., Ben-Abdallah, H.: Towards naturalistic programming: Mapping language-independent requirements to constrained language specifications. Science of Computer Programming 166, 89–119 (2018)

39. Nowakowski, W., Śmiałek, M., Ambroziewicz, A., Straszak, T.: Requirements-level language and tools for capturing software system essence. Computer Science and Information Systems 10(4), 1499–1524 (2013)
40. Object Management Group: Semantics of Business Vocabulary and Business Rules, version 1.4, formal/2017-05-05 (2017)
41. Object Management Group: Meta Object Facility (MOF) Core Specification, version 2.5.1, formal/2019-10-01 (2019)
42. Pahl, C.: Semantic model-driven architecting of service-based software systems. Information and Software Technology 49(8), 838–850 (2007)
43. Papotti, P.E., Do Prado, A.F., de Souza, W.L., Cirilo, C.E., Pires, L.F.: A quantitative analysis of model-driven code generation through software experimentation. In: International Conference on Advanced Information Systems Engineering. pp. 321–337 (2013)
44. Parreiras, F.S., Staab, S., Schenk, S., Winter, A.: Model driven specification of ontology translations. In: ER'08. Lecture Notes in Computer Science, vol. 5231, pp. 484–497 (2008)
45. Potel, M.: MVP: Model-View-Presenter the Taligent programming model for C++ and Java. Tech. rep., Taligent Inc. (1996)
46. Pulido-Prieto, O., Juárez-Martínez, U.: A survey of naturalistic programming technologies. ACM Comput. Surv. 50(5), 35p (2017)
47. Richardson, C., Rymer, J.R., Mines, C., Cullen, A., Whittaker, D.: New development platforms emerge for customer-facing applications (2014), Forrester report
48. Richardson, C., Rymer, J.R., Mines, C., Cullen, A., Whittaker, D.: Vendor landscape: The fractured, fertile terrainof low-code application platforms (2016), Forrester report
49. Rodriguez-Echeverria, R., Preciado, J.C., Rubio-Largo, A., Conejero, J.M., Prieto, A.E.: A pattern-based development approach for Interaction Flow Modeling Language. Scientific Programming 2019, 1–15 (apr 2019)
50. Ruscio, D.D., Kolovos, D., de Lara, J., Pierantonio, A., Tisi, M., Wimmer, M.: Low-code development and model-driven engineering: Two sides of the same coin? Software and Systems Modeling pp. 1–10 (jan 2022)
51. Rybiński, K., Śmiałek, M.: Beyond low-code development: Marrying requirements models and knowledge representations. In: 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS). pp. 919–928 (2022)
52. Sahay, A., Indamutsa, A., Di Ruscio, D., Pierantonio, A.: Supporting the understanding and comparison of low-code development platforms. In: 46th Euromicro Conference on Software Engineering and Advanced Applications. pp. 171–178 (2020)
53. Slimani, T.: Ontology development: A comparing study on tools, languages and formalisms. Indian Journal of Science and Technology 8(24), 1–12 (2015)
54. Slonneger, K., Kurtz, B.L.: Formal Syntax and Semantics of Programming Languages. Addison-Wesley (1995)
55. Śmiałek, M., Jarzebowski, N., Nowakowski, W.: Runtime semantics of use case stories. In: Visual Languages and Human-Centric Computing (VL/HCC), 2012 IEEE Symposium on. pp. 159–162. IEEE (2012)
56. Śmiałek, M., Nowakowski, W.: From Requirements to Java in a Snap: Model-Driven Requirements Engineering in Practice. Springer (2015)
57. Smialek, M., Straszak, T.: Facilitating transition from requirements to code with the ReDSeeDS tool. In: Requirements Engineering Conference (RE), 2012 20th IEEE International. pp. 321–322 (2012)
58. Soylu, A., De Causmaecker, P.: Merging model driven and ontology driven system development approaches: Pervasive computing perspective. In: 24th International Symposium on Computer and Information Sciences. pp. 730–735 (2009)
59. Stevenson, G., Dobson, S.: Sapphire: Generating java runtime artefacts from OWL ontologies. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp. 425–436. Springer (2011)

60. Thalheim, B., Jaakkola, H.: Model-based fifth generation programming. In: Information Modelling and Knowledge Bases XXXI, pp. 381–400. IOS Press (2020)
61. Trigo, A., Varajão, J., Almeida, M.: Low-code versus code-based software development: Which wins the productivity game? IT Professional 24(5), 61–68 (2022)
62. Valderas, P., Pelechano, V.: A survey of requirements specification in model-driven development of web applications. ACM Trans. Web 5(2), 1–51 (2011)
63. Vincent, P., Iijima, K., Driver, M., Wong, J., Natis, Y.: Magic quadrant for enterprise low-code application platforms. Gartner report (2019)
64. Völkel, M., Sure, Y.: RDFReactor-from ontologies to programmatic data access. In: Poster Proceedings of the Fourth International Semantic Web Conference (2005)
65. Vuorimaa, P., Laine, M., Litvinova, E., Shestakov, D.: Leveraging declarative languages in web application development. World Wide Web 19(4), 519–543 (2016)
66. Wakil, K., Jawawi, D.N.: Model driven web engineering: A systematic mapping study. e-Informatica Software Engineering Journal 9(1), 107–142 (2015)
67. Woo, M.: The rise of no/low code software development—no experience needed? Engineering 6(9), 960–961 (2020)
68. Yamouni Khelifi, N., Śmiałek, M., Mekki, R.: Generating database access code from domain models. In: 2015 Federated Conference on Computer Science and Information Systems. pp. 991–996 (2015)

Kamil Rybiński is Adjunct Professor at the Institute of the Theory of Electrical Engineering and Applied Informatics of the Warsaw University of Technology. He obtained his Ph.D. in software engineering from the Faculty of Electrical Engineering at the same university. His research interests includes requirements engineering, model-driven software development and knowledge representation.

Michał Śmiałek is Professor of Software Engineering at the Warsaw University of Technology. He obtained his habilitation (higher doctorate) degree in informatics from the Warsaw Military University and graduated from the Warsaw University of Technology (M.Sc. and Ph.D.) and the University of Sheffield (M.Sc.). Since 1991 he has worked in the industry as Software Developer, Project Manager and Professional Coach. His current research interests include model-driven software development, requirements engineering, software reuse, software language engineering and large-scale computing. He published and edited several books and over 100 papers in various journals and conference proceedings. He coordinated two European-level research projects, chaired several international conferences and reviewed for major computer science journals.

Received: July 15, 2023; Accepted: May 10, 2024.

Computer Science and Information Systems 21(3):725–757 https://doi.org/10.2298/CSIS230328013G

A Practical and UC-Secure Decentralized Key Management and Authentication Scheme Based on Blockchain for VNDN

Xian Guo*, Sarah Almadhehagi, Tao Feng, Di Zhang, Yongbo Jiang, and Junli Fang

School of Computer and Communication, Lanzhou University of Technology
730050 Lanzhou, China
{Xian Guo}iamxg@163.com

Abstract. In Vehicular Named Data Networking (VNDN), most of the existing key management mechanisms still rely on the hierarchical key trust model or the Public Key Infrastructure (PKI), in which the centralized certificate authority is used as a Trusted Third Party (TTP) to provide a signature for the user's public key. Thus, the TTP becomes vulnerable to attacks and maybe there exists a singlepoint failure problem. In addition, the in-network caching in the VNDN may incur a threat to the system and make it is vulnerable to the DoS attack caused by Interest flooding aimed at the Content-Store. To tackle these security issues, we firstly propose an efficient decentralized key management solution based on blockchain for VNDN. Secondly, based on the proposed key management scheme, a lightweight mutual authentication scheme and a key agreement protocol for V2X are respectively proposed in this paper. Finally, we analyze the security attributes of our solutions in the Universally Composable (UC) framework. Our analysis results show that our novel schemes can meet the security requirements of our solutions. In addition, our experimental results show our new schemes have higher efficient, lower computation and communication costs than other related schemes.

Keywords: Vehicular Named Data Networking (VNDN), Key management, Blockchain, Authentication, Key agreement, Universally Composable (UC).

1. Introduction

The implementation of a new network architecture with an appropriate security mechanism should strive to minimize the cost of computing, storage, and transmission and while to ensure the confidentiality, integrity, and authenticity of a transmitted message. The present Internet of Vehicles (IoV) [1] uses an IP-based communication pattern for this information delivery method, where messages are shared with other vehicles that have IP addresses. Most of IoV applications are point-to-multipoint (P2M) in design, where communications produced by one vehicle are broadcasted and shared with other vehicles nearby. Moreover, because IP-based communication is more vulnerable to threat and data loss, it violates the security requirements of IoV. Thus, the current network architecture based on IP network cannot fully adapt to the dynamic topology environment of the vehicular networks. An alternative communication architecture that suits IoV applications and services must consider the requirements such as security and Quality of

* Corresponding author

Service. The design principle of the Information-Centric Networking (ICN), e.g. Named Data Networking (NDN) [2] differs significantly from the traditional IP network. The NDN focuses on the transmitted content itself instead of the location of the content generator. Therefore, the integrating of vehicular networks with the NDN architecture, which is often called Vehicular Named Data Networking (VNDN) [3], is considered a hot research area. Due to the sensitivity of the data shared between the vehicles, the NDN fully complies with the nature of IoV where transmitted data is of more important than the location of the data generator. The VNDN enables a vehicle to access critical information within the network. Based on this critical information obtained from a network, vehicles can take correct decisions in time to ensure transportation security and safety.

In VNDN, the authenticity and integrity of the information spread are important issues. In order to guarantee the integrity and authenticity of the data transmitted on a network, the content producer will sign the content carried in a data packet using its private key, which securely and effectively links a name to the data. A consumer (or router) can then verify the signature using the public key of the producer and assert the data's provenance in this way. The customer now trusts the integrity and authenticity of the data and doesn't need to worry about where or how it is acquired. In addition, this scheme promotes fine-grained trust and enables a consumer to verify if a public key owner is a reliable producer of a specific piece of data in a certain situation. Therefore, the secret key of a content producer is the same critical material as in traditional IP network. Thus, it is very important to securely manage these cryptographic materials so as to avoid attacks to the transmitted data. Most of the existing NDN key managements use a hierarchical key trust model [4], in which the root key is used as a trust anchor to provide a key signature for the user's public key. To verify the authenticity of the public key, we can retrieve the key chain by using the key's name. This method can avoid generation of a false message. However, in the application of VANET, most of the existing key management schemes are based on traditional Public Key Infrastructure (PKI) [5], which adopts the centralized management model, the root key may be vulnerable to attacks and a tampered threat. So, a scheme based on PKI can lead to a single point of failure. In addition, PKIs such as X.509 PKIX [6] and Web of Trust [7], are designed to certify that a public key indeed belongs to a user with a name and the principle behind the name, and depend on Certificate Authorities (CAs). Due to the unconstrained privileges of CAs, they become central points of failure of the entire network. If a CA is compromised, the attacker can bind a name to an unauthorized public key and produce false data, which may result in severe security problems. Therefore, the key management in NDN is still an important security problem.

In recent years, using the distributed, tamper-proof, traceable and publicly verifiable characteristics of blockchain [8]to explore blockchain-based decentralization solutions has become a research hotspot in various fields [9,10,11,12]. The blockchain make use of distributed storage, consensus mechanisms, smart contracts, and cryptography [13]. So, the communication quality of messages and the pace of convergence are crucial challenges in peer-to-peer (P2P) networks. The blockchain technology over IP still has some significant issues, such as a lack of hierarchical access efficiency. These issues can be effectively resolved by an adoption of blockchain technology over NDN, which provides a decentralized system and streamlines the design. The integration [14,15] of the NDN network with the blockchain technology improves the actual security of the stored infor-

mation and the forwarding process, and further enhances the overall performance of the network, and maximizes forwarding effectiveness. So, the combining of blockchain with VNDN [16] can address some security challenges in the VNDN system, which can clearly improve the efficiency and security of VNDN. The blockchain-based VNDN facilitates trust establishment between different entities in VNDN, and can ensure the transmission of the interest packet and the data packet in a secure way, and prevents attackers from leveraging the built-in design features, such as the utilization of the distributed and uncontrolled in-network caching, to launch attacks at the cache store, and from supplying illegitimate or unrestricted interests to launch an Interest flooding attack.

To solve the security issues of the traditional key managements, some schemes depend on the features of the blockchain technology are proposed to manage the user's keys [17,18,19]. The scheme in [17] introduces a key management mechanism for VNDN. The proposed scheme is used to solve the mutual trust establishment issue between different entities. However, this scheme does not support key agreement suited to fast and dynamic applications in VNDN. In addition, the scheme in [18] proposes a key management based on blockchain for NDN to tackle the issue of reduction of the mutual trust between entities, but this scheme does not support key registration and key update. As for the scheme [19], it proposes a decentralized public key management for NDN based on blockchain. However, this scheme does not support efficiently key management in terms of updating and revoking of the public key. In addition, conducting security evaluations based on the informal analysis method is a common method in the existing studies [17,18,19]. The informal security analysis cannot accurately reflect security concerns. Actually, using of the blockchain technology has potential advantages and defects. On the one hand, a trust chain for a public key can be built using the tamper-proof attribute of blockchain. However, this attribute poses certain challenges for key management, including key update and revocation.

The Universal Composable (UC) framework [20,21] allows modular design and analysis of complex cryptography protocols and ensures security when any multiple protocol instances are executed concurrently. It has become the theoretical basis and methodological guidance for the design and analysis of various composable protocols. Ran Canetti et al. [20,21] propose the concept of the 'Ideal Functionality' that used to capture the security properties of basic cryptographic primitives and propose the security evaluation methods in the UC framework. These ideal functionalities are widely used as standard ideal functionalities to modularly design and analyze various composable protocols. The UC framework is suitable to the design and analysis of the complex system combined the VNDN and the blockchain.

Given the limitations of the above existing schemes of key management, we explore the possibility of implementing a decentralized key management mechanism for VNDN by combining the blockchain technology with the NDN in this paper. In addition, we present a lightweight mutual authentication and key agreement solution based on the proposed key management mechanism to resolve privacy and security issues in the system. To tackle the above shortcomings of the blockchain-based key managements, smart contracts [22] are used to automatically manage the user's keys and implement the various functions of registering, updating, and revoking for the user's keys. Finally, we analyze our novel schemes in UC framework.

The remainder of the paper is organized as follows: We discuss the related works in Section 2.The preliminary knowledge adopted in this study is described in Section 3. Our system model is described in Section 4. In Section 5, we detail our novel solution. In Section 6, we analyze the scheme's security requirements by using the UC framework. In Section 7, we evaluate the scheme's performance. In Section 8, we conclude our works in this paper.

2. Related Work

The centralized management model was used by many existing key management schemes based on traditional PKI [5]. Consequently, PKI-based schemes can lead to a single point of failure. Moreover, PKIs, such as X.509 [6] and web of trust [7], rely on certificate authorities (CAs) to prove that a public key belongs exactly to a name. Given that CA privileges are unaffected, they become the central point of network failure. An attacker can bind a key name to an unauthorized public key if a CA is compromised, causing serious security issues.

In 2021, Hao Liu et al. proposed blockchain-based key management and green routing scheme for VNDN [17]. A blockchain-based key management scheme was presented to solve the mutual trust problem between domain nodes. The number of signature verifications is reduced using the scheme. Moreover, the process of key acquisition and verification is accelerated, and the NDN is suitable for IoV.

A blockchain-based key management scheme in NDN was proposed by Junjun Lou et al. [18] to address the lack of mutual trust between sites without trust anchors. The NDN public key content objects and the scheme for storing, verifying, and revocation are redesigned.

In 2018, Kan Yang et al. proposed BC-PKM [19] for NDN. It utilizes the decentralized and tamper-proof design of blockchains to register, query, update, validate, and revoke the public keys of important principals. This system takes advantage of the decentralized and tamper-proof nature of blockchains. Adversaries that compromise less than half of the public key miners can resist various attacks by using this scheme. The BC-PKM scheme was proposed to address the compromised CA problem.

In 2021, Anhao Xiang et al. [23] proposed a lightweight anonymous device authentication scheme. This scheme is based on NDN and is a representative implementation of ICN. It provides security features, such as mutual authentication, session key agreement, defense against cyberattacks, anonymity, and resilience against device capture attacks. The results of security analysis and performance evaluation indicated that the proposed scheme has lower computational and communication overheads than other state-of-the-art schemes. Compagno et al. [24] proposed an authentication protocol for the ICN network. The proposed OnboardICNg is a symmetric key cryptographic protocol. Given its design, it can reduce the number of packets forwarded in the network.

For NDN-VANET, Xian Guo et al. [25] proposed a receiver-forwarding decision scheme based on Bayesian to solve the broadcast storm problem caused by blind flooding of interest packets. A received interest packet can be adaptively forwarded by their solution. Experimental simulations demonstrate that the BRFD algorithm greatly reduces the redundancy of interest packets.

In 2023, an intelligent forwarding solution with privacy awareness PABRFD for NDN-VANET was proposed by Xian Guo et al. [26] by including homomorphic encryption (HE) into the enhanced BRFD. In PABRFD, the security and privacy concerns of information transferred among car nodes are addressed using a secure Bayesian classifier. They explicitly demonstrate that the new approach can meet security demands, and they put their solution into practice using the HE standard libraries CKKS and BFV. The experimental findings demonstrate that PABRFD can meet their anticipated performance needs.

Table 1. Notations of the paper

Notation	Description
/ndn/VNDN/V	The name prefix of the vehicle's nodes
/ndn/VNDN/RSU/node	The name prefix of the blockchain nodes
HF	The hash function
Sig	The signature algorithm
sK	The secret value
T_Z	The timestamp of the entity Z
K_Z	The authentication key value of the entity Z
PbK	The public key
PrK	The private key
enc	The public encryption algorithm
$Senc$	The symmetric encryption algorithm
SEK_{X-Y}	The session key between entities X and Y
M_r	The request message
M_s	The response message
SEK'_Z	The session key parameter of the entity Z

3. Preliminaries

3.1. Bivariate Polynomial Theory

Blundo et al. [27] proposed a bivariate polynomial in a key distribution scheme. This scheme uses the symmetry of the polynomial to establish the key and ensure communication security.In Eq. (1), a polynomial with two variables (x, y) and degree n has the properties of a bivariate polynomial:

$$f(x,y) = \sum_{i,j=0}^{n} a_{ij} x^i y^j \,. \tag{1}$$

A unique identifier is assigned to each node of the network. For example, node b has a polynomial segment $g_b\,(y)$, which can be obtained by the following function:

$$G_b\,(y) = f\,(b,y)\,. \tag{2}$$

Node b stores n coefficient $g_j\ (0 \leq j \leq n)$:

$$g_j = \sum_{i=0}^{n} a_{ij} b^i \ (0 \leq j \leq n) \,. \tag{3}$$

where b is the ID of the node, and g_i is the coefficient of y^j in the polynomial $f\,(b, y)$. Nodes b and p need to establish a symmetric key between each other by exchanging node IDs first. Then, let $y = b$; node p calculates $f\,(b, y)$. Let $y = p$; node b calculates $f\,(b, y)$. Given that $f\,(b, p) = f\,(p, b)$, nodes b and p compute the same value that can serve as the session key between them.

3.2. NDN Architecture

The rise of NDN [28] has attracted the attention of researchers. Using caching technology to improve the cache utilization in the network by caching data packets, simplifying the coordination between nodes, and reducing the delay for users to obtain data have become urgent concerns. NDN relies on the content of the data instead of the IP address. The data are retrieved by using the name of the data. In NDN, network communication is driven by receivers (that is, data consumers), mainly involving two kinds of packets: an interest packet and a data packet. The interest packet is a data packet sent when an NDN user initiates a request, whereas a data packet is a response from a network service provider to a user's request to satisfy a request. An NDN nod maintains the three data structures Forwarding Information Base (FIB), Pending Interest Table (PIT), and Content Store (CS). In case something goes wrong, it's a good idea to have a backup strategy. Interest packet outgoing interfaces are stored in a forwarding table known as FIB. The arrival interfaces of pending interest packets are stored in the PIT table, which also functions as a forwarding table for data packets.

3.3. Blockchain Technology

A blockchain [29] is a distributed public digital ledger that records transactions in distributed decentralized networks. It uses a peer-to-peer (P2P) network. The blockchain data structure is interpreted as a chain record of transaction blocks. It is organized and can be kept as a ledger or in an ordinary database. Each block is determined by a hash created using the SHA256 cryptographic hash algorithm on the block header.

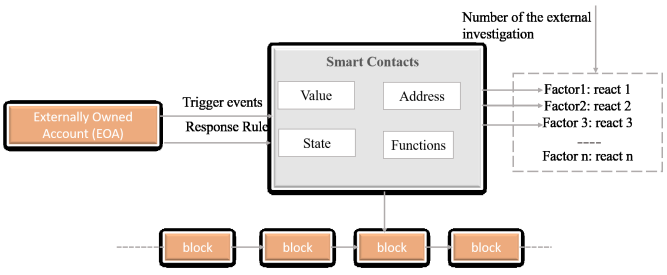


Fig. 1. The model of smart contracts

3.4. Smart Contracts

A public blockchain platform called Ethereum [22] enables anyone to create smart contracts and tokens. On Ethereum, decentralized virtual machines handle P2P agreements using their currencies. A Turing-complete scripting language is also available on Ethereum's programmable blockchain. The smart contract is triggered and executed when used. The model of smart contracts is shown in Figure 1.

3.5. Blockchain Over NDN

A few research projects have been created recently to incorporate the blockchain system into the NDN architecture, including [14,15]. These preliminary investigations mainly focus on how to develop forwarding and synchronization techniques for enabling all of the fundamental and essential functions of NDN as well as how to improve connections between NDN and blockchain.

The blockchain over NDN system is depicted in Figure 2. This system consists of users, NDN routers, and miners who can either be producers or consumers when constructing blockchain blocks depending on whether they are receiving blocks or transactions in real time. Also, although typically not in real-time, the user can assume the roles of a producer when generating a transaction or a consumer while receiving a block.

When a transaction is generated, the user assumes the role of the consumer and sends the transaction to the nearby router by encasing it in an interest packet. Each router forwards this interest packet to all of the outbound interfaces as soon as it receives it. When the interest reaches the miners, they can parse it to retrieve the transaction. Miners won't send any responses, and after the timeout, each router's PIT entries will be deleted.

3.6. UC Framework

The UC framework [20,21] is a paradigm for analyzing and guaranteeing the security of cryptographically composable protocols. The structure of the UC framework is formulated with some entities, such as an environment machine that interacts with the execution protocol and adversaries.The notion of this framework is based on emulation. For any real-world protocol π and an adversary A, an ideal protocol ϕ and an adversary S exist. In the UC framework, a protocol π performs UC simulation on protocol ϕ for an ideal functionality F if an ideal adversary S on the network exists for any possible adversary A on the network. In particular, for any environment ε, the probability that ε can tell the difference between the execution of the protocolπ with the adversary A and the execution of the protocol ϕ with the adversary S is neglible at most. However, the UC simulation ensures the strong correspondence between protocols.

The ideal functionality F is considered the most important part of the UC framework. It is a trusted party for achieving the security requirements of cryptographic protocols. It guarantees the secrecy, authenticity, or delivery of the communicated information. Most of the ideal functionalities have already been formulated, such as the signature functionality F_{SIG}, the public key encryption functionality F_{PKE}, the secure message transmission functionality F_{SMT}, the key exchange functionality F_{KE}, and the message authentication functionality F_{AUTH}.

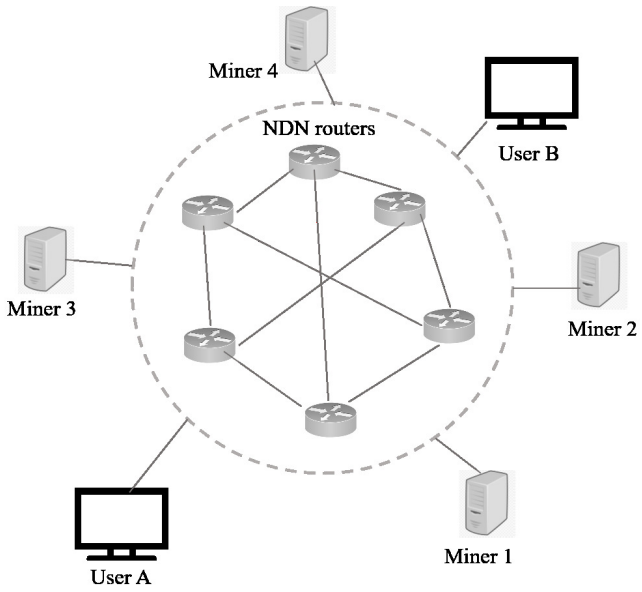


Fig. 2. Blockchain over NDN

3.7. Notations

The notations in the paper are described in Table 1.

4. System Model

4.1. System model

Our system divides into three levels: the first level includes vehicles, the second levle includes Vehicle Trusted Authority (VTA), the third level includes NDN–blockchain network (NDN-BCN). The system model is shown in Figure 3.

1) Vehicles: Vehicles are the main components in our considered VNDN network. Each vehicle is equipped with an onboard unit (OBU), which supports wireless communication for V2X. Every OBU has a tamper-proof device (TPD) to ensure that no private information is exposed or leaked. A hardware security module (HSM) is also installed in RSU's agent. Thus, a vehicle can securely store cryptographic materials. In our scheme, each vehicle can be a content consumer or a producer by sending a generated interest packet or a data packet of the owned content.
2) VTA: In our scheme, the VTA acts as a trusted authority. It is responsible for registering all the information of vehicles in advance, deploying a BCN and smart contract, initializing some system cryptography parameters, generating a transaction, and sending the transaction to the smart contract to register, update, and revoke vehicle information.

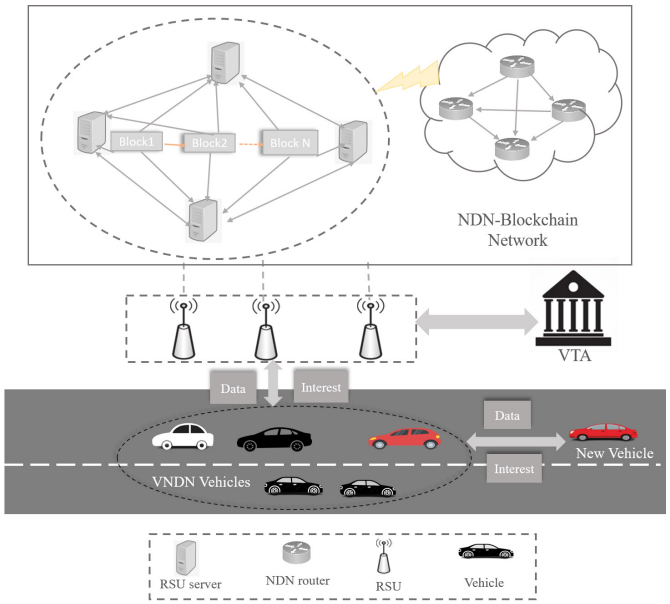


Fig. 3. System model

3) NDN-BCN: The blockchain in our scheme is implemented with the NDN protocol. NDN-BCN is a P2P network consisting of an RSU set, SRSU. Its function is to maintain and manage a public ledger for storing public keys and the identifier information of vehicles without the need for central PKI. Each RSU node plays a router role with the name prefix ndn/VNDN/RSU_i ($RSU_i \in SRSU$).

4.2. Threat Model

In our threat model, the attacks can be divided into two types: passive and active. The passive attack only eavesdrops a communication channel to gather sensitive information without modifying the messages eavesdropped. Two types of passive attacks are the release of message contents and traffic analysis. We denoted a passive attacker by the symbol Atc_1 in our proposed system.

The active attacks based on the Dolev-Yao attacker model [30]] can intercept messages, modify messages, or replay old messages to obtain confidential information. We denoted an active attacker by the symbol Atc_2 in our proposed system. The Atc_2 can intercept a transmitted interest/data packet, and try to obtain the intended information to conduct some attack. The Atc_2 even can masquerade or impersonate a legal new vehicle or RSU node to generate an interest packet or a data packet. Sometimes, the Atc_2 can compromise a RSU to obtain the legal identity of some vehicles. The Atc_2 also can flood the system with interest packets to crash the network or halt communication between different entities in the VNDN.

5. Proposed Scheme

Our proposed scheme consists of six stages: system initialization, registration stage, authentication stage, key agreement stage, public key update stage, and public key revocation stage. These stages are illustrated as follows:

5.1. System Initialization

In this stage, the VTA determines the algorithms and the parameters needed in our solution, such as hash function (e.g., SHA-256), EC-Schnorr signature algorithm, AES, and a bivariate polynomial function. The VTA deploys smart contracts on the blockchain to manage the public keys of the users.

Step 1: Building NDN-BCN and deploying of smart contracts

The VTA initializes a P2P BCN with an NDN prefix name. It also needs to create an Ethereum account for each NDN-BCN node by using an Ethereum wallet, such as MetaMask. The VTA creates smart contracts by using its account private key and address. Once the VTA has successfully deployed the smart contracts, a contract address is automatically created by NDN-BCN.

Step 2: Initialization of System Algorithms and Parameters

The VTA needs to determine a system parameter set $X = (p, a, b, G, n, h)$ to realize the public key cryptography system. These parameters determine an elliptic curve E on a finite field F_p, where p is a large prime number, and a and b are the parameters that define the elliptic curve E of a form $y^2 \pmod p = x^3 + ax + b \pmod p$. G is a generator denoted by a point (G_x, G_y) chosen from the elliptic curve, n is the order of the generator, h is SHA-256 hash function, and $h : \{0,1\}^* \longrightarrow F_p$ is a secure hash function. All the system parameters (p, a, b, G, n, h) are public to all the entities in the network.

Step 3: Bivariate polynomial initialization

A bivariate polynomial is introduced into the authentication stage in our solution to prevent a DoS attack and conduct a secure session key agreement. Moreover, two session keys are generated based on the polynomial for V2X. In the system initialization stage, VTA must set the relevant parameters of the bivariate polynomial. First, the VTA generates m bivariate polynomials $f(x, y)$ of degree n. They are denoted as $f_1, f_2, f_3, \ldots, f_m$. Then, the VTA generates n authentication key values K set $setK$. Each polynomial corresponds to an authentication key value K, and a polynomial fragment F is generated for the authentication value K using Eq. (3). Finally, the VTA randomly assigns an authentication key value $K(K \in setK)$ and a corresponding polynomial segment F to an RSU or a vehicle V when it needs to register on the VTA.

5.2. Registration Stage

The vehicle's information can be registered offline on the VTA in advance when a new vehicle needs to join the VDNN network and wants to communicate with other vehicles or access network services (RSU). Thus, the registration processes are illustrated as follows to ensure the legitimacy of the new vehicle:

Step 1: The vehicle owner submits all the registration information, such as the name of the vehicle owner, phone number, driving license, and physical vehicle number ID. The owner registers the vehicle's information to the VTA via a secure channel.

Step 2: The VTA generates a unique ID and a validity period (VP) for vehicle V. Then, the VTA selects $L(1 \leq L \leq m)$ polynomials from m bivariate polynomials. A corresponding authentication key value $K_v(K_v \in setK)$ and polynomial fragment F are distributed to the vehicle owner. According to the cryptography parameters (p, a, b, G, n, h), the user of the vehicle V computes a pair key (private key PrK and public key PbK). The vehicle V sends the public key PbK to the VTA. Afterward, the VTA generates a unique name $(PbKname)$ for the received public key PbK.

Step 3: The VTA binds the tuple $(ID, PbKname, PbK, andVP)$ and encodes it as hexadecimal codes. Then, it is compressed in JSON format. A new transaction is generated by the VTA. The generated hexadecimal code is embedded in its data field. Then, the VTA sends the transaction to the NDN-BCN for execution. It triggers the smart contract function register_ PK for vehicle registration. After the smart contract is successfully executed and deployed to the NDN-BCN, the transaction is stored in the NDN-BCN.

Step 4: After the registration is successfully executed, the secret materials $(ID, PrK, andVP)$ and the polynomial parameters are sent by the VTA to the vehicle. They are stored in the HSM, which is installed on the vehicle. Then, the vehicle V is implemented with the NDN prefix name /ndn/VNDN/V. The operation of the registration is described in Algorithm 1.

Algorithm 1 Aregister_PK function

Input ID, $PbKname$, PbK, VP
Output bool

1: **if** $msg.sender \neq VTA$ **then**
2: revert (); // Define an error and pass it while reverting a transaction
3: **else**
4: $users[ID] \leftarrow user(ID, PbKname, PbK, VP)$;
5: emit registerPK $(ID, PbKname, PbK, VP)$;
6: **end if**
return true;

5.3. Authentication Stage

After the registration stage is successfully completed, the authentication protocol is vital to realizing the system's security. This stage is described as follows and illustrated in Figure 4.

Step 1: When a new vehicle wants to participate and share information in VNDN, the RSU in the vehicle's communication range should first authenticate the legitimacy of the vehicle before allowing it to join the VNDN network. Given the broadcasted message of the RSU, the new vehicle V can obtain the public key PbK_{RSU} and the authentication key value K_{RSU} of the RSU, which are in the same communication range as the new vehicle V. At the beginning, the vehicle V sends a request message M_r in an interest packet to the RSU. The vehicle V first calculates a shared secret value sK_v between the vehicle V and RSU, as shown in Eq. (4), by using the symmetric bivariate polynomial function in Eq. (3). Then, the vehicle V computes a hash value HF on the message

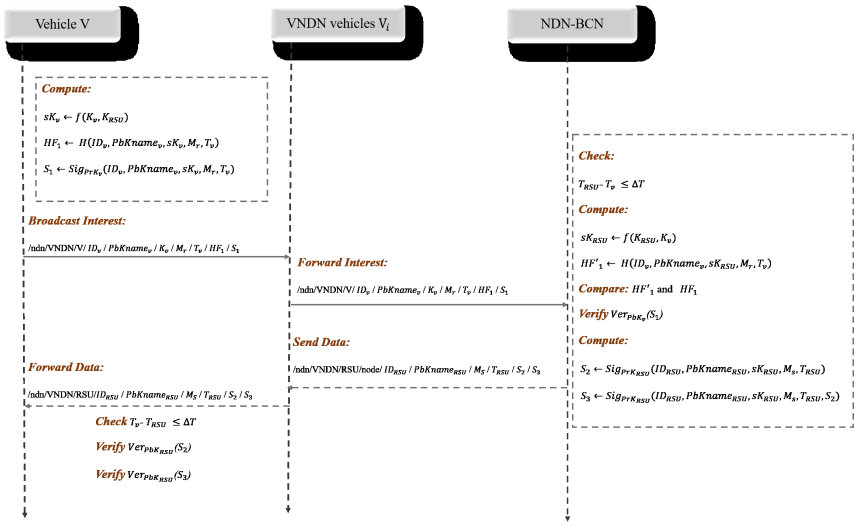


Fig. 4. An illustration of the authentication stage

$(ID_v, PbKname_v, sK_v, M_r, T_v)$ with HF_1 to guarantee the integrity of the message request, where the algorithm $H : \{0,1\}^* \longrightarrow \{0,1\}^h$ is a collusion-resistant hash function. The hash algorithm is described in Eq. (5), where M_r is the request message, and T_v is the vehicle current time stamp.

$$sk_v \leftarrow f\left(k_v, K_{RSU}\right). \tag{4}$$

$$HF_1 \leftarrow H\left(ID_v, PbKname_v, sK_v, M_r, T_v\right). \tag{5}$$

Step 2: The vehicle V needs to sign the request message to guarantee the authenticity of the message. The vehicle V signs the message $(ID_v, PbKname_v, sK_v, M_r, T_v)$ with its private key PrK_v, as shown in Eq. (6).

$$S_1 \leftarrow Sig_{PrK_v}\left(ID_v, PbKname_v, sK_v, M_r, T_v\right). \tag{6}$$

Afterward, the vehicle V generates an interest packet with the name /ndn/VNDN/V/ID_v /$PbKname_v$/K_v/M_r/T_v/HF_1/S_1 , where K_v is the authentication key value assigned to the vehicle V by the VTA. Then, this interest packet is broadcasted to the VNDN network, and a new PIT entry is created with the name prefix /ndn/VNDN/V/ID_v/$PbKname_v$/K_v /M_r/T_v/HF_1/S_1 in the router that received the interest packet copies.

Step 3: Upon receipt of the interest packet, the RSU calculates a shared secret value sK_{RSU} by using the same method used by the vehicle V, as shown in Eq. (7). Then, the RSU extracts HF_1 and S_1 from the interest packet to check the integrity and authenticity of the interest packet. If $T_{RSU} - T_v \leq \triangle T$, the RSU computes a commitment value $HF_1^{'}$ based on the secret key sK_{RSU}, as described in Eq. (8). If the $HF_1^{'}$ does not match the received HF_1, the session stops at this step.

$$sK_{RSU} \leftarrow f\left(K_{RSU}, K_v\right). \tag{7}$$

$$HF_1^{'} \leftarrow H\left(ID_v, PbKname_v, sK_{RSU}, M_r, T_v\right). \tag{8}$$

If the $HF_1^{'}$ matches the received HF_1, then the RSU checks the legitimacy of the new vehicle's public key by calculating the hash value of its public key in the received interest packet. It is compared with the hash value of the public key retrieved from the blockchain by using the public key's name $PbKname_v$. If these hash values are equal, the RSU believes that this new vehicle V is the owner of this public key, and it has been registered on the VTA before. After the legitimacy of the public key PbK_v is verified, the RSU uses PbK_v to verify the message signature S_1 to guarantee the authenticity of the requested message.

If the verification is successful, then RSU believes that the vehicle V is a legitimate vehicle. Moreover, the vehicle V is allowed to participate and share information within VNDN.

Step 4: The new vehicle in the VNDN also needs to authenticate the RSU that communicates with it to guarantee the legitimacy of this RSU and prevent an adversary from masquerading as an RSU. Thus, the RSU is required to sign a message M_s in a data packet forwarded to the vehicle V. The RSU signs the message $(ID_{RSU}, PbKname_{RSU}, sK_{RSU}, M_s, T_{RSU})$ with its private key PrK_{RSU} with S_2 to ensure the authenticy of the message, as shown in Eq. (9). Then, the RSU prepares a data packet with the message. The RSU also signs the data packet with S_3 to guarantee the authenticity of the data packet, as shown in Eq. (10). Then, the data packet with the name /ndn/VNDN/RSU/node/ID_{RSU}/$PbKname$/K_{RSU}/M_s/T_{RSU}/S_2/S_3 is sent. The data packet is appended in PIT and forwarded to the vehicle V.

$$S_2 \leftarrow Sig_{PrK_RSU}\left(ID_{RSU}, PbKname_{RSU}, sK_{RSU}, M_s, T_{RSU}\right). \tag{9}$$

$$S_3 \leftarrow Sig_{PrK_RSU}\left(ID_{RSU}, PbKname_{RSU}, sK_{RSU}, M_s, T_{RSU}, S_2\right). \tag{10}$$

Step 5: When the vehicle V receives the data packet, it first checks if $T_{RSU} - T_v \leq \triangle T$. Then, it retrieves the public key PbK_{RSU} of the RSU from the blockchain by using the name $PbKname_{RSU}$ to check the legitimacy and validity of the RSU's public key PbK_{RSU} by using the hash algorithm. After the legitimacy of PbK_RSU is verified, the vehicle V extracts S_2 and S_3 from the data packet. Then, it uses PbK_{RSU} to verify the message signature S_2 and the signature S_3 on the data packet to guarantee the authenticity of the response message and the data packet. If they are correct, the vehicle V believes that the RSU is legal. Otherwise, the vehicle V stops at this step. Finally, the lightweight mutual authentication between the vehicle V and RSU is completed.

5.4. Key Agreement Stage

In this stage, a vehicle $V_i (i = 1, 2, 3, \ldots n)$ in the VNDN helps complete a communication process between the vehicle V_i and an RSU node. Two session keys are established for V2X to guarantee the confidentiality of the content in an interest packet and a data packet containing the relevant parameters of the polynomial used in our solution. These keys are used for future communications based on the symmetry of the bivariate polynomial, as described in Section 3. The key agreement protocol is described in Figure 5. In this stage, the communications for V2X are based on symmetric key encryption. The key agreement process is described as follows:

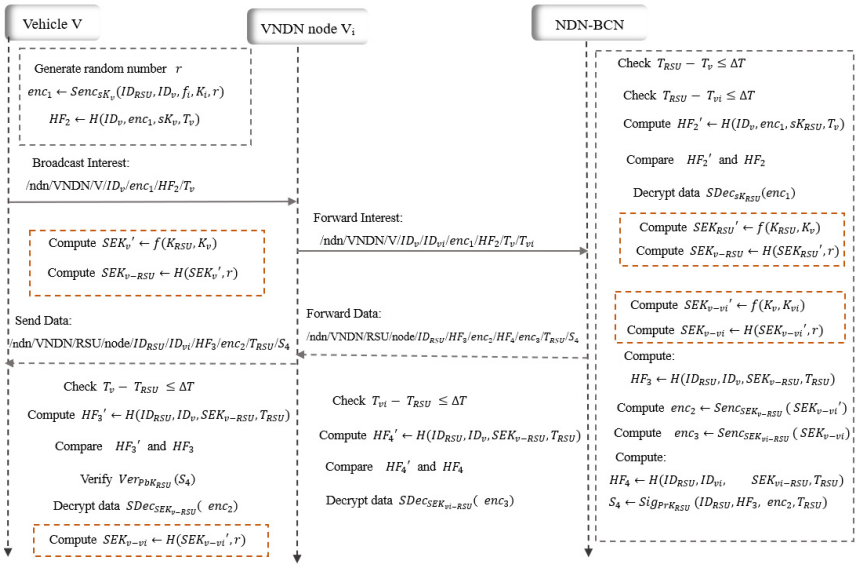


Fig. 5. An illustration of the key agreement stage

Step 1: After the above authentication process between the vehicle V and an RSU is completed, an agreement process of a session key SEK must be established between the vehicle V and RSU as SEK_{v-RSU}. The vehicle V first generates a random number r. Then, it uses the secret value sK_v described in Eq. (4) to encrypt the relevant message $(ID_{RSU}, ID_v, (f_i, K_i), r)$ described in Eq. (11), where (f_i, K_i) represents a random polynomial, which is chosen from the polynomial parameters assigned to the vehicle V by the VTA, as described in the registration stage. $Senc_{sK_v}(m)$ is a symmetric key encryption algorithm on message m using the secret value sK_v of the vehicle V. Similar to the case in the HF_1 generation process in the authentication stage, the vehicle V also needs to compute a hash authentication function HF_2 to prevent DoS attack, as described in Eq. (12).

$$enc_1 \leftarrow Senc_{sK_v}(ID_{RSU}, ID_v, f_i, K_i, r). \tag{11}$$

$$HF_2 \leftarrow H(ID_v, enc_1, sK_v, T_v). \tag{12}$$

Afterward, the vehicle V generates an interest packet with the name /ndn/VNDN/V/ID_v /enc_1/HF_2/T_v. A PIT entry is created with the name prefix /ndn/VNDN/V/ID_v/enc_1/HF_2 /T_v. Then, this interest packet is broadcasted to the VNDN network.

Step 2: Upon the receipt of the broadcasted interest packet, a trusted neighbor vehicle V_i, which is authenticated while it joins the VNDN network, helps complete the communication process. The vehicle V_i generates a new interest packet with a name prefix /ndn/VNDN/V/ID_v/ID_{vi}/enc_1/HF_2/T_v/T_{vi}. The new interest packet is added to the PIT. The interest packet is forwarded to the RSU located in its communication range.

Step 3: After the RSU receives the interest packet /ndn/VNDN/V/ID_v/ID_{vi}/enc_1/HF_2 /T_v/T_{vi} from the vehicle V_i, the RSU extracts enc_1 and HF_2 from the interest packet. If

$T_{RSU} - T_v \leq \triangle T$ and $T_{RSU} - T_{vi} \leq \triangle T$, then the RSU computes a hash function $HF_2^{'}$ based on the message in the received interest packet. If the $HF_2^{'}$ is equal to the received HF_2, the RSU decrypts enc_1 to obtain the relevant parameters for a session key establishment by using the secret value sK_{RSU} described in Eq. (7).

Step 4: First, the RSU computes a session key parameter $SEK_{RSU}^{'}$ using the corresponding polynomial parameters f_i and the fragments F simultaneously to generate a session key. The vehicle V computes a session key parameter $SEK_v^{'}$ using the same method used by the RSU. Then, the RSU uses the hash function H with the $SEK_{RSU}^{'}$ and r to compute an agreement session key SEK_{v-RSU}, as in Eq. (13). The vehicle V uses the hash function H with the $SEK_v^{'}$ and r to compute an agreement session key SEK_{v-RSU}, as in Eq. (14). On the basis of the information on the vehicle V and the vehicle V_i that the RSU obtained during the communication, the RSU helps the vehicle V_i to generate an agreement session key SEK_{v-vi} between it and the vehicle V. Thus, the RSU should compute a session key parameter $\left(SEK_{v-vi}^{'}\right)$. Then, it computes the generated agreement session key SEK_{v-vi}, as in Eq. (15).

$$SEK_{v-RSU} \leftarrow H\left(SEK_{RSU}^{'}, r\right). \tag{13}$$

$$SEK_{v-RSU} \leftarrow H\left(SEK_{v}^{'}, r\right). \tag{14}$$

$$SEK_{v-vi} \leftarrow H\left(SEK_{v-vi}^{'}, r\right). \tag{15}$$

Step 5: The RSU needs to communicate with the vehicle V_i to confirm that the session key is successfully established. The RSU generates a hash value HF_3, as described in Eq. (16). Moreover, the RSU computes enc_2 by using the session key SEK_{vi-RSU}, as described in Eq. (17). Meanwhile, the RSU generates a hash value HF_4, as described in Eq. (18), to guarantee the integrity of the message between the RSU and the vehicle V_i. The RSU also computes enc_3 by using SEK_{v-RSU} to encrypt the message $\left(SEK_{v-vi}^{'}\right)$, as in Eq. (19).

$$HF_3 \leftarrow H\left(ID_{RSU}, ID_v, SEK_{v-RSU}, T_{RSU}\right). \tag{16}$$

$$enc_2 \leftarrow Senc_{SEK_{v-RSU}}\left(SEK_{v-vi}^{'}\right). \tag{17}$$

$$HF_4 \leftarrow H\left(ID_{RSU}, ID_{vi}, SEK_{vi-RSU}, T_{RSU}\right). \tag{18}$$

$$enc_3 \leftarrow Senc_{SEK_{vi-RSU}}\left(SEK_{v-vi}\right). \tag{19}$$

Afterward, the RSU prepares a data packet with the message $(ID_{RSU}, HF_3, enc_2, enc_3, HF_4)$, which is included in the content. Then, the RSU signs the data packet by using its private key PrK_{RSU} with S_4. The RSU sends the data packet with a name prefix /ndn/VNDN/RSU/ID_{RSU}/HF_3/enc_2/HF_4/enc_3/T_{RSU}/S_4 to the vehicle V_i. The data packet with the name /ndn/VNDN/RSU/ID_{RSU}/HF_3/enc_2/HF_4/enc_3 /T_{RSU}/S_4 is appended to the PIT.

Step 6: After the vehicle V_i receives the data packet from the RSU, it extracts HF_4 and enc_3 from the data packet. If $T_{vi} - T_{RSU} \leq \triangle T$, the vehicle V_i first computes $HF_4^{'}$. Then, it compares the $HF_4^{'}$ with the received HF_4. If they are correct, the vehicle V_i

decrypts enc_3 by using SEK_{vi-RSU} to obtain the session key SEK_{v-vi}. The vehicle V_i forwards the data packet /ndn/VNDN/RSU/ID_{RSU}/ID_{vi}/HF_3/enc_2/T_{RSU}/S_4 to the vehicle V.

Step 7: When the vehicle V receives the data packet from the vehicle V_i, it first extracts HF_3 and enc_2 from the data packet. If $T_v - T_{RSU} \leq \triangle T$, the vehicle V computes the hash function $HF_3^{'}$ based on the message in the received data packet. If $HF_3^{'}$ is equal to the received HF_3, the vehicle V also verifies the signature S_4 of the received data packet by using the public key of the RSU. If it is correct, the vehicle V believes that the session key SEK_{v-RSU} is shared successfully between V and RSU. Then, the vehicle V decrypts enc_2 to obtain the session key parameters SEK_{v-vi}. The vehicle V computes the agreement session key SEK_{v-vi} between the vehicle V and the vehicle V_i by using the same method by the RSU using a hash function with $SEK_{v-vi}^{'}$ and r.

5.5. Public Key Update Stage

In this stage, the user requests to update the public key data. The users can apply for a key update if the VP of the current public key is about to expire. In the registration stage, the VTA conducts a detailed review of the vehicle user. Thus, at this stage, the VTA does not need to review the user's identity information in detail. The processes of the update public key are described in Algorithm 2.

Algorithm 2 update_PK function

Input $ID, new_PbKname, new_PbK, new_VP$
Output bool

1: **if** $msg.sender \neq VTA$ **then**
2: revert (); // Define an error and pass it while reverting a transaction
3: **else**
4: $users[ID] \leftarrow user(ID, new_PbKname, new_PbK, new_VP)$;
5: emit updatePK $(ID, new_PbKname, new_PbK, new_VP)$;
6: **end if**
return true;

The user sends a request message to the VTA for an updated public key. The request includes $(ID, PbKname, PbK, andVP)$. The VTA performs a simple verification on the user. After the verification is passed, the user generates a new key pair (new_PbK, new_PrK) in the registration stage. Then, the user of the vehicle V sends the new public key n_PbK to the VTA. Then, the VTA generates a new PbKname and new VP for the new public key.

Afterward, the VTA binds the tuple $(ID, new_PbKname, new_PbK, new_VP)$ and encodes it as hexadecimal codes. Then, it is compressed in JSON format. A new transaction is generated by the VTA, and the generated hexadecimal code is embedded in its data field. Then, the VTA sends the transaction to the NDN-BCN for execution. The smart contract function update_PK is triggered for the update of the public key of the vehicle V. After the smart contract is successfully executed and deployed to the NDN-BCN, the transaction is stored in the NDN-BCN.

After the updated public key data are successfully uploaded to the chain, the VTA also needs to update the polynomial previously assigned to the vehicle; that is, VTA selects new $L(1 \leq L \leq m)$ polynomials from m polynomials and distributes a new authentication key value $K_v(K_v \in setK)$ and polynomial fragment F to the user of the vehicle V. Finally, the VTA securely sends secret materials and the relevant parameters of the new polynomial to the vehicle user. Thus far, the public key update process of the vehicle user has been completed.

5.6. Public Key Revocation Stage

In the management mechanism proposed in this study, the security of the public and private keys also means the security of the VNDN system. Disclosing or stealing the user's private key seriously impacts the entire system. Therefore, an effective detection mechanism and a revocation mechanism for invalid public keys are urgently needed.

Step 1: When the RSU receives the broadcast interest packet from the vehicle user, the RSU takes the malicious user's ID and sends it to the VTA if the interest packet has serious authenticity deviation. The VTA defines this user as a malicious user.

Algorithm 3 revoke_PK function

Input ID
Output bool

1: **if** $msg.sender \neq VTA$ **then**
2: revert (); // Define an error and pass it while reverting a transaction
3: **else**
4: $users[ID] \leftarrow user\,(0)$;
5: number=number-1;
6: emit revokeID (ID);
7: **end if**
return true;

Step 2: After the VTA determines the identity of the malicious user, the VTA sends the revoked transaction to the NDN-BCN and executes the revoke_PK function in the contract to remove the user's identity and the binding $(PbKname, PbK)$. The transaction record is removed after the smart contract is successfully executed and mined. On the NDN-BCN, the malicious user's public key is identified as invalid and returned to the VTA for execution results. Thus far, the public key revocation process of the vehicle user has been completed. Algorithm 3 describes the revocation function operation of the user's public key.

6. Security Analysis

6.1. Informal Security Analysis

Our proposed scheme in this paper uses blockchain to manage the user's identity and public key. The decentralized key management solution can effectively resolve the problems

of the traditional solution based on the centralized PKI. And the proposed scheme satisfies the following security requirements:

1) **Entity Authentication:** In our proposed scheme, mutual authentication between different communication entities by using our authentication protocol is required to verify the legitimacy of these entities such as the vehicles or RSUs. So, our scheme can provide confidence that an entity can't perform the active attacks such as either a masquerade or an unauthorized replay.
2) **Data Confidentiality:** In our VNDN system, a sensitive data transmitted on the network is required to be encrypted by using a symmetric encryption to prevent a passive attack mentioned in section 4.2. So, a session key needs to be generated by using our key agreement protocol before two authenticated entities communicate. We denote the session key between a vehicle V and a RSU by the symbol SEK_{v-RSU} and denotes a session key between the vehicle V and its neighbor vehicles V_i by the symbolSEK_{v-vi}. These two session keys are created based on bivariate polynomial parameters, and are stored in the HSM. Of course, any adversary cannot obtain them to compromise the data confidentiality.
3) **Message Authentication:** The EC-Schnorr signature algorithm is used to ensure the integrity authenticity of a message transmitted between different entities in VNDN system. The private keys (PrK_v, PrK_{RSU}) used in the EC-Schnorr algorithm are stored in HSM. Thus, an adversary cannot obtain these cryptographic materials to forge a valid message signature. In addition to, our proposed scheme also can prevent the active attack mentioned in section 4.2 by flooding a large number of interest packets.

6.2. Security Analysis in the UC Framework

We follow the approach of the UC framework to analyze the security attributes of our protocols in our VNDN system. Our main protocols include an authentication protocol and a key agreement protocol. The notion of this framework is based on simulation. That is to say, for any real-world protocol π and an adversary A, if there exist an ideal protocol ϕ, ideal functionality F and a simulator S can simulate running of the protocol π, then we say that the protocol π . In other words, for any environment ε, if the probability that the ε can tell the difference between the execution of the protocol π with the adversary A and the execution of the ideal protocol ϕ with the adversary S is neglible at most, the protocol π is a UC-secure protocol. Therefore, the UC security ensures a composable protocol of any UC-secure protocols is also secure.

The ideal functionality F is considered the most important part of the UC framework. It is a trusted party for achieving the security requirements of cryptographic protocols. It guarantees the secrecy, authenticity, or delivery of the communicated information. Most of the ideal functionalities have already been formulated, such as the signature functionality F_{SIG}, the public key encryption functionality F_{PKE}, the secure message transmission functionality F_{SMT}, the key exchange functionality F_{KE}, and the message authentication functionality F_{AUTH}.The interested reader can refer to the related literatures. These ideal functionalities can be used as the standard ideal functionality in our security analysis.

A. Security Analysis of the Authentication Stage

The real authentication protocol π_{AUTH} in the real world is described in Figure 4. The difference with the ideal authentication protocol π_{AUTH} described in Figure 6 is that the ideal protocol should be executed in (F_{SIGN}, F_{REG})-hybrid model, where the two functions (F_{SIGN}, F_{REG}) in F_{AUTH} are selected to realize the composition theorem. . In particular, the authentication protocol π_{AUTH} uses the signature ideal functionality F_{SIGN} to achieve the signature process. The registration ideal functionality F_{REG} is used for registering the key material parameters such as K_v and K_{RSU} and so on. The detail of the ideal authentication protocol ϕ_{AUTH} is described in Figure 6.

Ideal Protocol φ_{AUTH} for $\mathcal{F}_{AUTH}$

Let $SID = (ID_v, ID_{RSU}, sid)$ where sid is the session identity of ID_v and ID_{RSU}.

1) The vehicle V is activated with input (Send, SID, Interest packet) to the ideal functionality $\mathcal{F}_{AUTH}$.
2) Upon receiving (Sent, SID, new Interest packet), the RSU sends the response (Send, SID, Data packet) to the functionality $\mathcal{F}_{AUTH}$, which may forward through VNDN nodes to the vehicle V.

Fig. 6. Ideal Authentication Protocol ϕ_{AUTH}

Theorem 1. *The UC framework can be used to prove the security attributes of the authentication protocol π_{AUTH} in the (F_{CERT}, F_{REG})-hybrid model, i.e., the real protocol π_{AUTH} can securely access the ideal protocol ϕ_{AUTH} for the functionality F_{AUTH}.*

Proof. Let π_{AUTH} be the real protocol in the real world. We say that the protocol π_{AUTH} securely realizes an ideal protocol ϕ_{AUTH} for the ideal functionality F_{AUTH} if an ideal process adversary S exists for any real-world adversary A. In particular, no environment ε can tell the difference between a real process and an ideal process with a nonnegligible probability.

An adversary S in the (F_{SIGN}, F_{REG})-hybrid model is responsible for delivering the message from the copies of F_{CERT} and F_{REG}. This message is forwarded to a real adversary A. The adversary S can be a simulator for the adversary A. Thus, adversary S can activate a real adversary A. Moreover, the simulator S forwards the instruction from environment ε to adversary A and copies the output of the adversary A to the environment ε. The detailed of adversary S is as follows:

1) When the vehicle V is activated with input (**Send**, SID, Interest packet), the adversary S simulates the forwarding process of the interest packet with the name prefix /ndn/VNDN/V/ID_v/$PbKname_v$/K_v/M_r/T_v/HF_1/S_1 for A from the vehicle V to the RSU. First, the message $(ID_v, PbKname_v, K_v, M_r, T_v, HF_1, S_1)$ is extracted from the interest packet. Upon receiving (**Send**, SID, Interest packet), the adversary S first wants to retrieve the key value parameters K_v and K_{RSU}. It sends (**Retrieve**, sid, V)

and (**Retrieve**, sid, RSU). Then, it obtains the responses (**Retrieve**, sid, V, K_v) and (**Retrieve**, sid, RSU, K_{RSU}) by the interaction with the F_{REG}. After the adversary S obtains the key values parameters, it computes the secret value sK_v according to the bivariate polynomial $sK_v \leftarrow f(K_v, K_{RSU})$. Once the secret values are obtained, the adversary S computes the hash value $HF_1 \leftarrow H(ID_v, PbKname_v, sK_v, M_r, T_v)$. Then, the adversary S interacts with F_{SIGN} to obtain S_1, when it receives (**Sign**, sid, V, $(ID_v, PbKname_v, sK_v, M_r, T_v)$) from F_{SIGN}. The adversary S forwards the message to the adversary A. When the adversary A outputs (**Signature**,sid, V,$(ID_v$, $PbKname_v, sK_v, M_r, T_v), S_1$), the adversary S assigns S_1 as the signature of the vehicle V. Then, the message is sent to the functionality F_{SIGN}.

2) When the adversary A delivers the interest packet /ndn/VNDN/V/ID_v/$PbKname_v$/K_v /M_r/T_v/HF_1/S_1 from the vehicle V, it is forwarded through the VNDN nodes to RSU. The adversary S simulates for the adversary A the receiving interaction with F_{SIGN}. When the adversary S receives the message (**Verify**,sid, V,$(ID_v, PbKname_v, sK_v$ $, M_r, T_v), S_1$) from F_{SIGN}, this message is forwarded to the adversary A. Then, the adversary S obtains the response (**Verify**,sid, V,$(ID_v, PbKname_v, sK_v, M_r, T_v), \sigma_1$) from the adversary A. Afterward, S sends the response message obtained from the adversary A to F_{SIGN}. If F_{SIGN} outputs (**Verified**,sid, V,$(ID_v, PbKname_v, sK_v, M_r, T_v)$ $, \sigma_1, f = 1$) to RSU, the adversary S sends (**Sent**, SID, Interest packet) to the authentication functionality F_{AUTH}. Otherwise, it does not do anything.
3) When the RSU is activated with the input (**Send**, SID, Data packet), the adversary S simulates for A the data packet /ndn/VNDN/RSU/node/ID_RSU/$PbKname_{RSU}$/ K_{RSU}/M_s/T_{RSU}/S_2/S_3 from RSU that is forwarded to the vehicle V. The adversary S also has sK_{RSU}. Then, it obtains S_2 and S_3 by the interaction with F_{SIGN}. When the adversary S receives the message (**Sign**, sid, RSU,$(ID_RSU, PbKname_{RSU}$, $sK_{RSU}, M_s, T_{RSU})$) from F_{SIGN}, the adversary S forwards the message to A. When the adversary A outputs (**Signature**,sid, RSU,$(ID_{RSU}, PbKname_{RSU}, sK_{RSU}, M_s$ $, T_{RSU}), S_2$), the adversary S assigns S_2 as the signature of the RSU and then sends the message to F_{SIGN}. Afterward, S also receives the message (**Sign**, sid, RSU, $(ID_{RSU}, PbKname_{RSU}, sK_{RSU}, M_s, T_{RSU}, S_2)$) from F_{CERT}. S forwards the message to A. When the adversary A outputs (**Signature**,sid, RSU,$(ID_{RSU}, PbKname_{RSU}$ $, sK_{RSU}, M_s, T_{RSU}, S_2), S_3$), the adversary S assigns S_3 as the signature of the data packet of RSU and then sends the message to F_{SIGN}.
4) When the adversary A delivers the data packet /ndn/VNDN/RSU/node/ID_{RSU}/ $PbKname_{RSU}$/K_{RSU}/M_s/T_{RSU}/S_2/S_3, RSU forwards it through the VNDN node to V. S also simulates for the adversary A the receiving interaction with F_{SIGN}. When the adversary S receives the messages (**Verify**,sid,RSU,$(ID_{RSU}, PbKname_{RSU}$ $, sK_{RSU}, M_s, T_{RSU}), S_2$) and (**Verify**,sid,RSU,$(ID_{RSU}, PbKname_{RSU}, sK_{RSU}$, $M_s, T_{RSU}, S_2), S_3$) from F_{SIGN}, it forwards these messages to the adversary A. Then, the adversary S obtains the responses (**Verify**,sid, RSU,$(ID_{RSU}, PbKname_{RSU}$ $, sK_{RSU}, M_s, T_{RSU}), \sigma_2$) and (**Verify**,sid, RSU,$(ID_{RSU}, PbKname_{RSU}, sK_{RSU}$, $M_s, T_{RSU}, S_2), \sigma_3$) from the adversary A. Afterward, S sends the response messages obtained from the adversary A to F_{SIGN}, if F_{SIGN} outputs (**Verified**,sid, RSU,$(ID_{RSU}, PbKname_{RSU}, sK_{RSU}, M_s, T_{RSU}), \sigma_2, f = 1$) and (**Verified**,sid, RSU,$(ID_{RSU}, PbKname_{RSU}, sK_{RSU}, M_s, T_{RSU}, S_2), \sigma_3, f = 1$) to RSU. Finally, the adversary S sends (**Sent**, SID, Data packet) to F_{AUTH}. Otherwise, it does not do anything.

According to the symmetry property of the polynomial function and the logic of F_{REG} that used to register the key values parameters, the adversary A cannot change K_v and K_{RSU}, which are registered in F_{REG}. Thus, the adversary A cannot modify the secret values sK_v and sK_{RSU} that are calculated based on the bivariate polynomial function. However, based on the resistant collusion of the hash function HF_1 and the signatures obtained from F_{SIGN}, the adversary A can attack the real protocol π_{AUTH} and forge the hash message and the signature message. Adversary A can create a forge interest packet as /ndn/VNDN/V/$ID_{v'}$/$PbKname_{v'}$/K_v/$M_r^{'}$/$T_{v'}$/HF_1/S_1 when it obtains the interest packet /ndn/VNDN/V/ID_v/$PbKname_v$/K_v/M_r/T_v/HF_1/S_1. Moreover, the adversary A can create a forge data packet /ndn/VNDN/RSU/node/$ID_{RSU'}$/$PbKname_{RSU'}$/$K_{RSU'}$/ $M_s^{'}$/$T_{RSU'}$/S_2/S_3. The verification process indicate that the authentication of the interest packet and data packet of the real protocol π_{AUTH} is different from that of the interest packet and data packet of the ideal protocol ϕ_{AUTH} for functionality F_{AUTH}.

Collusion in the hash function is hard to find. The probability of obtaining the collusion and forging the signature is negligible at most. However, the resistant collusion in the hash function and the logic of F_{SIGN} is unforgeability against chosen-message attacks. Therefore, we can say that environment ε can distinguish between the real protocol π_{AUTH} in the (F_{SIGN}, F_{REG})-hybrid model and the ideal protocol ϕ_{AUTH} for functionality F_{AUTH} with a negligible probability, at most.

B. Security Analysis of the Key Agreement Stage

The real key agreement protocol π_{KA} in the real world is described in Figure 5. The main difference with the ideal key agreement protocol ϕ_{KA} in the ideal process is that the protocol π_{KA} should be in the $(F_{PKE}, F_{SIGN}, F_{REG})$-hybrid model. In particular, the key agreement protocol π_{KA} uses the ideal functionality F_{PKE} for realizing the encryption algorithm. The ideal functionalities F_{SIGN} and F_{REG} are used in the authentication protocol. The detail of the ideal protocol ϕ_{KAG} is described in Figure 7.

Theorem 2. *The UC framework can be used to prove security attributes of the key agreement protocol π_{KA} in the $(F_{PKE}, F_{SIGN}, F_{REG})$-hybrid model, i.e., the real protocol π_{KA} can securely access the ideal protocol ϕ_{KA} for the functionality F_{KE}.*

Proof. Let π_{KA} be the real protocol in the real world. We say that the protocol π_{KA} securely realizes an ideal protocol ϕ_{KA} for the ideal functionality F_{KE} if an ideal-process adversary S exists for any real-world adversary A. In particular, no environment ε can tell the difference between the real process and ideal process with a nonnegligible probability.

The details of the adversary S is as follows:

1) When the vehicle V is activated with input (**CreateSEK**, SID_1), the adversary S first retrieves the key values parameters K_v and K_{RSU}. It sends (**Retrieve**, sid_1, V) and (**Retrieve**, sid_1, RSU). Then, it obtains the responses (**Retrieve**, sid_1, V, K_v) and (**Retrieve**, sid_1, RSU, K_{RSU}) through the interaction with the F_{REG}. After the adversary S obtains the key values parameters, it computes the secret values sK_v and sK_{RSU} based on the polynomial function f defined in the system initialization stage. Then, the simulator S selects random polynomials f_i and K_i and a random number r. It sends (**Encrypt**, sid_1, RSU, $(ID_{RSU}, ID_v, f_i, K_i, r)$) to the functionality F_{PKE}. When the vehicle V receives the response (**Ciphertext**, sid_1,

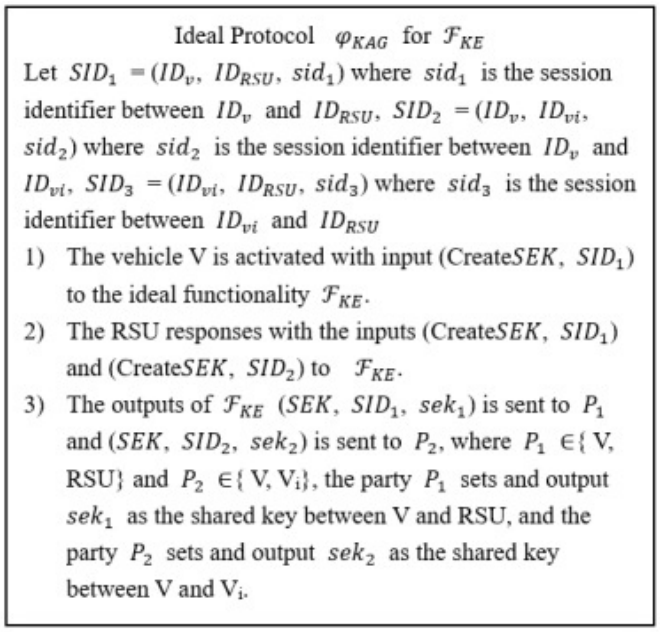


Fig. 7. Ideal Key Agreement Protocol ϕ_{KAG}

RSU, enc_1) from the functionality F_{PKE}, the adversary S computes the hash value $HF_2 \leftarrow H(ID_v, enc_1, sK_v, T_v)$ and simulates for the adversary A the forwarding process of the interest packet with the name prefix /ndn/VNDN/V/ID_v/enc_1/HF_2/T_v from the vehicle V through the vehicle V_i to the RSU.

2) When the adversary A delivers the interest packet /ndn/VNDN/V/ID_v/enc_1/HF_2/T_v from the vehicle V to the RSU, the simulator S simulates for the adversary A the receiving interaction with F_{PKE}. When the adversary S receives the message (**Decrypt**, sid_1, RSU, enc_1) from the simulated process of RSU, it forwards this message to the functionality F_{PKE}. Then, the adversary S obtains the response (**Plaintext**, sid_1, RSU, $(ID_{RSU}, ID_v, f_i, K_i, r)$) from the F_{PKE} to RSU. Afterward, the adversary S computes $SEK'_{(v-RSU)} \leftarrow f(K_{RSU}, K_v)$. Finally, the adversary S assigns the session key $SEK_{(v-RSU)}$ as $SEK_{v-RSU} \leftarrow H\left(SEK'_v, r\right)$ in the simulation of the real protocol π_{KA}.
3) When the RSU is activated with the inputs (**CreateSEK**, SID_1) and (**CreateSEK**, SID_2), the adversary S first retrieves $SEK_{(vi-RSU)}$. It sends (**Retrieve**, sid_3, V_i, RSU). Then, it obtains the response (**Retrieve**, sid_3, V_i, RSU, SEK_{vi-RSU}) through the interaction with the F_{KE} function. The adversary S computes HF_3, HF_4, and SEK'_{v-vi}. Then, S assigns the session key as $SEK_{v-vi} \leftarrow H\left(SEK'_{v-vi}, r\right)$ in the simulated of the real protocol π_{KA}. The adversary S sends (**Encrypt**, sid_1, V, SEK'_{v-vi}), and (**Encrypt**, sid_3, V_i, (SEK_{v-vi})) to the functionality F_{PKE}. When the RSU receives the response (**Ciphertext**, sid_1, V, enc_2) and (**Ciphertext**, sid_3, V_i, enc_3) from the functionality F_{PKE}, S obtains S_4 by interacting with F_{SIGN}. When

S receives the message (**Sign**, sid_1, V, $(ID_{RSU}, HF_3, enc_2, T_{RSU})$) from F_{SIGN}, S simulates for adversary A the data packet with the name prefix /ndn/VNDN/RSU/ID_{RSU} /HF_3/enc_2/HF_4/enc_3/T_{RSU}/S_4 from RSU to the vehicle V_i. The forwarding data packet /ndn/VNDN/RSU/ID_{RSU}/HF_3/enc_2/T_{RSU}/S_4 from the vehicle V_i to the vehicle V is also simulated.

4) When the adversary A delivers the data packet /ndn/VNDN/RSU/ID_{RSU}/HF_3/enc_2 /HF_4/enc_3/T_{RSU}/S_4 from RSU to the vehicle V_i and the forwarding data packet /ndn/VNDN/RSU/ID_{RSU}/HF_3/enc_2/T_{RSU}/S_4 from the vehicle V_i to the vehicle V, the adversary S simulates for the adversary A the receiving interaction with the functionality F_{PKE} and functionality F_{SIGN}. When the adversary S receives the messages (**Decrypt**, sid_1, V, enc_2) from the simulated process of the vehicle V and (**Decrypt**, sid_3, V_i, enc_3) from the simulated process of the vehicle V_i, it forwards these messages to the functionality F_{PKE}. Then, the adversary S obtains the responses (**Plaintext**, sid_1, V, $(SEK^{'}_{v-vi})$) from F_{PKE} to the vehicle V and (**Plaintext**, sid_3, V_i, (SEK_{v-vi})) from the functionality F_{PKE} to the vehicle V_i. When receiving (**Verify**, sid_1, V, $(ID_{RSU}, HF_3, enc_2, T_{RSU})$) from F_{SIGN}, the adversary S forwards this message to the adversary A and obtains the response (**Verify**, sid_1, V_i, $(ID_{RSU}, HF_3, enc_2, T_{RSU})$,$\sigma_4$) from the adversary A. Moreover, the responses are forwarded to the functionality F_{SIGN}.

5) For the outputs $(SEK, sid_1, P_1, SEK_{v-RSU})$ and $(SEK, sid_2, P_2, SEK_{v-vi})$ of the simulation of the real protocol π_{KA}, the adversary S forwards them to the functionality F_{KE}. If the vehicle V, vehicle V_i, and RSU are not corrupted, the vehicle V_i and RSU output (SEK, sid_1, P_1, sek_1). Moreover, the vehicle V and vehicle V_i output (SEK, sid_2, P_2, sek_2) according to the logic of the functionality F_{KE} in the ideal protocol ϕ_{KAG}.

According to the symmetry property of the polynomial function described in Section 3 and the logic of F_{REG} used to register the key values parameters, the adversary A cannot change K_v, K_{RSU}, and K_{vi}, which are registered in F_{REG}, and the logic of F_{PKE}. Thus, the adversary A cannot modify the secret values sK_v, sK_{vi}, and sK_{RSU}, which are computed based on bivariate polynomial functions. Moreover, the adversary A cannot modify the session key SEK_{vi-RSU} that is created in F_{KE}. However, the resistant collusion of the hash functions HF_2, HF_3, and HF_4 and the signatures that are obtained from F_{SIGN} indicate that the adversary A can attack the real protocol π_{KA} and forge the hash message and the signature message. When the adversary A obtains the interest packet /ndn/VNDN/V/ID_v/enc_1/HF_2/T_v, it can create a forge interest packet as /ndn/VNDN/V/SID_1/$ID_{v'}$ /enc_1/HF_2/$T_{v'}$. Moreover, the adversary A can create a forge data packet /ndn/VNDN/RSU/SID_1/SID_2/SID_3/$ID_{RSU'}$/HF_3/enc_2/HF_4/enc_3/$T_{RSU'}$/S_4.

In addition, the difference between the session keys SEK_{v-RSU} and SEK_{v-vi} in the real protocol π_{KA} and the session keys and in the ideal protocol ϕ_{KAG} that the environment ε can distinguish is a negligible. That because if the environment ε can tell the difference between the session keys in the real protocol and ideal protocol with just a non-negligible probability, the ciphertexts enc_1 and enc_2 can be distinguish by constructing an adversary A based on the environment ε , where this will conflict with the logic of the functionality F_{PKE}, as we know the UC-secure public key encryption that implemented the ideal functionality F_{PKE} is a CCA-secure encryption.

Collusion in the hash function is difficult to find, and the probability of obtaining the collusion and forging the signature is negligible at most. However, the resistant collusion in the hash function and the logic of F_{SIGN} and F_{PKE} are unforgeability against the chosen-message attacks. Therefore, we can say that environment ε can distinguish between the real protocol π_{KA} in the $(F_{SIGN}, F_{PKE}, F_{REG})$-hybrid model and the ideal protocol ϕ_{KAG} for functionality F_{KE} with a negligible probability at most.

7. Performance Analysis

7.1. Implementation of our scheme

The implementation of our scheme depends on the ndn-geth project to implement the blockchain nodes with the NDN protocols. A private blockchain is generated in virtual machine of the Ubuntu 20.04.4 LTS with 4 GB RAM on a laptop. To implement the NDN protocol on the laptop, first, the NFD and ndn-cxx are installed to obtain their requirements libraries. Then, the ndn-geth is cloned from GitHub. In this study, the blockchain nodes are maintained by the RSUs In the VNDN system. The private blockchain environment contains two nodes node1 and node2. After each node is generated, it is necessary to ensure that the NFD is running. the prefix name /ndn/VNDN/RSU1/node1 is assigned to the first node1, and the prefix name /ndn/VNDN/RSU2/node2 is assigned to the another node node2. These two nodes are configured with the same genesis file, Figure 8 shows the JISON configuration file of the prefix name and the block IDs of the NDN's host communication interface. These two nodes can communication by using the NDN protocol, and their information will be added the router list of the NDN. The peer-to-peer host configuration of NDN-blockchain is shown in Figure 9 and the router list of the NDN host is shown in Figure 10.



Fig. 8. JISON configuration file

7.2. Gas Consumption

In the proposed key management framework, the smart contracts are used to register, update and revoke the identity and public key data of vehicles and RSUs. In order to analyze the economic cost of this scheme, this paper uses the Ethereum test network (Rinkeby)

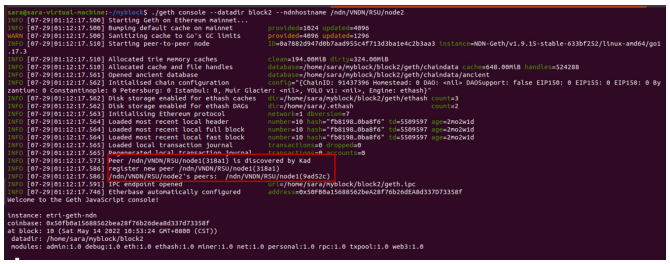


Fig. 9. Peer-to-peer host configuration NDN-blockchain



Fig. 10. Router list of the NDN host

to compile and deploy smart contracts. Rinkeby is an official test network provided by Ethereum. This test network can achieve the same functions as the real Ethereum network. The experiment was performed on July 2022. The gas price was set to 1 Gwei, where 1 Gwei = 0.000000001 Ether and 1 Ether = 1,065 USD. The transactions of the smart contracts are found at https://rinkeby.etherscan.io/ by using the account address of the VTA (0xb42819AEdB1b9928199F9E42202fF492c37072D8). The gas cost analysis of the smart contracts is shown in Table 2.

Table 2. The Cost Performance of The Smart Contracts

Functions	Gas Consumption	Ether's Cost	Cost (USD)
Contract Creation	1323932	0.001323932	1.4099
register_PK	166483	0.000166483	0.1773
update_PK	59910	0.000059910	0.0638
revoke_PK	35088	0.000035088	0.0374

As shown in Table 2, the contract creation is executed just once, and the cost is approximately $1.4099. The cost of the register_PK process is $0.1773, and the cost of the update_PK process is $0.0638. The cost of the revoke_PK process is approximately $0.0374. When the number of users increases, the cost of the computation increases in the process of registration, update, and revocation. The cost of the computation process with various users is shown in Figure 11.

We compare our scheme with the other existing key management schemes in the NDN and VNDN, as shown in Table 3. Some existing key management schemes in the NDN [5,6,7] are centralized schemes that depend on the CA. Other schemes [17,18,19] make use of the benefits provided by the blockchain.

Table 3 shows that the scheme [19] uses the blockchain to update and revoke a public key. However, it creates some problems in the key management because of the tamper-proof feature of the blockchain. Therefore, we use the smart contract to avoid these prob-

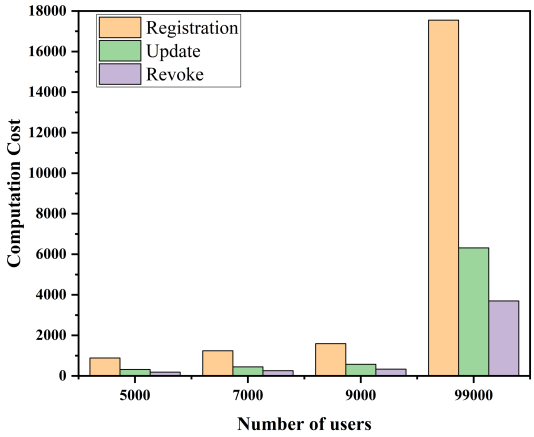


Fig. 11. The computation cost process of the variety of the users

lems by automatically managing the user's public key and identity and presenting the efficient key management in VNDN in an automatic way.

Table 3. Compression with the existing schemes

Existing sSchemes	Structure	Based On	Smart Contract	Key Update	Key Revocation	Authentication	Key Agreement
Chaoyi [4]	Centralized	Trusted Anchor	-	No	No	No	No
A. Albarqi [5]	Centralized	PKI	-	No	Yes	No	No
S. Santesson [6]	Centralized	PKI	-	No	Yes	No	No
Yingdi [7]	Decentralized	Web-of-Trust (WoT)	-	No	Yes	Yes	No
Hao [17]	Decentralized	Blockchain	No	No	No	Yes	No
Junjun [18]	Decentralized	Blockchain	No	No	Yes	Yes	No
Kan [19]	Decentralized	Blockchain	No	Yes	Yes	No	No
Our scheme	Decentralized	NDN with blockchain	Yes	Yes	Yes	Yes	Yes

7.3. Computation Cost

The computation cost of the authentication and key agreement stage in our scheme is analyzed in this section. The operation time of basic cryptographic algorithms used in our scheme is firstly tested in our experimental environment and is shown in Table 4. The notations used in the computation complexity measurement are illustrated in Table 5. We make a comparison between our proposed scheme and the scheme in [24]. The comparison results of the average computation complexity of two different stages in these solutions are shown in Table 6.

In our proposed scheme, the main costs of the authentication stage include the signing time and verification time used in the EC-Schnorr signature algorithm. Thus, the time cost of the authentication stage is $3(T_{sign} + T_{ver})$. In the key agreement stage, two different session keys are generated based on the bivariate polynomial function. The AES-based

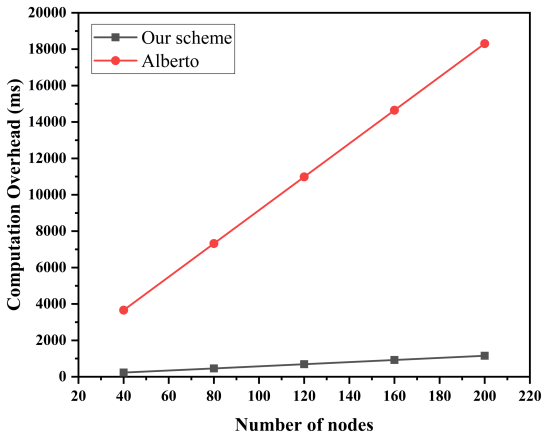


Fig. 12. The computation cost of the authentication stage over the number of nodes

Table 4. The execution times

Operations	Average Time (ms)
EC-Schnorr(signing)	0.144
EC-Schnorr (verification)	1.777
Hash function (SHA 256)	0.008
AES-128 (encryption)	0.019
AES-128 (Decryption)	0.418

Table 5. The Notation Explanation

Notations	Description
T_{sign} and T_{ver}	The average time of the signature and verification of our scheme's signature solution
T_{ssign} and T_{sver}	The average time of the signature and verification of other scheme's signature solution
T_{Senc} and T_{SDec}	The average time of the encryption and decryption of the symmetric cryptography
T_{Biv} and T_K	The average time of the bivariate polynomial and key derivation

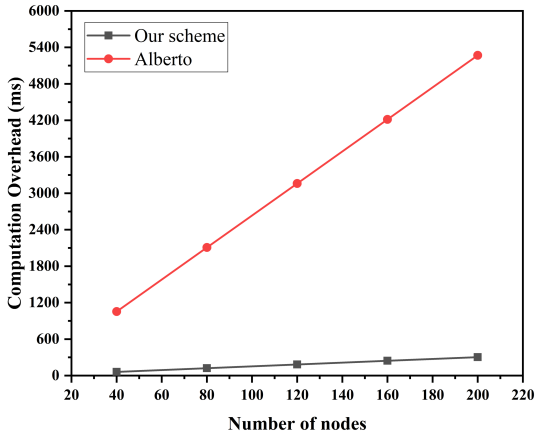


Fig. 13. The computation cost of the key agreement stage over the number of nodes

encryption is used to transmit securely the relevant parameters of the polynomial. Thus, the computation cost of the key agreement contains the cost of the AES algorithm and bivariate polynomial. The total time of the key agreement stage is $T_{Senc} + T_{SDec} + T_{Biv}$. The computation complexity of the authentication stage and key agreement stage over the different number of nodes is respectively shown in Figure 12 and Figure 13. The computation overhead against the DoS attack in the authentication stage is shown in Figure 14. The result in Figure 14 shows our scheme can efficiently mitigate DoS attacks aimed at the cache store by the cause of interest flooding.

Table 6. Computation complexity

Schemes	Authentication Stage Complexity		Key Agreement Stage	Complexity
Our Scheme	$3T_{sign} + 3T_{ver}$	5.765	$3T_{Senc} + 3T_{SDec} + 2T_{Biv}$	1.523
Alberto [24]	$2T_{ssign} + 2T_{sver}$	91.55	$4T_{Senc} + 4T_{SDec} + 2T_K$	26.35

7.4. Communication Cost

The communication cost of this study is evaluated during the authentication stage and key agreement stage in terms of the amount of interest and data packets shared in the network and the number of bytes sent and received by vehicle nodes and RSUs nodes. This thesis depends on IEEE 802.15.4 standard [31], which offers the fundamental lower network layers of a type of wireless personal area network (WPAN) which focuses on low-cost, low-speed ubiquitous communication between devices. The maximum size of the IEEE 802.15.4 frame is 127 bytes, the frame contains three basic fields such that header, payload, and footer. This study considers the 1 +0 encoding which is proposed for

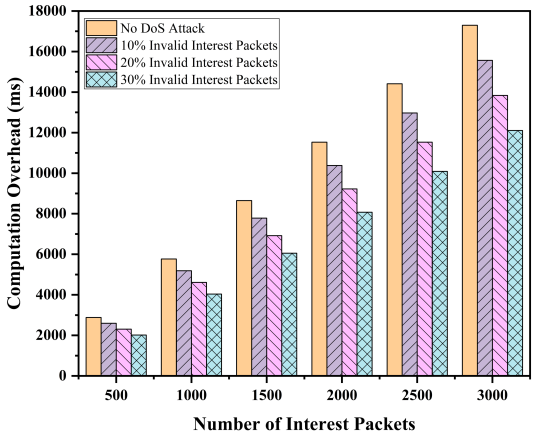


Fig. 14. The computation cost of the authentication protocol over DoS attacks

interest and data packets for NDN networks [32]. The size of the interest and data packets in each filed of the IEEE 802.15.4 frame is shown in Table 7.

Table 7. The size of NDN Interest (I)/Data packets(D)

Field	Field size	I	D
802.15.4 PHY header	6B	✓	✓
802.15.4 MAC header	23B	✓	✓
802.15.4 SEC header	5B	✓	✓
Packet type TL	1B	✓	✓
Name TL	1B	✓	✓
Name component TLVs	SN	✓	✓
Content TLV	1B (TL) +SC		✓
Signature info TL	1B		✓
Signature type TLV	1B (TL) + 1B (V)		✓
Signature value TLV	1B (TL) + 16B (V)		✓
802.15.4 Signature footer	16B	✓	✓
802.15.4 CRC footer	2B	✓	✓

As shown in Table 7, the number of bytes in the header and footer fields are 52 bytes. In the authentication stage, the vehicles who are on the same communication range help in the authentication process between a new vehicle and RSU in the VNDN. The communication between the new vehicle and other vehicles in the network is untrusted, therefore, the number bytes of the header and footer included in the 802.15.4 frame which is exchanged between the new vehicle and other vehicles in the network are 36 bytes because the signature field of the footer will not be included in the frame. On other hand,

the communication between the vehicles on the network who are already on the network and RSU is trusted, therefore, the number bytes of the header and footer included in the 802.15.4 frame which is exchanged between the vehicles and RSU are 52 bytes. According to the Table 7, the cost communication is computed, where SN is the total size of the name values, SC is the total size of the data packet content. In order to compute the cost communication, this study assumes values for the variables which are used in this study, where (ID, PbKname, M, T) are 4 bytes, the hash, signature and encryption are 16 bytes, the prefix name is 1 byte.

Table 8. Number of bytes for interest (I) and data (D) packets for authentication and key agreement stages

Stage	Communication path	Number of bytes (Interest/ Data)
Authentication stage	Vehicle V/ Vehicles Vi	91B (I)/ 92B(D)
	Vehicles Vi/ RSU	107B(I)/ 108B(D)
Key agreement stage	Vehicle V/ Vehicles Vi	79B(I)/ 104B(D)
	Vehicles Vi/ RSU	103B(I)/ 148B(D)

Table 9. The comparison of communication cost

Scheme	Number of bytes transmitted
This study	835 bytes
Alberto [24]	867 bytes

According to the above assumption, the number of bytes for interest (I) and data (D) packets which are transmitted in both authentication and key agreement stages are shown in Table 8. The comparison of cost communication of this study with the Alberto [24] scheme is shown in Table 5.8. As shown in Table 9, the number of bytes that are shared in this system are less than in scheme [24].

8. Conclusion

In this paper, a decentralized key management solution based on blockchain for the VNDN is firstly proposed. Smart contracts are used to manage the identity and public key information of entities in the VNDN. Then, a lightweight mutual authentication scheme between a vehicle V and RSU is proposed. In the key agreement stage, the session keys based on the bivariate polynomial between different entities are generated. Our scheme can prevent the passive attack by adopting the encryption scheme. Our scheme can also prevent the active attack, for example the DoS attack incurred by the interest flooding. In addition, we also explore the security analysis methods in the UC framework. The security requirements of the authentication stage and key agreement stage of the proposed scheme are analyzed in the UC framework. We also implement our scheme in the NDN-

blockchain platform and we experimentally analyze performance of our scheme. We will further improve our scheme in the future work.

Acknowledgments. This work is supported by NSFC No. 61461027; Gansu province science and technology plan project under grant No. 20JR5RA467; Innovation Promotion Education Fund of Ministry of Education No. $2018A05003$; Graduate Fine-designed Course of Lanzhou University of Technology.

References

1. Baofeng Ji, Xueru Zhang, Shahid Mumtaz, Congzheng Han, Chunguo Li, Hong Wen, and Dan Wang. Survey on the internet of vehicles: Network architectures and applications. *IEEE Communications Standards Magazine*, 4(1):34–41, 2020.
2. Alex Afanasyev, Jeff Burke, Tamer Refaei, Lan Wang, Beichuan Zhang, and Lixia Zhang. A brief introduction to named data networking. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)*, pages 1–6. IEEE, 2018.
3. Chaker Abdelaziz Kerrche, Farhan Ahmad, Mohamed Elhoseny, Asma Adnane, Zeeshan Ahmad, and Boubakr Nour. Internet of vehicles over named data networking: Current status and future challenges. *Emerging Technologies for Connected Internet of Vehicles and Intelligent Transportation System Networks: Emerging Technologies for Connected and Smart Vehicles*, pages 83–99, 2020.
4. Chaoyi Bian, Zhenkai Zhu, Alexander Afanasyev, Ersin Uzun, and Lixia Zhang. Deploying key management on ndn testbed. *UCLA, Peking University and PARC, Tech. Rep*, 2013.
5. Aysha Albarqi, Ethar Alzaid, Fatimah Al Ghamdi, Somaya Asiri, Jayaprakash Kar, et al. Public key infrastructure: A survey. *Journal of Information Security*, 6(01):31, 2014.
6. Stefan Santesson, Michael Myers, Rich Ankney, Ambarish Malpani, Slava Galperin, and Dr. Carlisle Adams. X.509 Internet Public Key Infrastructure Online Certificate Status Protocol - OCSP. RFC 6960, June 2013.
7. Yingdi Yu, Alexander Afanasyev, Zhenkai Zhu, and Lixia Zhang. An endorsement-based key management system for decentralized ndn chat application. *University of California, Los Angeles, Tech. Rep. NDN-0023*, 2014.
8. Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Decentralized business review*, page 21260, 2008.
9. Moutaz Alazab, Salah Alhyari, Albara Awajan, and Ayman Bahjat Abdallah. Blockchain technology in supply chain management: an empirical study of the factors affecting user adoption/acceptance. *Cluster Computing*, 24:83–101, 2021.
10. Deepa Pavithran, Khaled Shaalan, Jamal N Al-Karaki, and Amjad Gawanmeh. Towards building a blockchain framework for iot. *Cluster Computing*, 23(3):2089–2103, 2020.
11. Hui Li, Lishuang Pei, Dan Liao, Xiong Wang, Du Xu, and Jian Sun. Bddt: use blockchain to facilitate iot data transactions. *Cluster Computing*, 24:459–473, 2021.
12. Tejasvi Alladi, Vinay Chamola, Nishad Sahu, Vishnu Venkatesh, Adit Goyal, and Mohsen Guizani. A comprehensive survey on the applications of blockchain for securing vehicular networks. *IEEE Communications Surveys & Tutorials*, 2022.
13. Ahmed Elkhalil, Jiashu Zhang, and Rashad Elhabob. An efficient heterogeneous blockchain-based online/offline signcryption systems for internet of vehicles. *Cluster Computing*, pages 1–18, 2021.
14. Khizra Asaf, Rana Asif Rehman, and Byung-Seo Kim. Blockchain technology in named data networks: A detailed survey. *Journal of Network and Computer Applications*, 171:102840, 2020.

15. Quang Tung Thai, Namseok Ko, Sung Hyuk Byun, and Sun-Me Kim. Design and implementation of ndn-based ethereum blockchain. *Journal of Network and Computer Applications*, 200:103329, 2022.
16. Hakima Khelifi, Senlin Luo, Boubakr Nour, Hassine Moungla, Syed Hassan Ahmed, and Mohsen Guizani. A blockchain-based architecture for secure vehicular named data networks. *Computers & Electrical Engineering*, 86:106715, 2020.
17. Hao Liu, Rongbo Zhu, Jun Wang, and Wengang Xu. Blockchain-based key management and green routing scheme for vehicular named data networking. *Security and Communication Networks*, 2021:1–13, 2021.
18. Junjun Lou, Qichao Zhang, Zhuyun Qi, and Kai Lei. A blockchain-based key management scheme for named data networking. In *2018 1st IEEE international conference on hot information-centric networking (HotICN)*, pages 141–146. IEEE, 2018.
19. Kan Yang, Jobin J Sunny, and Lan Wang. Blockchain-based decentralized public key management for named data networking. In *The international conference on computer communications and networks (ICCCN 2018)*, 2018.
20. Ran Canetti. Universally composable security: A new paradigm for cryptographic protocols. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 136–145. IEEE, 2001.
21. Michel Abdalla, Dario Catalano, Céline Chevalier, and David Pointcheval. Efficient two-party password-based key exchange protocols in the uc framework. In *Cryptographers' Track at the RSA Conference*, pages 335–351. Springer, 2008.
22. Zibin Zheng, Shaoan Xie, Hong-Ning Dai, Weili Chen, Xiangping Chen, Jian Weng, and Muhammad Imran. An overview on smart contracts: Challenges, advances and platforms. *Future Generation Computer Systems*, 105:475–491, 2020.
23. Anhao Xiang and Jun Zheng. A lightweight anonymous device authentication scheme for information-centric distribution feeder microgrid. *Computers, Materials & Continua*, 69(2), 2021.
24. Alberto Compagno, Mauro Conti, and Ralph Droms. Onboardicng: a secure protocol for onboarding iot devices in icn. In *Proceedings of the 3rd ACM Conference on Information-Centric Networking*, pages 166–175, 2016.
25. Xian Guo, Yuxi Chen, Laicheng Cao, Di Zhang, and Yongbo Jiang. A receiver-forwarding decision scheme based on bayesian for ndn-vanet. *China Communications*, 17(8):106–120, 2020.
26. Xian Guo, Baobao Wang, Yongbo Jiang, Di Zhang, and Laicheng Cao. Homomorphic encryption based privacy-aware intelligent forwarding mechanism for ndn-vanet. *Computer Science and Information Systems*, 20(1):1–24, 2023.
27. Carlo Blundo, Alfredo De Santis, Amir Herzberg, Shay Kutten, Ugo Vaccaro, and Moti Yung. Perfectly-secure key distribution for dynamic conferences. In *Annual international cryptology conference*, pages 471–486. Springer, 1992.
28. Junji Takemasa, Yuki Koizumi, and Toru Hasegawa. Data prefetch for fast ndn software routers based on hash table-based forwarding tables. *Computer Networks*, 173:107188, 2020.
29. Dylan Yaga, Peter Mell, Nik Roby, and Karen Scarfone. Blockchain technology overview. *arXiv preprint arXiv:1906.11078*, 2019.
30. Carlo Blundo, Alfredo De Santis, Luisa Gargano, and Ugo Vaccaro. On the information rate of secret sharing schemes. In *Advances in Cryptology—CRYPTO'92: 12th Annual International Cryptology Conference Santa Barbara, California, USA August 16–20, 1992 Proceedings*, pages 148–167. Springer, 2001.
31. Andreas F Molisch, Kannan Balakrishnan, Chia-Chin Chong, Shahriar Emami, Andrew Fort, Johan Karedal, Juergen Kunisch, Hans Schantz, Ulrich Schuster, and Kai Siwiak. Ieee 802.15. 4a channel model-final report. *IEEE P802*, 15(04):0662, 2004.

32. Marcel Enguehard, Ralph E Droms, and Dario Rossi. On the cost of geographic forwarding for information-centric things. *IEEE Transactions on Green Communications and Networking*, 2(4):1150–1163, 2018.

Xian Guo is a professor in School of Computer and Communication, Lanzhou University of Technology, China. He is a visiting scholar at University of Memphis. He received his PhD at Lanzhou University of Technology in 2011, and he received his BS at Northwest Normal University. His current research interests include cryptography, design and analysis of security protocol, and blockchain. E-mail: iamxg@ 163.com.

Sarah Almadhehagi is a graduate student in School Computer and Communication at Lanzhou University of Technology, China. She received her Bachelor degree from Taiz University of Computer Science, Yemen, in 2017, and started her master studying in 2020. Her research interests are blockchain, network and information security, design and analysis of security protocol, etc. E-mail: sara.almad2015@gmail.com.

Tao Feng is a professor in School of Computer and Communication at Lanzhou University of Technology, China. He received his PhD in Xidian University in 2008. His current research interests include network and information security, design and analysis of security protocol, etc.

Di Zhang is an associate professor in School of Computer and Communication at Lanzhou University of Technology, China. He received MS and PhD in Communication University of China, China, in 2013 and 2016, respectively. His current research interests include network and information security, and blockchain etc.

Yongbo Jiang is an associate professor in School of Computer and Communication at Lanzhou University of Technology, China. He received his PhD in Xidian University in 2013. His current research interests include network and information security, and Information-Centric Networking etc.

Junli Fang is a lecturer in School of Computer and Communication at Lanzhou University of Technology, China. She received her MS in Beijing Jiaotong University, China, in 2011. His current research interests include cryptography and blockchain, etc.

Received: March 26, 2023; Accepted: January 10, 2024.

Computer Science and Information Systems 21(3):759–780 https://doi.org/10.2298/CSIS230923018P

SSEPC Cloud: Carbon Footprint Aware Power Efficient Virtual Machine Placement in Cloud Milieu

Bivasa Ranjan Parida[1,⋆], Amiya Kumar Rath[1,2], Bibudhendu Pati[3], Chhabi Rani Panigrahi[3], Hitesh Mohapatra[4], Tien-Hsiung Weng[5,⋆], and Rajkumar Buyya[6]

[1] Department of Computer Science and Engineering, Veer Surendra Sai University of Technology, Burla, Odisha, India
bivasa.parida@silicon.ac.in
[2] Vice-Chancellor, Biju Patnaik University of Technology, Rourkela, Odisha, India
akrath_cse@vssut.ac.in
[3] Department of Computer Science, Rama Devi Women's University, Bhubaneswar, Odisha, India
patibibudhendu@rdwu.ac.in
panigrahichhabi@gmail.com
[4] School of Computer Engineering, KIIT (Deemed to be) University, Bhubaneswar, Odisha, India
hiteshmahapatra@gmail.com
[5] Department of Computer Science and Information Engineering, Providence University, Taiwan
thweng@gm.pu.edu.tw
[6] Cloud Computing and Distributed Systems (CLOUDS) Lab, School of Computing and Information Systems, University of Melbourne, Australia
rbuyya@unimelb.edu.au

Abstract. The consumption of energy and carbon emission in cloud datacenters are the alarming issues in recent times, while optimizing the average response time and service level agreement (SLA) violations. Handful of researches have been conducted in these domains during virtual machine placement (VMP) in cloud milieu. Moreover it is hard to find researches on VMP considering the cloud regions and the availability zones along with the datacenters, although both of them play significant roles in VMP. Hence, we have worked on a novel approach to propose a hybrid metaheuristic technique combining the salp swarm optimization and emperor penguins colony algorithm, i.e. SSEPC to place the virtual machines in the most suitable regions, availability zones, datacenters, and servers in a cloud environment, while optimizing the mentioned quality of service parameters. Our suggested technique is compared with some of the contemporary hybrid algorithms in this direction like Sine Cosine Algorithm and Salp Swarm Algorithm (SCA-SSA), Genetic Algorithm and Tabu-search Algorithm (GATA), and Order Exchange & Migration algorithm and Ant Colony System algorithm (OEMACS) to test its efficacy. It is found that the proposed SSEPC is consuming 4.4%, 8.2%, and 16.6% less energy and emitting 28.8%, 32.83%, and 37.45% less carbon, whereas reducing the average response time by 11.43%, 18.57%, and 26% as compared to its counterparts GATA, OEMACS, and SCA-SSA respectively. In case of SLA violations, SSEPC has shown its effectiveness by lessening the value of this parameter by 0.4%, 1.2%, and 2.8% as compared to SCA-SSA, GATA, and OEMACS respectively.

Keywords: Virtual Machine Placement, Energy Consumption, Carbon Emission, Salp Swarm Optimization, Emperor Penguins Colony Algorithm.

⋆ Corresponding authors

1. Introduction

Cloud computing has proved itself as an indispensable part of technologies in recent times being a utility as pay-per-use model to deliver the resources to the cloud users from the datacenters virtually [1]. The significant cost of expense in the datacenters is the cost of energy consumption [2]. In 2020, the datacenters globally used 1% of world electricity [3]. If the remedial measures have not taken and the technological trends remain unchanged for the coming decade, then between 2016 and 2030, the energy consumption by the datacenters may exceed with 12% approx. [4]. Some countries also enforce tax on carbon footprint for the environmental sustainability [5]. Carbon footprint can be reduced by implementing more and more renewable energy in cloud datacenters [6]. Due to the erratic availability of renewable energy round the clock, the hybrid model of power supply including the brown energy obtained from the fossil fuels along with the green energy from the renewable sources are used in the cloud datacenters [6]. Hence, the consumption of energy and carbon emission in cloud datacenters are the genuine issues to deal with, while optimizing the average response time and service level agreement (SLA) violations for better quality of service (QoS). The strategy of virtual machine placement (VMP) directly influences the consumption of energy and carbon footprints of a datacenter [7] along with the mentioned QoS parameters.

1.1. Motivation and Contributions

Although a handful of researches have been conducted to minimize the energy consumption and carbon emission in cloud datacenters, efficient methodologies are still scarce. Energy consumption by the networks and the cooling system in a datacenter are ignored in most of the research works. Carbon footprint estimation and minimization is a matter of concern for maintaining the environmental sustainability. Addressing these issues through VMP in the cloud datacenters while maintaining the average response time along with SLA violations, are the motivating factors behind this research. Although availability zones are purely logical entities in the cloud regions [8], both of them play significant roles in mapping the requests onto the physical servers in the datacenters. The user account in service provider's domain determines the list of cloud regions accessible to that user by default. Some other regions may be enabled by the user and some others may be restricted depending on global policies. If the region selection is not done by the user, then the process is automated by the service provider through certain searching mechanisms to allot a default region [9]. Multiple availability zones as the isolated locations exist in each region. Expansion of an availability zone may be constrained due to the increasing number of zones in a region. So, for the new users, the constrained availability zones may be removed from the list. Two different users may access different set of availability zones while availing services from the same cloud region. If user doesn't insist on an availability zone, then the service provider selects one zone for the user. It is always recommended for the initial request of a user to continue with the default availability zone selected by the service provider through certain searching mechanisms [9]. It is hard to find works conducted so far for VMP considering cloud regions and availability zones along with datacenters and the servers to the best of our knowledge. Hence, we are motivated to contribute a little to this vast field of emerging aspect of research in cloud milieu. The following points are the key contributions of this work.

- This proposed methodology introduces a hybrid nature inspired optimization technique to solve the problem of VMP in the cloud milieu in two phases. In the first phase, the Salp Swarm Algorithm (SSA) [10] is used to explore geographically in the global search space to find out the suitable cloud region and further converge to the most appropriate availability zone with the aim to make the best use of green energy. Then it exploits to find out the best datacenter possible for VMP with the availability of maximum green energy. The Emperor Penguins Colony (EPC) algorithm [11] enriched with the potential of local search has been implemented in the second phase, to trace the most suitable server in the datacenter for placing the virtual machines (VMs). Both the algorithms are hybridized in the proposed algorithm Salp Swarm based Emperor Penguins Colony (SSEPC) for overcoming the limitations of each other to establish a robust methodology.
- SSEPC places the VMs appropriately to minimize energy consumption and carbon emission of a datacenter with respect to average response time and SLA violations.
- Proposed VMP model considers the energy consumption due to servers, networks and the cooling system of the datacenters simultaneously to make the approach more realistic in the cloud environment.
- Unlike other existing VMP proposals, this model portrays both initial as well as the runtime VMP in the cloud milieu considering the priority of user requests.

The rest of the paper is organized as follows. Section 2 reviews the literature of VMP, consumption of energy, and carbon emission in cloud environment along with the QoS parameters. The system model along with the mathematical formulations are elucidated in Section 3. The proposed algorithm is explored in the Section 4 and Section 5 deliberates the simulation environment, as well as analyses the experiments conducted in heterogeneous cloud environment and the obtained results. Finally, the conclusive remarks along with the future directions in the research in this domain are illuminated in the Section 6.

2. Related Work

The process of allocating a VM to a server in the datacenter to avail the required resources to host the application requested by the user is called as VMP [12]. From the existing works related to VMP in the cloud milieu, it has been observed that although works have been conducted on inter datacenter or multi datacenter VMP [2], [6], [7], [15], [16] considering the assignment of VMs to the appropriate datacenters followed by mapping onto the suitable servers, we didn't come across any VMP proposal considering mapping of VMs onto the compatible cloud regions and the availability zones. Technically a cloud region contains multiple availability zones, an availability zone may be a collection of single or many datacenters, and a datacenter physically holds huge number of servers [8].

The energy consumption by a datacenter may be broadly categorized into two types; i.e. (i) The consumption of energy by the IT equipment like servers and networks, (ii) The usage of energy by the non-IT equipment like cooling system [2]. Hence, none of the components can be ignored while designing an energy conservation model [2]. Minimization of carbon emission in the datacenters is becoming a necessity now-a-days, as carbon taxes are imposed in some countries and in others carbon footprints are taken seriously for the environmental sustainability [5]. The response time [6], [7], [17] and SLA

violations [6], [7], [14], [15], [16], [17], [19] play vital role in cloud milieu to meet the quality of experience (QoE) of the users during VMP in cloud datacenters.

Table 1 compares the proposed work SSEPC with the other relevant proposals of recent times in terms of different performance metrics like datacenter energy consumption based on server, networks, and cooling equipment, carbon emission, response time, SLA violations, and use of renewable energy sources for environmental sustainability. Hybridizing multiple methodologies helps in overcoming the limitations of one another [24]. The hybridization is one of the most suitable strategies accepted widely for promoting diversity in the search space to achieve global optimum [24]. Hence, we have hybridized two recently proposed metaheuristic techniques SSA (2017) [10] and EPC (2019) [11] to search the global space for VMP. Table 2 analyses some existing hybrid algorithms for VMP in cloud domain for comparison with the proposed SSEPC methodology.

Table 1. Analogy of related works with proposed work in terms of various metrics

Proposed By	**VMP Implemented**	**Hybrid Approach**	**Server Based Energy Consumption**	**Network Based Energy Consumption**	**Cooling Based Energy Consumption**	**Carbon Emission**	**Response Time**	**SLA Violation**	**Use of Green Energy**
Feng et al. [2]	✓	✓	✓	✓	✓				
Xu et al. [6]	✓		✓		✓	✓	✓	✓	✓
Abbasi-khazaei et al. [7]	✓		✓			✓	✓	✓	✓
Liu et al. [12]	✓	✓	✓						
Zhao et al. [13]	✓	✓	✓						
Samriya et al. [14]	✓		✓					✓	
Gharehpasha et al. [15]	✓	✓	✓					✓	
Zhao et al. [16]	✓	✓	✓			✓		✓	✓
Xu et al. [17]	✓		✓				✓	✓	
Tang et al. [18]	✓	✓	✓	✓					
Yadav et al. [19]	✓		✓	✓				✓	
Khodayarseresht et al. [20]	✓		✓			✓			
Belabed et al. [21]	✓			✓					
Justafort et al. [22]	✓	✓	✓	✓	✓	✓			
Anusooya et al. [23]		✓	✓			✓			
Proposed Algorithm	✓	✓	✓	✓	✓	✓	✓	✓	✓

3. System Model

This section elaborates the system components and the system parameters with mathematical elucidations for designing the proposed model for VMP to satisfy the QoS metrics. Figure 1 portrays the system model of the proposed methodology.

Table 2. Existing hybrid algorithms of VMP considered for comparison

Authors	**Name of the Hybrid Algorithm**	**Algorithms involved in Hybridization**	**Purpose of Hybridization**
Liu et al. [12]	OEMACS	Order Exchange and Migration algorithm and Ant Colony System algorithm	OEM swaps VMs between servers locally, whereas ACS allocates VMs in minimum number of servers through global search
Zhao et al. [13]	GATA	Genetic Algorithm and Tabu-search Algorithm	The algorithm of tabu-search works like a mutation operator of genetic algorithm for improving its local search ability
Gharehpasha et al. [15]	SCA-SSA	Sine Cosine Algorithm and Salp Swarm Algorithm	SCA explores and exploits the space of searching for finding the optimal solution, whereas SSA manages leader and followers in the population to reach at the optimal solution

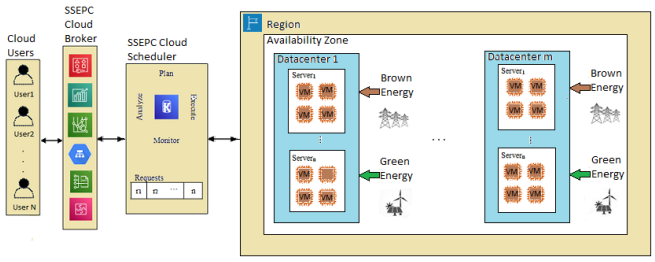


Fig. 1. SSEPC based Cloud System Model for Virtual Machine Placement

3.1. System Components

The cloud service providers, cloud brokers, and cloud users are the entities of a cloud system. The infrastructure of the cloud is built around physically separated regions in different geographic locations [8]. Regions contain multiple availability zones to ensure resiliency [25]. Each availability zone consists of one or more datacenters [8]. A datacenter contains $50,000$ to $80,000$ physical servers [26]. Cloud broker works like an interface for the cloud users to communicate with the service provider. Here, the SSEPC cloud broker plays the major role of routing the user requests to the most suitable resources for the minimization of the consumption of energy and emission of the carbon footprints while maximizing the renewable energy utilization, in such a way that the average response time and SLA violations remain as minimum as possible. The process of scheduling the requests to the servers of the different datacenters belongs to the MAPE-K [27] loop in Figure 1. This loop consists of different phases, such as Monitor, Analyse, Plan, and Execute, managed by the Knowledge module. Lower the number of rejection of the requests to create VMs, higher is the QoE of the cloud users.

The set of u cloud regions is represented as a set $CR = \{cr_1, cr_2, cr_3, \ldots, cr_u\}$, where each $cr_x \in CR$, such that $1 \leq x \leq u$. A region contains v availability zones as a

set $AZ = \{az_1, az_2, az_3, \ldots, az_v\}$, where each $az_y \in AZ$, such that $1 \leq y \leq v$. The set of m datacenters in an availability zone is represented as a set $D = \{d_1, d_2, d_3, \ldots, d_m\}$, where each $d_i \in D$, such that $1 \leq i \leq m$. Each datacenter has n number of physical servers, i.e., $S = \{s_1, s_2, s_3, \ldots, s_n\}$, where $s_j \in S$, such that $1 \leq j \leq n$. Each server possesses p number of VMs portrayed as set $VM = \{vm_1, vm_2, vm_3, \ldots, vm_p\}$, where $vm_k \in VM$, such that $1 \leq k \leq p$. R is the set of requests, i.e., $R = \{r_1, r_2, r_3, \ldots, r_q\}$, where $r_l \in R$, such that $1 \leq l \leq q$. Hence, the relationship between the cloud regions, availability zones, datacenters, servers, and the VMs can be defined as follows.

$\{\{\{\{\{vm_1, vm_2, \ldots, vm_p\}, s_2, \ldots, s_n\}, d_2, \ldots, d_m\}, az_2, \ldots, az_v\}, cr_2, \ldots, cr_u\}$

3.2. System Parameters

This subsection discusses the system parameters mentioned as follows.

3.2.1 Energy Consumption of Datacenters The components mainly responsible for the consumption of energy in a datacenter are the servers, networks, and cooling system [2]. Hence, energy consumed by all the datacenters E_D can be illustrated as in Eqn. (1).

$$E_D = \sum_{i=1}^{m} E_{d_i} = \sum_{i=1}^{m} (E_{i,Server} + E_{i,Network} + E_{i,Cooling}) \tag{1}$$

Here, the energy consumed in total by a datacenter is E_{d_i}, where $E_{i,Server}$, $E_{i,Network}$, and $E_{i,Cooling}$ are the amount of energy consumption by the servers, networks, and the cooling system of a datacenter respectively.

3.2.1.1 Server Energy Consumption Model The energy consumed by all the servers of a datacenter d_i is denoted as $E_{i,Server}$, which is given [28] as in Eqn. (2).

$$E_{i,Server} = \sum_{j=1}^{n} E_{i,s_j} = \sum_{j=1}^{n} (E_{j,Idle} + (E_{j,Max} - E_{j,Idle}) \times u_{s_j}) \tag{2}$$

Here, consumption of energy by a server s_j in a datacenter d_i can be represented as E_{i,s_j}, where $E_{j,Idle}$ and $E_{j,Max}$ are the energy consumed by the server s_j at idle time and peak time respectively, u_{s_j} is the utilization of jth server.

3.2.1.2 Network Energy Consumption Model Let the communication graph of VMs be $G_C(VM, L_C)$. Here, VM represents the set of VMs and L_C stands for the set of edges or links between the VMs for communication among themselves, such that $v_1, v_2 \in VM$ and edge$(v_1, v_2) \in L_C$. The bandwidth required for this communication is $bw(v_1, v_2)$. Let the datacenter network topology graph be $G_T(S, L_T)$. Here, S represents the set of servers and L_T stands for the set of edges or links between servers, such that $s_1, s_2 \in S$ and edge$(s_1, s_2) \in L_T$, only if s_1 links to s_2 in network topology. The bandwidth associated with each link can be represented as $bw(s_1, s_2)$. Let f_v be the function to determine the result of VMP, i.e., $s_1 = f_v(v_1)$, if and only if the VM v_1 is placed on the server s_1. Let the length of shortest path between v_1 and v_2 be $P(v_1, v_2)$. The consumption of energy by the network of a datacenter d_i is denoted as $E_{i,Network}$, which can be formulated [2] as in Eqn. (3).

$$E_{i,Network} = \sum_{(v_1,v_2)\in L_C} bw(v_1,v_2) \times P(v_1,v_2) \tag{3}$$

3.2.1.3 Cooling Energy Consumption Model The cooling system of a datacenter is called as Computer Room Air Conditioning (CRAC) [2]. Let the maximum temperature of air flow supplied by CRAC is $T_{Sup(CRAC)}$. It is inversely proportional to the energy consumption of cooling $E_{i,Cooling}$. Wang et al. [29] found that up to $4.3 - 9.8\%$ energy can be conserved per 1°C rise in $T_{Sup(CRAC)}$. The cooling system energy consumption $E_{i,Cooling}$ can be calculated based on the proposed model of Tang et al. [30] and is given in Eqn. (4).

$$E_{i,Cooling} = \frac{E_{i,Server} + E_{i,Network}}{0.0068 \times (T_{Sup(CRAC)})^2 + 0.008 \times T_{Sup(CRAC)} + 0.458} \tag{4}$$

3.2.2 Renewable Energy Availability Cloud service providers are swiftly moving towards the renewable energy sources to derive electricity for the datacenters to reduce the utilization of brown energy due to its high energy cost, carbon footprint cost and alarming negative impact on environment. Although most of the renewable energy sources are intermittent in nature, this proposed technique considers the use of renewable or green energy (E^G) in cloud datacenters with highest priority as compared to the brown energy (E^B). Datacenters will utilize the renewable energy first as long as it is available before using the brown energy.

3.2.3 Carbon Emission of Datacenters The approach for minimizing the carbon emission in datacenters using SSEPC is based on the carbon intensity value instead of relying on weather data for VMP, as weather data may be an inaccurate metric. According to National Grid Electricity System Operator, carbon intensity is the number of grams of CO2 emitted per unit of electricity at one kilowatt hour [31]. Carbon intensity value is based on the type of brown energy, where this value is zero for green energy [7]. If the energy required by the datacenter E_{d_i} is greater than the green energy $E^G_{d_i}$ available in that datacenter, then the brown energy $E^B_{d_i}$ usage will have a positive value. On the contrary, if the available green energy is more than the energy requirement of the datacenter, then the brown energy usage value will be zero. Hence, the total carbon emission [6] from all the datacenters C_D is defined as in Eqn. (5).

$$C_D = \sum_{i=1}^{m} C_{d_i} = \sum_{i=1}^{m} (E^B_{d_i} \times I^C_{d_i}) = \sum_{i=1}^{m} (max(E_{d_i} - E^G_{d_i}, 0) \times I^C_{d_i}) \tag{5}$$

Where C_{d_i} is the total carbon emission of the datacenter d_i and $I^C_{d_i}$ is the carbon intensity of datacenter d_i.

3.2.4 Response Time For the better QoE of the cloud users, response time plays a vital role along with other parameters. To make the proposed model realistic one, the response time of the requests is considered along with other parameters. Let the response

time for request r_l allocated to d_i is T_{l,d_i}. Due to insufficient green energy, if the request is forwarded to another datacenter $d_i^{'}$, then the response time will be $T_{d_i,d_i^{'}}$. Hence, the response time of r_l is $T_{r_l} = T_{l,d_i} + T_{d_i,d_i^{'}}$. The average response time of all requests T_D can be calculated as in Eqn. (6).

$$T_D = \frac{1}{|R|} \sum_{r_l \in R} T_{r_l} \tag{6}$$

3.2.5 Total Cost of Execution The purpose of designing the proposed model is to minimize the consumption of energy and emission of carbon footprints in a cloud datacenter, while optimizing the average response time. Hence, the total cost of execution of the requests in cloud datacenters C_T can be composed of the total energy consumed by the datacenters E_D, the total carbon emission from the datacenters C_D, and the average response time of all requests in the system T_D and is given in Eqn. (7).

$$C_T = E_D + C_D + T_D \tag{7}$$

3.2.6 Capacity of a Server The capacity of a server s_j in a datacenter d_i can be represented as Cap_{i,s_j} in Eqn. (8) [32].

$$Cap_{i,s_j} = \sum_{k=1}^{p} Cap_{j,vm_k} = \sum_{k=1}^{p} (num(PEs)_{vm_k} \times MIPS(PEs)_{vm_k} + BW_{vm_k}) \tag{8}$$

Where the capacity of a VM vm_k in a server s_j is Cap_{j,vm_k}, $num(PEs)_{vm_k}$ is denoted as the count of processing elements in k^{th} VM, $MIPS(PEs)_{vm_k}$ stands for the million instructions per second of all the processing elements of kth VM, and BW_{vm_k} represents the bandwidth of kth VM.

3.2.7 SLA Violations SLA violations in datacenters arise due to several reasons. Overuse of a server or the resources requested by a VM not allocated by the server etc. may be the reasons of SLA violation related to our current context. Minimization of SLA violation is required for better QoE of the users. The overall SLA violation value SLA_{vv} is calculated as in Eqn. (9) [14].

$$SLA_{vv} = \frac{MIPS_{TR} - MIPS_{TA}}{MIPS_{TR}} \tag{9}$$

Here, $MIPS_{TR}$ is the total million instructions per second requested and $MIPS_{TA}$ is the total million instructions per second assigned to VMs on the basis of user demands.

3.2.8 Total Optimum Value On regular intervals, the proposed algorithm calculates the values of energy consumption, SLA violations, and capacity of the servers of the datacenter for the VMP. The value of energy consumption of a server s_j in a datacenter d_i is calculated in Eq.(2) as E_{i,s_j} and SLA violation value SLA_{vv} is calculated as in Eqn. (9).

$$O_T = E_{i,s_j} + SLA_{vv} \tag{10}$$

Here, O_T is the total optimum value, which is the combination of energy consumption value of a server E_{i,s_j} and the SLA violation value SLA_{vv}.

4. Salp Swarm optimization based Emperor Penguins Colony (SSEPC) Algorithm for Virtual Machine Placement

This section elaborates the proposed SSEPC algorithm. It is of multi-stage collaborative hybrid type [24] comprising of two phases, where SSA performs global search in Phase-I of SSEPC and the EPC algorithm does the local search in SSEPC Phase-II. Both SSA and EPC are swarm inspired stochastic collective algorithms [10].

4.1. SSEPC Phase-I based on the Salp Swarm Algorithm

In SSEPC Phase-I, the VMs are considered as the salps in the population foraging the food sources, where the computing resources in the datacenters mimic the foods.The salps compete among themselves to discover the best source of food in the ocean. Similarly, the VMs also compete during VMP for availing the most suitable regions, availability zones, and the datacenters with respect to maximizing the green energy utilization with the objective of minimization of consumption of energy, emission of carbon footprints, and average response time.

According to Eqn. (11) [10], the position of the leader salp in the salp chain foraging the food sources is updated.

$$X_j^1 = \begin{cases} F_j + r_1\big((UB_j - LB_j)r_2 + LB_j\big), & r_3 \geq 0, \\ F_j - r_1\big((UB_j - LB_j)r_2 + LB_j\big), & r_3 < 0. \end{cases} \tag{11}$$

Where the leader's position at j^{th} dimension is represented as X_j^1 and F_j portrays the location of destined resource in the same dimension j. The upper bound as well as the lower bound of jth dimension are portrayed as UB_j and LB_j respectively. r_2 and r_3 are the random numbers generated evenly between 0 and 1. The r_1 is also a random value, which acts as the significant factor for trading off between exploration and exploitation. The Eqn. (12) is used to calculate r_1 [10].

$$r_1 = 2e^{\left(-\frac{4t}{T_{max}}\right)^2} \tag{12}$$

Where T_{max} stands for the number of iterations at max and the iteration at an instant is mimicked by t. With the increasing values of the number of iterations, the parameter r_1 goes down. In this way it becomes capable of maintaining the uniformity among the exploration and the exploitation.

The Newton's law of motion inspires the Eqn. (13) to update the position of the follower VMs [10].

$$X_j^i = \frac{(X_j^i + X_j^{i-1})}{2} \tag{13}$$

Where i^{th} follower's position in j^{th} dimension is X_j^i, such that $i \geq 2$. The SSA is meant for providing solutions to the optimization problems of continuous nature. But VMP in cloud computing is of binary characteristic, as the VMs migrate in a restricted fashion confined to 0 or 1. Hence, the generated solutions are required to be transformed into discrete values. Transfer function (TF) is a convenient way for the transformation of continuous solutions into the discrete ones [33]. In our work, we have used sigmoid transfer function for this purpose in Eqn. (14).

$$S(X_j^{i+1}) = sig(X_j^{i+1}) = \frac{1}{1 + e^{-X_j^{i+1}}} \tag{14}$$

Where

$$X_j^{i+1} = \begin{cases} 1, & if S(X_j^{i+1}) > \text{ran.nextInt(2)}, \\ 0, & otherwise. \end{cases}$$

Computational overhead for searching the global space significantly decreases due to the implementation of salp swarm optimization in the first phase, as the requests from the same user follow the path of the first request without wasting time for further searching. The objective is to search the appropriate datacenter in an availability zone of a particular region of the cloud milieu for VMP by minimizing the total cost as per the Eqn. (7) in the first phase. Hence, the multi objective fitness function of the Phase-I of proposed SSEPC, F_{val_1} is as in Eqn. (15).

$$F_{val_1} = min.(C_T) \tag{15}$$

4.2. SSEPC Phase-II based on the Emperor Penguins Colony Algorithm

The SSEPC methodology implements EPC algorithm in its second phase for VMP to search the most suitable server in a datacenter locally. Here, the servers mimic the herd of the penguins and the VMs imitate the penguins. List of VM requests (penguins) approaching the datacenter is collected. They search for the appropriate server (huddle). A penguin intends to be warmed goes to a nearby huddle randomly. Likewise a VM is allotted to a server in the datacenter arbitrarily. If the huddle has maximum number of penguins, then the new penguin arrived recently will remain at the periphery and can't be warmed enough. So this penguin will again move to another huddle having less than maximum number of penguins to be warmed appropriately. But if a huddle contains very less number of penguins, then the penguins loss a lot of energy and enough heat can't be generated in that huddle. So, those penguins prefer to move to an existing huddle with sufficient capacity to be accommodated, by dissolving their huddle. Similarly, a new VM migrates to an appropriate server accommodating less than maximum VMs, as it can't be allocated with resources by a server already reached to its maximum capacity. Likewise, if a server is allocated with very less number of VMs, it lets the VMs to migrate to another server and goes to sleep mode to conserve energy. This is the process of VM migration in running VMP.

The capacity of a server s_j at an instance of time should be within a threshold value depending on the physical resource availability of the server to allow a new VM to be allocated to that server. If Cap_{i,s_j} will go beyond the upper bound threshold value (Th_{UB}) after adding the new VM, then instead of allocating resources to that VM, the server will

forward it to the VM migration list. Similarly, if Cap_{i,s_j} will fall below the lower bound threshold value (Th_{LB}) in a server, then the allocated VMs will be added to the VM migration list and the server will go to sleep mode to conserve energy. According to Xu et al. [6], sleep mode is for 0% server utilization, idle mode is when server utilization remains less than or equal to 50% and 100% server utilization denotes the maximum. Hence, the servers being under utilized are sent to sleep mode after migrating their allocated tasks help in conserving energy and reducing computational cost.

On the basis of priority, the VMs in the VM migration list will be allocated to other suitable servers. The priority has been calculated by using priority-aware VM allocation (PAVA) algorithm [34] and set to a binary value, i.e. either critical or not. The objective is to minimize O_T, while satisfying the server capacity. A multi objective fitness function F_{val_2} can be considered for VMP inside a datacenter using the Phase-II of proposed SSEPC algorithm and is given in Eqn. (16).

$$F_{val_2} = \underset{s.t.\ Th_{LB} \leq Cap_{i,s_j} \leq Th_{UB}}{min.} O_T \tag{16}$$

4.3. Proposed SSEPC Algorithm

The proposed SSEPC for VMP is presented in the Algorithm 1. From a user, the first request to search an appropriate cloud region, followed by the availability zone and then the most suitable datacenter, is treated as the leader salp. Then the other requests from the same user behave like follower salps on the basis of updated data received from the leader, as proposed in the Phase-I of SSEPC. Algorithm 2, Algorithm 3, and Algorithm 4 are the various modules of Phase-I to search the cloud regions, availability zones, and the datacenters respectively. Algorithm 5 represents the VM migration during running VMP among the servers of a datacenter based on EPC algorithm, which is implemented in the Phase-II of proposed SSEPC. The Algorithm 1 invokes Algorithm 2, Algorithm 3, Algorithm 4, and Algorithm 5 as different modules as and when required.

Algorithm 2 deliberates the pseudo code for searching the appropriate cloud region cr_x. Each region is a distinguished geographic area of the cloud datacenters on earth, which is independent of the other regions. The pseudo code for searching the appropriate availability zone az_y is given in Algorithm 3. Here, our proposed algorithm does the exploration to find out the suitable region and the availability zone with utmost care for the maximization of green energy utilization and minimization of overall response time.

Algorithm 4 represents the pseudo code for searching the appropriate datacenter d_i. These datacenters are the physical infrastructures of the cloud providers. The process of exploitation in the first phase of SSEPC is executed to find out the appropriate datacenter for VMP. The VM request from the user reached at the availability zone is mapped to the datacenter having maximum amount of green energy, provided the average response time should not exceed threshold value. If sufficient green energy is not available in any of the datacenters, then the brown energy is taken into consideration. A list is maintained and updated on regular intervals to monitor the consumption of energy by the datacenters. The list contains the datacenters in increasing order of energy consumption. The datacenter consuming minimum energy is chosen for VMP, provided the average response time should not exceed the threshold value. Else the datacenter has been chosen based on the distance of the user from the datacenter to minimize the response time.

Algorithm 5 portrays the second phase of proposed SSEPC on the basis of emperor penguins colony algorithm to search the most suitable server s_j in a datacenter to host the virtual machine. Servers accommodate the VMs till they have not reached the upper bound threshold value of a server's capacity. Servers having less number of VMs, where server capacity falls below the lower bound, let the VMs migrate to other servers and go to sleep mode for conserving energy. Even the servers crossing the upper bound of their capacity while accommodating the VMs, let the excess VMs migrate to other servers to maintain the balance of load on each server with respect to their capacity. In this way we have implemented the VM migration technique for considering the running VMP along with the initial VMP implemented in the first phase of SSEPC. In both the phases of SSEPC, we have considered the priority of user requests in VMP for a better QoE.

The time to find the suitable cloud region among the set of u regions is $\Theta(ulogu)$. Similarly, the time for finding an appropriate availability zone in a cloud region out of the v numbers of availability zones is $\Theta(vlogv)$. The time complexity of searching a datacenter in an availability zone and then searching a server in a particular datacenter are $\Theta(mlogm)$ and $\Theta(nlogn)$ respectively due to the m number of datacenters in an availability zone and n number of servers in a datacenter. Hence, the time complexity to find a region globally, an availability zone in the region, a datacenter therein, and a server in the datacenter to place a VM request from a user is $\Theta(ulogu+vlogv+mlogm+nlogn)$.

5. Performance Evaluation

This section illustrates the setup for the experimental work for the evaluation of the proposed algorithm and its counterparts. The results obtained from the experiments are analyzed and presented as graphs to validate the efficiency of the proposed methodology in comparison to some other recent metaheuristic techniques for VMP.

5.1. Experimental Setup

An Infrastructure-as-a-Service (IaaS) cloud environment is required to evaluate the proposed methodology. As it is difficult to evaluate on real cloud infrastructure, the experiments have been conducted using CloudSim [35]. To simulate the work, three AWS cloud regions, i.e. US East (N. Virginia), US West (N. California), and Europe (Ireland) are considered with codes us-east-1, us-west-1, and eu-west-1 respectively [9]. The availability zone is coded as its region code followed by a letter for identification, e.g. us-east-1a [9]. Availability zones are mapped independently to each AWS account, i.e. us-east-1a availability zone for one AWS account may not be the same physical location in a region as us-east-1a for another AWS account. Here, the datacenters are considered of comprising four numbers of servers or the hosts with dissimilar configurations. Each host encompasses some cores of processing, speed of processing, bandwidth, RAM, and storage. 300 VMs with disparate specifications are deliberated for mapping onto the servers. The VM specifications are portrayed in the Table 3. This research considers a heterogeneous cloud milieu, where the count of servers concurrently vary with the count of VMs to endorse the efficacy of the proposed metaheuristic technique.

The QoS parameters such as the energy consumption, carbon emission, response time, and SLA violations are taken into consideration for validating the proposed metaheuristic

Algorithm 1 Salp Swarm based Emperor Penguins Colony (SSEPC) algorithm for the virtual machine placement

Require: The set of cloud regions CR with size u, the list AZ of v number of availability zones in each region, set of datacenters D of size m in each availability zone, list of servers S in each datacenter with size n, set of p number of virtual machines VM, set of q number of requests R, time interval t, response time threshold Th_R.

Ensure: Requests allocated to destination.

for r_l in R **do**
  Call CloudRegion_Search()
  Call AvailabilityZone_Search()
  Call Datacenter_Search()
  if F_{val_1} satisfies according to Eqn. (15) **then**
    Allocate r_l to $cr_x \rightarrow az_y \rightarrow d_i$
    The leader request in the request queue of user_account_id (r_l) = r_l
    user_account_id = user_account_id (r_l)
    Add r_l to user_account_id list
    for r_{l+1} in R **do**
      if user_account_id (r_{l+1}) == user_account_id **then**
        r_{l+1} is the follower request, which will follow the search path of r_l
        Allocate r_{l+1} to $cr_x \rightarrow az_y \rightarrow d_i$
        Add r_{l+1} to user account id list
      end if
      $l = l + 1$
    end for
  end if
  for r_l in user_account_id list **do**
    Sort r_l in descending order of priority
  end for
  for r_l in user_account_id list **do**
    for s_j in d_i **do**
      if $Capi, s_j > Th_{UB}$ **then**
        $j = j + 1$
      else if $Capi, s_j < Th_{UB}$ **then**
        $j = j + 1$
      else
        if F_{val_2} satisfies according to Eqn. (16) **then**
          Allocate r_l to s_j
        else
          $j = j + 1$
        end if
      end if
    end for
  end for
  Update R by removing all r_l allocated to servers
  Call VM_Migration()
end for

Algorithm 2 Algorithm of **CloudRegion_Search( )**

Require: The set CR of cloud regions cr_x with size u, the green energy value of a cloud region $E^G_{cr_x}$, the set R of requests r_l, time interval t, threshold of response time Th_R. If $cr_x$is 100% enabled with green energy, then let us assume $E^G_{cr_x} = 1$. Let the response time for request r_l allocated to cr_x is T_{l,cr_x}.

Ensure: Request r_l is allocated to appropriate cloud region cr_x.

Generate the list of regions CR accessible to r_l
for cr_x in CR **do**
  if $E^G_{cr_x} \geq 0.5$ **then**
    Add cr_x to the sorted list CR^G arranged in descending order based on $E^G_{cr_x}$
  else
    Add cr_x to the sorted list CR^D organized in rising order based on distance between the region and the user
  end if
end for
if $CR^G \neq \phi$ **then**
  Allocate r_l to the cr_x in CR^G such that $T_{l,cr_x} \leq Th_R$
else
  Allocate r_l to the cr_x in CR^D
end if
return cr_x

Algorithm 3 Algorithm of **AvailabilityZone_Search( )**

Require: The set AZ of availability zones az_y with size v, the green energy value of an availability zone $E^G_{az_y}$, the set R of request r_l, time interval t, threshold of response time Th_R. If az_y is 100% enabled with green energy, then let us assume $E^G_{az_y} = 1$. Let the response time for request r_l allocated to cr_x and az_y are T_{l,cr_x} and T_{l,az_y} respectively.

Ensure: Request r_l allocated to appropriate availability zone az_y.

Generate the list of unconstrained availability zones AZ available to r_l
for az_y in cr_x **do**
  if $E^G_{az_y} \geq 0.5$ **then**
    Add az_y to the sorted list AZ^G arranged in descending order based on $E^G_{az_y}$
  else
    Add az_y to the sorted list AZ^D arranged in the ascending order based on the distance of availability zone az_y from the user
  end if
end for
if $AZ^G \neq \phi$ **then**
  Allocate r_l to the az_y in AZ^G such that $(T_{l,cr_x} + T_{l,az_y}) \leq Th_R$
else
  Allocate r_l to the az_y in AZ^D
end if
return az_y

technique SSEPC. The average response time data is taken from [36], which has been collected in two days with interval of 15 minutes through a real cluster. Facebook data [37] has been fetched and analysed to estimate daily requests from the active users, assuming

Algorithm 4 Algorithm of **Datacenter_Search( )**

Require: The set D of datacenters d_i with size m, the green energy value of a datacenter $E^G_{d_i}$, the value of consumption of energy of a datacenter E_{d_i}, the set of request r_l, time interval t, threshold of response time Th_R. If d_i is 100% enabled with green energy, then let us assume $E^G_{d_i} = 1$. Let the response time for request r_l allocated to cr_x, az_y, and d_i are T_{l,cr_x}, T_{l,az_y}, and T_{l,d_i} respectively.

Ensure: Request r_l allocated to appropriate datacenter d_i.

for d_i in az_y **do**
  if $E^G_{d_i} \geq 0.5$ **then**
    Add d_i to the sorted list D^G organized in descending order based on $E^G_{d_i}$
  else if $E_{d_i} < 0.5$ **then**
    Add d_i to the sorted list D^E arranged in the increasing order of E_{d_i}
  else
    Add d_i to the sorted list D^D organized in the ascending order based on the distance between datacenter d_i and the user
  end if
end for
if $D^G \neq \phi$ **then**
  Allocate r_l to the d_i in D^G such that $(T_{l,cr_x} + T_{l,az_y} + T_{l,d_i}) \leq Th_R$
else if $D^E \neq \phi$ **then**
  Allocate r_l to the d_i in D^E such that $(T_{l,cr_x} + T_{l,az_y} + T_{l,d_i}) \leq Th_R$
else
  Allocate r_l to the d_i in D^D
end if
return d_i

Table 3. Virtual Machine Specification

Type of CPU	Number of Cores for Processing	Speed in MIPS	RAM in GB	Storage in GB	Bandwidth in MIPS
Core_i7_Extreme_ Edition	1-4	500	2	20	1024

that one request is submitted per user on daily basis. The count of requests are estimated in millions from various time zones. The carbon intensity data of the considered cloud locations has been collected on hourly basis obtained on 9th January 2019 [11]. In this work, the solar power is considered as the green energy, whose data from different time zones for the considered cloud locations has been collected [38].

For evaluating the efficacy of SSEPC VMP algorithm, we have chosen OEMACS, GATA, and SCA-SSA as three recent hybrid algorithms of VM placement proposed by Liu et al. [12], Zhao et al. [13], and Gharehpasha et al. [15] respectively for comparison with the proposed methodology. As these three algorithms are evolved by the hybridization of two other existing algorithms just like the hybridization applied in this proposed work, we have chosen them as the appropriate contemporaries in the same direction for comparison.

Algorithm 5 Algorithm of **VM_Migration( )**

Require: The set S of servers s_j with size n, the set VM of virtual machines vm_k with size p, time interval t, upper threshold of server capacity Th_{UB}, and lower threshold of server capacity Th_{LB}.

Ensure: Runtime VM migration for VMP.

for s_j in d_i **do**
  if $Cap_{i,s_j} < Th_{LB}$ **then**
    for vm_k in s_j **do**
      Insert vm_k to vm_migration list in descending order of priority
    end for
    s_j goes to sleep() mode
  else if $Cap_{i,s_j} > Th_{UB}$ **then**
    for vm_k in s_j **do**
      if vm_k is requesting for scaling up resources **then**
        Insert vm_k to vm_migration list in descending order of priority
      end if
    end for
  else
    $j = j + 1$
  end if
end for
for vm_k in vm_migration list **do**
  for s_j in d_i **do**
    if $Th_{LB} \leq Cap_{i,s_j} \leq Th_{UB}$ **then**
      if F_{val_2} satisfies according to Eqn. (16) **then**
        Allocate vm_k to s_j
      else
        $j = j + 1$
      end if
    else
      $j = j + 1$
    end if
  end for
end for

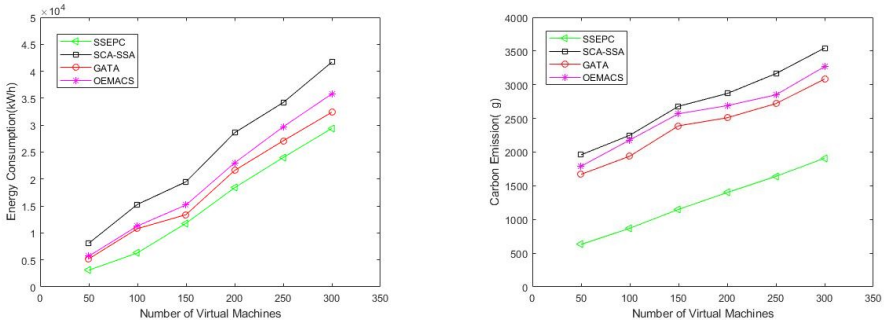
5.2. Obtained Results and Discussion

After simulating the experiments in CloudSim, the results are obtained for validating this work. The results obtained are analysed and are portrayed as graphs in this sub-section.

In SCA-SSA and OEMACS, out of the two participating algorithms of hybridization, one algorithm, i.e. SCA and ACS respectively are conducting the entire searching mechanism, although the search space broadens in this experiment. In GATA, the genetic algorithm takes the charge of searching, where tabu search helps in improving the local search ability of GA as an integrated method. These three algorithms are found to be slowing down in performance with the increase of the search space. In contrast, the proposed methodology SSEPC divides the entire search space into two parts and implements two algorithms in multi-stage collaborative approach for them separately by excelling in performance.

Figure 2(a) shows, SSEPC consumes less energy as compared to SCA-SSA, GATA, and OEMACS. The values of average energy consumption while implementing SSEPC, GATA, OEMACS, and SCA-SSA are found to be $1.48 \times (10)^4$ KWh, $1.7 \times (10)^4$ KWh, $1.89 \times (10)^4$ KWh, and $2.31 \times (10)^4$ KWh respectively. It is already stated above that SSEPC is more efficient in searching than others, for which it avails most suitable resources with better utilization and implements the runtime VMP too. So, it produces optimum result in energy consumption as compared to other algorithms.

Figure 2(b) portrays, SSEPC emits less carbon as compared to its counterparts, i.e. 1230g, 2382g, 2543g, and 2728g of carbon footprints are emitted on an average in case of SSEPC, GATA, OEMACS, and SCA-SSA respectively. SSEPC gives priority to the green energy availability while mapping VMs to the appropriate resources. Each datacenter runs on green energy till its availability and then switches to brown energy. As SCA-SSA, GATA, and OEMACS have not considered the renewable energy sources to operate the datacenters, they are emitting more carbon.



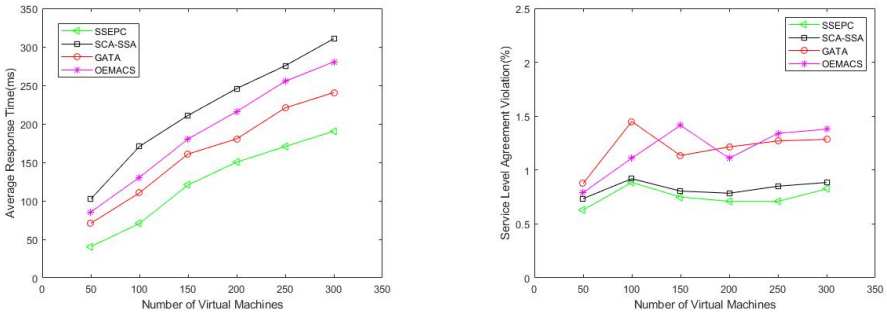
(a) Comparison of Energy Consumption of SSEPC with other VMP Algorithms

(b) Comparison of Carbon Emission of SSEPC with other VMP Algorithms

Fig. 2. Comparison of Energy Consumption and Carbon Emission of SSEPC with others

Due to prioritization of user requests during runtime VMP in SSEPC, along with the implementation of leader-follower mechanism, the proposed algorithm gives optimum average response time with a better QoE for the users than the other metaheuristic algorithms considered for comparison. This can be observed in Figure 3(a). The average response time during the implementation of SSEPC, GATA, OEMACS, and SCA-SSA are obtained as 117.32ms, 157.48ms, 182.15ms, and 208ms respectively.

GATA and OEMACS have not taken care of SLA violations as a QoS parameter. Although SCA-SSA has considered the minimization of SLA violations during VMP, it slows down in a larger search space resulting a bit higher SLA violations as compared to SSEPC. Figure 3(b) depicts the comparison of SLA violations of SSEPC and others. From Figure 3(b), it can be viewed that SSEPC minimizes the SLA violations more in comparison to others. The average SLA violations in percentage for SSEPC, SCA-SSA, GATA, and OEMACS are found to be 0.8%, 0.81%, 0.83%, and 0.87% respectively.



(a) Comparing Average Response Time of SSEPC with other VMP Algorithms

(b) Comparison of SLA Violation of SSEPC with other VMP Algorithms

Fig. 3. Comparison of Average Response Time and SLA Violation of SSEPC with others

The results obtained clearly indicate that the proposed SSEPC outperforms its counterparts for the QoS parameters like energy consumption, carbon emission, average response time, and SLA violations. To keep the brevity of this paper, we have reduced the detailed description in this section.

6. Conclusion and Future Works

This work proposes a multi-stage collaborative hybrid algorithm for the solution of VMP problem in the cloud milieu. In this work, the search space is broaden by the incorporation of cloud regions and the availability zones along with the datacenters and the servers to make it more realistic. The proposed SSEPC methodology explores in the first phase to search the appropriate region as well as the availability zone and then exploits to search the suitable datacenter therein for the placement of the VMs by utilizing salp swarm optimization. In the second phase, it implements emperor penguins colony algorithm to search locally in a datacenter to find the most suitable server to place the VMs. Along with initial VMP, runtime VMP is also applied in this work. Some vital QoS metrics like energy consumption, carbon emission, average response time, and SLA violations are considered to appraise the efficacy of the recommended hybrid metaheuristic technique in comparison with some recent VMP algorithms, where it outperforms others. Utilization of green energy is prioritized during the process of VMP for the environmental sustainability. User requests are processed on the basis of priority for better QoE of the users.

As part of our future work, we plan to extend our approach to include the cost of energy consumption and carbon footprint cost along with calculating their amounts to analyze the evaluation from both qualitative and quantitative perspective. Hybridization of some other techniques in this direction can be explored to optimize the QoS metrics better. The exploration of adaptive algorithms can be considered to learn and predict the optimal VMP over time, considering energy consumption, carbon footprint, and other cloud service demands.

References

1. Parida, S., Pati, B., Nayak, S. C., Panigrahi, C. R., Weng, T. H.: PE-DCA: Penalty elimination based data center allocation technique using guided local search for IaaS cloud. Computer Science and Information Systems, 19(2), 679-707. (2022)
2. Feng, H., Deng, Y., Li, J.: A global-energy-aware virtual machine placement strategy for cloud data centers. Journal of Systems Architecture, 116, 102048. (2021)
3. Wikipedia, Data center (2023).[Online]. Available: https://en.wikipedia.org/wiki/Data-center (current June 2023)
4. Koot, M., Wijnhoven, F.: Usage impact on data center electricity needs: A system dynamic forecasting model. Applied Energy, 291, 116798, 12-27. (2021)
5. Sarpong, K. A., Xu, W., Gyamfi, B. A., Ofori, E. K.: A step towards carbon neutrality in E7: The role of environmental taxes, structural change, and green energy. Journal of Environmental Management, 337, 117556. (2023)
6. Xu, M., Buyya, R.: Managing renewable energy and carbon footprint in multi-cloud computing environments. Journal of Parallel and Distributed Computing, 135, 191-202. (2020)
7. Abbasi-khazaei, T., Rezvani, M. H.: Energy-aware and carbon-efficient VM placement optimization in cloud datacenters using evolutionary computing methods. Soft Computing, 26(18), 9287-9322. (2022)
8. AWS, Regions and Availability Zones. [Online]. Available: https://aws.amazon.com/about-aws/global-infrastructure/regions_az/ (current June 2023)
9. AWS, Amazon Elastic Compute Cloud, Regions and Zones (2023). [Online]. Available: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html (current June 2023)
10. Mirjalili, S., Gandomi, A. H., Mirjalili, S. Z., Saremi, S., Faris, H., Mirjalili, S. M.: Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. Advances in engineering software, 114, 163-191. (2017)
11. Harifi, S., Khalilian, M., Mohammadzadeh, J., Ebrahimnejad, S.: Emperor Penguins Colony: a new metaheuristic algorithm for optimization. Evolutionary Intelligence, 12(2), 211-226. (2019)
12. Liu, X. F., Zhan, Z. H., Deng, J. D., Li, Y., Gu, T., Zhang, J.: An energy efficient ant colony system for virtual machine placement in cloud computing. IEEE transactions on evolutionary computation, 22(1), 113-128. (2016)
13. Zhao, D. M., Zhou, J. T., Li, K.: An energy-aware algorithm for virtual machine placement in cloud computing. IEEE Access, 7, 55659-55668. (2019)
14. Samriya, J. K., Chandra Patel, S., Khurana, M., Tiwari, P. K., Cheikhrouhou, O.: Intelligent SLA-aware VM allocation and energy minimization approach with EPO algorithm for cloud computing environment. Mathematical Problems in Engineering. (2021)
15. Gharehpasha, S., Masdari, M., Jafarian, A.: Power efficient virtual machine placement in cloud data centers with a discrete and chaotic hybrid optimization algorithm. Cluster Computing, 24(2), 1293-1315. (2021)
16. Zhao, D., Zhou, J.: An energy and carbon-aware algorithm for renewable energy usage maximization in distributed cloud data centers. Journal of Parallel and Distributed Computing, 165, 156-166. (2022)
17. Xu, M., Toosi, A. N., Buyya, R.: Ibrownout: an integrated approach for managing energy and brownout in container-based clouds. IEEE Transactions on Sustainable Computing, 4(1), 53-66. (2018)
18. Tang, M., Pan, S.: A hybrid genetic algorithm for the energy-efficient virtual machine placement problem in data centers. Neural processing letters, 41(2), 211-221. (2015)
19. Yadav, R., Zhang, W., Kaiwartya, O., Singh, P. R., Elgendy, I. A., Tian, Y. C.: Adaptive energy-aware algorithms for minimizing energy consumption and SLA violation in cloud computing. IEEE Access, 6, 55923-55936. (2018)

20. Khodayarseresht, E., Shameli-Sendi, A., Fournier, Q., Dagenais, M.: Energy and carbon-aware initial VM placement in geographically distributed cloud data centers. Sustainable Computing: Informatics and Systems, 100888. (2023)
21. Belabed, D., Secci, S., Pujolle, G., Medhi, D.: Striking a balance between traffic engineering and energy efficiency in virtual machine placement. IEEE Transactions on Network and Service Management, 12(2), 202-216. (2015)
22. Justafort, V. D., Beaubrun, R., Pierre, S.: A hybrid approach for optimizing carbon footprint in intercloud environment. IEEE Transactions on Services Computing, 12(2), 186-198. (2016)
23. Anusooya, G., Vijayakumar, V.: Reduced carbon emission and optimized power consumption technique using container over virtual machine. Wireless Networks, 27, 5533-5551. (2021)
24. Ting, T. O., Yang, X. S., Cheng, S., Huang, K.: Hybrid metaheuristic algorithms: past, present, and future. Recent advances in swarm intelligence and evolutionary computation, 71-83. (2015)
25. Microsoft, Azure, Regions and availability Zones (2023). [Online]. Available: https://docs.microsoft.com/en-us/azure/availability-zones/az-overview (current June 2023)
26. Forbes, With The Public Clouds Of Amazon, Microsoft And Google, Big Data Is The Proverbial Big Deal (2017). [Online]. Available: https://www.forbes.com/sites/johnsonpierr/2017/06/15/with-the-public-clouds-of-amazon-microsoft-and-google-big-data-is-the-proverbial-big-deal/?sh=4ba4ae7c2ac3 (current June 2023)
27. Arcaini, P., Riccobene, E., Scandurra, P.: Modeling and analyzing MAPE-K feedback loops for self-adaptation. In 2015 IEEE/ACM 10th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (pp. 13-23). IEEE. (2015)
28. Basmadjian, R., Ali, N., Niedermeier, F., De Meer, H., Giuliani, G.: A methodology to predict the power consumption of servers in data centres. In Proceedings of the 2nd international conference on energy-efficient computing and networking (pp. 1-10). (2011)
29. Wang, N., Zhang, J., Xia, X.: Energy consumption of air conditioners at different temperature set points. Energy and Buildings, 65, 412-418. (2013)
30. Tang, Q., Gupta, S. K. S., Varsamopoulos, G.: Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach. IEEE Transactions on Parallel and Distributed Systems, 19(11), 1458-1472. (2008)
31. National grid, What is carbon intensity? (2023). [Online]. Available: https://www.nationalgrid.com/stories/energy-explained/what-is-carbon-intensity (current June 2023)
32. Babu, D. LD., Venkata Krishna, P.: Honey bee behavior inspired load balancing of tasks in cloud computing environments. Appl. Soft Computer Journal, 13(5), 2292-2303. (2013)
33. Rizk-Allah, R. M., Hassanien, A. E., Elhoseny, M., Gunasekaran, M.: A new binary salp swarm algorithm: development and application for optimization tasks. Neural Computing and Applications, 31(5), 1641-1663. (2019)
34. Son, J., Buyya, R.: Priority-aware VM allocation and network bandwidth provisioning in software-defined networking (SDN)-enabled clouds. IEEE Transactions on Sustainable Computing, 4(1), 17-28. (2018)
35. Calheiros, R. N., Ranjan, R., Beloglazov, A., De Rose, C. A., Buyya, R.: CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Software: Practice and Experience, 41(1), 23-50. (2011)
36. Doyle, J., Shorten, R., O'Mahony, D.: Stratus: Load balancing the cloud for carbon emissions control. IEEE Transactions on Cloud Computing, 1(1), 1-1. (2013)
37. Atikoglu, B., Xu, Y., Frachtenberg, E., Jiang, S., Paleczny, M.: Workload analysis of a large-scale key-value store. In Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems (pp. 53-64). (2012)
38. E. Commission, Photovoltaic Geographical Information System (2022). [Online]. Available: https://re.jrc.ec.europa.eu/pvg_tools/en/tools.html (current June 2023)

Bivasa Ranjan Parida received the B.E. degree in CSE in 2004, and the M.Tech. degree in CSE in 2010 from Biju Patnaik University of Technology, Odisha, India. Currently, he is pursuing his Ph.D. in CSE as a scholar in Veer Surendra Sai University of Technology, Burla, India under the guidance of Prof. Amiya Kumar Rath. His research area includes cloud computing, grid computing, and wireless body area network. He has 18 years of teaching and research experience.

Amiya Kumar Rath received the B.E. degree in CSE from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad in 1990, the M.B.A. degree in Systems Management from Shivaji University in 1993, the M.Tech. degree in Computer Science from Utkal University in 2001, and the Ph.D. degree in Computer Science from Utkal University in 2005, with a focus on embedded systems. He is currently a Professor with the Department of Computer Science and Engineering, Veer Surendra Sai University of Technology, Burla, India. He has contributed over 80 research level papers to many national and international journals and conferences. He has published seven books by reputed publishers. His research interests include embedded systems, ad hoc networks, sensor network, power minimization, evolutionary computation, and data mining. Currently, he has been deputed as Vice-Chancellor of Biju Patnaik University of Technology, Odisha, India.

Bibudhendu Pati is an Associate Professor in the Department of Computer Science at Rama Devi Women's University, Bhubaneswar, India. He received his BE in CSE degree with Honours, ME in CSE from NITTTR, Chandigarh, PhD degree from IIT Kharagpur, India. He has around 25 years of experience in teaching and research. His current research interests include Wireless Sensor Networks, Mobile Cloud Computing, Internet of Things, and Advanced Network Technologies. He has been involved in many professional and editorial activities. He has got several papers published in reputed journals, conference proceedings, and books of International repute. He is the Life Member of ISTE, Life Member of CSI, and Senior Member of IEEE.

Chhabi Rani Panigrahi is an Assistant Professor in the Department of Computer Science at Rama Devi Women's University, Bhubaneswar, India. Prior to this she was associated with Central University of Rajasthan. She received her PhD degree in Computer Science & Engineering from Indian Institute of Technology Kharagpur, India. Her current interests include software testing, mobile cloud computing, Internet of Things, and machine learning. She has more than 23 years of teaching and research experience. She has published several scholarly articles in journals and conferences of international repute. She was invited to organize and chair many international conferences. She is the author of a text book and a few edited books. She is a life member of the Indian Society of Technical Education (ISTE) and a member of IEEE and the Computer Society of India (CSI).

Hitesh Mohapatra received the B.E. degree in Information Technology in 2006, and the M.Tech. degree in CSE in 2009 from Biju Patnaik University of Technology, Odisha, India. He received his Ph.D. in Computer Science & Engineering in 2021 from Veer Surendra Sai University of Technology, Burla, India. He has contributed 20 SCI and Scopus indexed research papers, 15 international/ national conferences and 2 books on Software Engineering and C Programming respectively. He has 12 years of teaching experience

both in industry and academia. His research interests include wireless sensor network, smart city, smart grid and smart water.

Tien-Hsiung Weng is currently working as a Professor in the Department of Computer Science and Information Engineering, Providence University, Taichung City, Taiwan. He received his Ph.D. in Computer Science from the University of Houston, Texas. His research interests include parallel computing, high performance computing, scientific computing, and machine learning.

Rajkumar Buyya is a Fellow of IEEE, Professor of Computer Science and Software Engineering and Director of the Cloud Computing and Distributed Systems (CLOUDS) Laboratory at the University of Melbourne, Australia. He is also serving as the founding CEO of Manjrasoft, a spin-off company of the University. He is one of the highly cited authors in computer science and software engineering worldwide (h-index=109, g-index=230, 58,300+ citations). Microsoft Academic Search Index ranked him as #1 author in the world (2005-2016) for both field rating and citations evaluations in the area of Distributed and Parallel Computing." A Scientometric Analysis of Cloud Computing Literature" by German scientists ranked him as the World's Top-Cited (#1) Author and the World's Most-Productive (#1) Author in Cloud Computing. Software technologies for Grid and Cloud computing developed under his leadership have gained rapid acceptance and are in use at several academic institutions and commercial enterprises in 40 countries around the world. Manjrasoft's Aneka Cloud technology developed under his leadership has received "2010 Frost & Sullivan New Product Innovation Award". Recently, he received "Mahatma Gandhi Award" along with Gold Medal for his outstanding and extraordinary achievements in Information Technology field and services rendered to promote greater friendship and India-International cooperation. He served as the founding Editor-in-Chief of the IEEE Transactions on Cloud Computing. He is currently serving as Co-Editor-in-Chief of Journal of Software: Practice and Experience, which was established over 45 years ago.

Received: September 23, 2023; Accepted: January 3, 2024.

Computer Science and Information Systems 21(3):781–806 https://doi.org/10.2298/CSIS230730015P

A Novel Approach to Source Code Assembling in the Field of Algorithmic Complexity

Đorđe Pešić[1], Milena Vujošević Janičić[2], Marko Mišić[1], and Jelica Protić[1]

[1] University of Belgrade - School of Electrical Engineering
Bulevar kralja Aleksandra 73, RS-11120 Belgrade, Serbia
pd235049p@student.etf.bg.ac.rs
{marko.misic, jelica.protic}@etf.bg.ac.rs
[2] University of Belgrade - Faculty of Mathematics
Studentski trg 16, RS-11000 Belgrade, Serbia
milena.vujosevic.janicic@matf.bg.ac.rs

Abstract. Computational complexity analysis plays an essential part in the education of computer and software engineers. For that reason, it is carefully studied in programming courses, as well as in the algorithms and data structures courses. The number of students who learn programming is rapidly growing, but the number of teachers cannot keep up with that trend. Therefore, it is necessary to develop tools that can ease and accelerate the daily tasks of teachers, especially for learning purposes and in the context of automating the processes of exam preparation. We propose a novel template- and rule-based approach and a corresponding software system for assembling synthetic source code segments of defined time complexity. Based on the developed grammar, the system can produce source code segments with a broad scope of different time complexities while guaranteeing the complexity of the generated segment. The system can be used for generating questions for exams as it can assemble a large number of different code segments that can be given as questions that have similar difficulty levels. The system was evaluated both by human experts and ChatGPT tool.

Keywords: automated source code assembling, computational complexity, time complexity, rule-based assembling.

1. Introduction

Software is becoming an essential part of everyday life. The software industry is growing constantly, with a high and increasing demand for well-trained programmers. Many countries invest a lot in the development of programming education, leveraging different innovative approaches from primary education [13] to higher education [60] levels. Computer science and engineering schools have an essential role in such education. For example, at the Department of Electrical Engineering and Department of Software Engineering of the School of Electrical Engineering at the University of Belgrade, introductory programming courses are taught in the first year of undergraduate studies. These courses attend more than 1000 students [33]. The number of students is expected to grow even more in the following years amid the increasing interest in those topics [48].

The number of teaching personnel does not follow this increased number of students. In addition, the number of exam periods is six times per course per year. This puts pressure on the teaching staff and makes automated test assembling and evaluation crucial in

maintaining the quality of the teaching process and helping the professors and teaching assistants in their everyday responsibilities.

Computational complexity and analysis of algorithms are fundamental disciplines in computer science. Computational complexity classifies computational problems according to their difficulty and assigns complexity classes to them [26], [14], [12]. The analysis of an algorithm describes how good it is from the quantitative point of view [26] and includes time and space complexity analysis. Both disciplines help programmers understand and estimate the resources necessary to execute programs. Time complexity approximates the time needed for the program execution depending on the problem size, while space complexity models the computer memory usage [50]. The problem size is usually input data size, noted as a dimension of the problem n.

Many researchers have studied these areas from the early days of computer science. Donald Knuth summarized concepts of computational complexity theory and the big O notation [25], first introduced a notation for lower-bounded functions [27], and created the term *analysis of algorithms* [26]. Papadimitriou represented complexity as a number-theoretic concept and examined P and NP problems [39]. Arora and Barak collected a substantial quantity of complexity classes [5]. They also analyzed space complexity and represented the theoretical concept of randomized and quantum computation [5].

Understanding the basic concepts of the computational complexity of algorithms, especially time complexity, is a substantial part of introductory courses on programming, algorithms, and data structures. There are well-known classes of computer programs with certain time complexities used in real-life applications, such as binary search, different sorting methods, tree and graph traversals, finding shortest paths in graphs, and similar. Most universities recognize the necessity of teaching these concepts. For example, Table 1 shows the first five universities on the ARWU (Shanghai) list [51] for the year 2022 and the courses that cover the topics on computational complexity as a part of their curriculum. The selected list sorts these universities by Computer Science and Engineering subject.

However, besides real-life examples, there is frequently a need for synthetic source code examples with defined time complexity. Based on our previous work and efforts [40,44], we propose a novel, template- and rule-based approach to source code assembling of predefined time complexity. The approach is implemented through a software system that generates code segments based on the defined code templates and interaction rules. One of the most important goals of the system is to improve both the teaching process and the quality of student examination. Assembled segments can be used in classes for teaching or given to students at exams. Teachers can choose one of the segments in manual mode or let the software system choose it in automated mode. Usage of the proposed software significantly decreases exam assembly time, as instructors configure it once and then use the assembled segments.

The software can help to prevent cheating with the possibility of generating many segments according to the given criteria. Therefore, each student can get a different question. In addition, the system does not contain a database of questions in its final form, which prevents possible database theft. To secure result correctness, the software calculates and verifies the time complexity of the generated segments.

Table 1. World Universities and their courses which study algorithm time complexity

University	Courses
Massachusetts Institute of Technology (MIT)	Introduction to Computer Science Programming in Python (6.001), Introduction to Algorithms (6.006), Automata, Computability and Complexity (6.045), Design and Analysis of the Algorithms (6.046)
Stanford University	Mathematical Foundations of Computing (CS103), Programming abstractions (CS106B), Introduction to automata and complexity theory (CS154), Design and analysis of algorithms (CS161), Computational complexity (CS254)
University of California, Berkeley	The Structure and Interpretation of Computer Programs (CS61A), Efficient Algorithms and Intractable Problems (CS170), Computability and Complexity (CS172)
Carnegie Mellon University	Great Ideas in Theoretical Computer Science (15251), Algorithms & Advanced Data Structures (15351), Algorithms (15451), Undergraduate complexity theory (15455)
Tsinghua University	Logic and Computation I, Combinatorics and Algorithms Design, Advanced Theoretical Computer Science

Keeping in mind the educational perspective of a large number of students, the necessity of automatizing the exam assembling process, and cheating prevention, the main contributions of the paper are:

- We developed a novel, formal approach for assembling code segments with verified target time complexity.
- We implemented a software system based on our formal approach and made it publicly available and open source [42].
- We evaluated the implemented system and showed that assembled segments resemble hand-crafted and that the tool can significantly decrease exam preparation time.

The paper is organized as follows. We introduce computational complexity analysis emphasizing time complexity analysis in Section 2. We define template- and rule-based proposed system principles, the assembling approach, and the time complexity calculation in Section 3. We present the system architecture in Section 4. We evaluate the system in Section 5. We discuss related work in Section 6. We briefly conclude and summarise directions for future work in Section 7.

2. Theoretical background

We introduce all the necessary definitions for understanding the proposed system's design principles. We also note the undecidability of the symbolic complexity calculation.

2.1. About time complexity calculation

We assume that a problem size depends on only one input parameter n. This assumption is not a limiting factor because the time complexity calculations are usually done in one

dimension [14], [12]. The function $T(n)$, which approximates the program execution time as a function of the problem size n, is defined with several elementary steps. Each step has a constant execution time [14]. Unfortunately, even for small programs, this function becomes very complex. Therefore, it is usually reduced to the function with the same asymptotic behavior for sufficiently large n. For example, the function $T(n) = n^2 + log(n) + n + 2^n$, is reduced to the function 2^n. This is formalized by big O notation, where $O(2^n)$ denotes a class of functions that have 2^n as their upper limit. The big O notation is defined by Definition 1 [12].

Definition 1. *For a given function* $T(n)$*,*

$$T(n) = O(f(n))$$

if there exist $c \in \mathcal{R}$ *and* $n_0 \in \mathcal{N}$ *such that* $c > 0$ *and* $n_0 > 0$ *and for all* $n \in \mathcal{N}$ *such that* $n > n_0$ *it holds*

$$0 \leq T(n) \leq c \cdot f(n)$$

.

Expression $T(n) = O(f(n))$ denotes that function $T(n)$ belongs to the class of functions $O(f(n))$. When this notation is used, the analysis of the complexity of an algorithm is reduced to the analysis of the complexity of the general structure of the algorithm.

Different characteristics can be proved directly from the definition [12]. Two of them, used in our approach for complexity calculation, are based on calculating complexity when two independent code segments are put into the sequence and when one code segment is nested into a loop. In the case of sequence, the resulting complexity is a sum of the two complexity classes. In the case of nesting, there are some cases where the resulting complexity is a product of two complexity classes.

It can be proved that the sum of the two complexity classes is equal to the larger complexity class [12], as shown in Equation 1

$$O(f(n)) + O(g(n)) = O(max(f(n), g(n))) \,. \tag{1}$$

The product of two complexity classes is the complexity class of the product function [12], as shown in Equation 2.

$$O(f(n)) \cdot O(g(n)) = O(f(n) \cdot g(n)) \tag{2}$$

Although time complexity function can have arbitrary form, there is a subset of function classes that are commonly used in real-world programming and therefore have exceptional importance in education: constant $O(1)$, logarithmic $O(log(n)$, linear $O(n)$, polynomial $O(n^k), k > 1$, exponential $O(k^n), k > 1$ and factorial $O(n!)$. Equation 3 compares these function classes.

$$O(1) < O(log(n)) < O(n) < O(n^k) < O(k^n) < O(n!) \tag{3}$$

2.2. Halting problem and the complexity calculation

In the general case, the time complexity calculation of an arbitrary source code is an undecidable problem. This is a direct consequence of the undecidability of the halting problem (proved by Alan. M. Turing [56]). The halting problem is the problem of determining for an arbitrary program and an arbitrary input whether the program will finish running or continue to run forever. Since it is not possible to construct an algorithm that can determine if each program stops its execution, then it is also not possible to make an algorithm that would count the number of executed statements, i.e., which would calculate the time complexity of an arbitrary program. The Listing 2.1 shows the example of the Collatz conjecture problem [11]: a program that looks very simple, but its time complexity cannot be calculated.

The Collatz conjecture program generates a sequence of numbers. For the arbitrary number n, there is no mathematical proof that the sequence will eventually reach number 1 and that the loop will terminate. Conway proved that Collatz-type problems are undecidable [11].

```
while(n != 1){
  if(n % 2 == 0){
    n = n / 2;
  } else {
    n = 3 * n + 1;
  }
  print(n);
}
```

Listing 2.1. Pseudo code of the Collatz conjecture problem

2.3. Time complexity calculation approaches in software

A time complexity calculation of a source code is a complex task, undecidable for an arbitrary source code. However, extensive research is done to estimate and predict time complexity or generate a code segment of the given complexity.

Early work in this context was done by McCabe, who proposed the concept of cyclomatic complexity [30]. Cyclomatic complexity is a software metric that measures the number of linearly independent paths in the control flow graph of the given source code. It is then used to estimate the complexity of the program. However, it does not quantify the number of passes through the paths in the execution context of the program. Hence, it cannot be used directly to estimate time complexity.

The Master theorem for divide-and-conquer problems was given in [7]. It provides an asymptotic analysis with big O notation for recurrence relations presented in typical divide-and-conquer algorithms. In general, the time complexity of the algorithm that takes the problem of the size n and partitions it in partitions each sized n/b can be calculated with Equation 4:

$$T(n) = a \cdot T(n/b) + f(n)\,. \tag{4}$$

where a is the number of partitions and $f(n)$ is the time to create partitions and combine their results. The time complexity of well-known algorithms such as binary search of a sorted array, binary tree traversal, and merge sort can be calculated with this approach. However, the approach is limited only to divide-and-conquer problems.

In the recent period, there were attempts to use machine learning methods to estimate the time complexity of the code. Those methods operate on hand-crafted features or appropriate code embeddings, which are used to learn the internal structure of the code. Code representations such as *code2vec* [4] are used in this context in [53] to predict the time complexity class of a given program written in Java. For training purposes, they collected annotated dataset CoRCoD: Code Runtime Complexity Dataset [32][53], extracted from online judges, with 932 source code examples with five different classes of complexities: $O(1)$, $O(log(n)$, $O(n)$, $O(n \cdot logn)$, $O(n^2)$. The results showed that code embeddings have comparable performances to hand-crafted features for classification using Support Vector Machines, which showed the best accuracy, precision, and recall on the test set. However, the accuracy was at most around 70% and varied across classes significantly.

New research on time and space complexity prediction can be found in [52],[34]. Siddiq et al. used a GPT3-based code generation tool *GitHub Copilot* to predict the runtime complexity of a given source code using zero-shot prompting [52]. The obtained accuracy ranges from 45.44% to 56.38%, depending on the suggestions given to the system. A transformer-based approach to space and time complexity is considered in [34]. The authors state that existing time complexity calculation libraries and tools only apply to limited use cases because of the lack of a well-defined rule-based system. They proposed using code-based language models, such as BERT, CodeBERT, GraphCodeBERT, CodeT5, and Longformer. They achieved prediction accuracy ranging from 72.55% for C++ codes to 92.08% for Java codes, depending on the model used.

Another machine learning approach was conducted in [36] using a dataset based on exam questions from the Programming 1 course at the University of Belgrade, School of Electrical Engineering. The dataset contained 61 exam questions and solutions, each representing a code segment written in Pascal and its time complexity. Features were crafted from the source code using three models: count vectorizer, term frequency-inverse document frequency, and embedding layers in neural networks. The results showed that classical machine learning models, such as random forests, perform better than more advanced models based on neural networks. However, the model's accuracy was not high due to the relatively small annotated data set used in training.

The authors of this paper adopted another approach to time complexity calculation. Through several papers [44], [45], [40], [41], we developed a set of methods for source code assembly of a given time complexity. In the first approach, a set of basic, handcrafted, and verified complexity segments was used to produce new segments with code parametrization and obfuscation based on simple rules given in Equation 1 and Equation 2 [44]. However, only basic complexity classes and combining techniques were supported. Further approach [45] introduced a more abstract assembly strategy modeled with XML and abstract syntax trees. The paper [40] was oriented toward educational and technical aspects of the system. These aspects are further elaborated in the extended version of the conference paper [41], where we introduce a formal approach to the segment assembling process with a prototype implementation described with the subset of C grammar. The system is based on templates and rules together with the time complexity calculation approach done by symbolic calculation using the Yacas tool. This significantly broadens the scope of assembled code segments and their complexity. Although promising, symbolic calculation has limitations due to its high time complexity and non-applicability for some

complexity classes. Preliminary analysis and assessment of the system for teaching and evaluation were given in [43]. The results suggest that the system produces code segments with similar semantics but different textual representations.

3. Segment assembling process and time complexity calculation

This section presents an algorithm for source code assembling and all the necessary concepts. The algorithm merges two code segments (templates) with known complexities using a chosen rule, which lets us calculate and guarantee the exact complexity of the resulting segment. The resulting segment can be used in further segment assembling. This enables rule cascading and complex segment assembling.

3.1. Basic building blocks, templates, and rules

A template is an arbitrary source code segment, assembled of building blocks, recognized from imperative programming. Building blocks are assignment statements and simple *while* loops. *If* statements are generated in the assembling process, as will be described. *For* and *do-while* loops are generated by transforming *while* loops at the end of the assembly process (not further discussed).

Figure 1 shows different template examples, denoted with T1 to T10. Templates T1, T3, and T5 are examples of assignment statements, and they are predefined. Templates T2 and T9 are examples of simple loops. Templates T2, T4, and T6 to T10 are products of the assembling process.

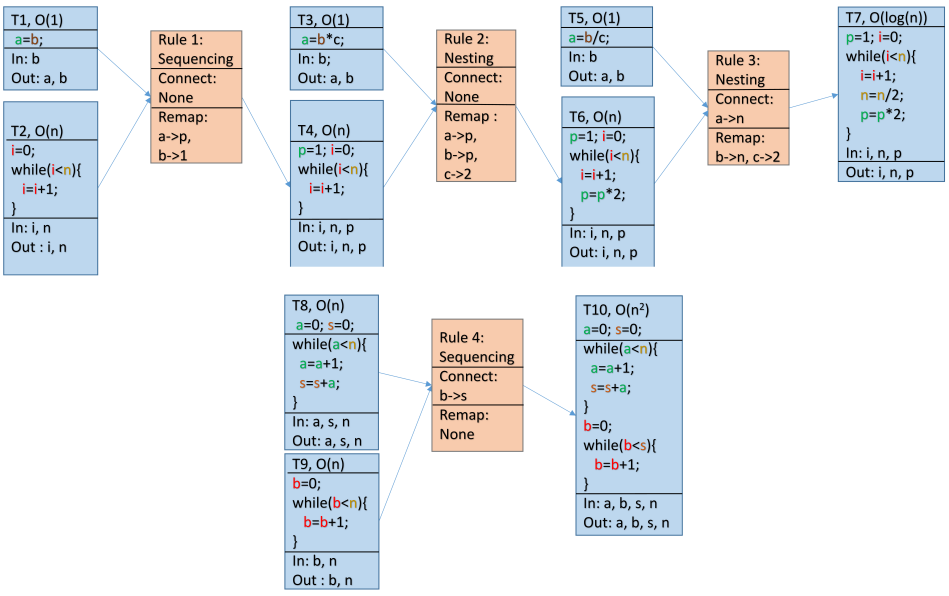


Fig. 1. Assembling process with different templates and rules

We support assignment statements that include expressions with addition, subtraction, multiplication, and division. We also support different kinds of simple loops. The structure of a simple loop is described with grammar. In the grammar, we use the symbols presented in Table 2 and abbreviations given in Listing 3.1. We also introduce the operator $*$ (star) to shorten the number of grammar rules (without information loss). The operator $*$ takes one permutation from the set of all permutations of the sequence of the terminal or non-terminal grammar symbols. Symbols presented in Table 2 are non-terminal symbols. They are mapped to variables and constants used in particular source code segments. For example, in Figure 1 variable i in template T1 is a specialization of symbol a_0, and constant 1 is a specialization of symbol z_1.

Table 2. Symbols used in the grammar

Symbols	Description
a_0, a_1	Variables holding lower loop bounds
b_0, b_1	Variables holding upper loop bounds
s_0	Variable holding a sum
v_0, v_1, ..., v_6	Auxiliary variables
c_0, c_1, c_2	Numerical constants greater than zero
z_0, z_1	Numerical constants greater than zero ordered such that $z_0 < z_1$
o_0, o_1	Numerical constants greater than one ordered such that $o_0 < o_1$

```
init(lower,upper) := (lower=z0; upper=n;)*
init(lower,upper,aux0) := (lower=z0; upper=o0; aux0=n;)*
body(lower,upper,aux0,aux1,aux2) := body1(lower,upper) |
                                    body2(lower,upper,aux0,aux1) |
                                    body3(lower,upper) |
                                    body4(lower,aux0,aux2)
body1(lower,upper) := (lower=lower+z1; [upper=upper+z0;])* |
                      (upper=upper-z1; [lower=lower-z0;])* |
                      (lower=lower+c0 | upper=upper-c1)
body2(lower,upper,aux0,aux1) := (lower=aux0*aux0;
                                 upper=aux1*aux1;
                                 aux0=aux0+z0;
                                 aux1=aux1-z0;)*
body3(lower,upper) := (lower=lower*o1;
                [upper=upper-z0; | upper=lower+z0; | upper=upper*o0;])*
                | (upper=upper/o1;
                [lower=lower-z0; | lower=lower+z0; | lower=lower/o0])*
body4(lower,aux0,aux1)  :=  (lower=aux0*aux0; aux0=aux0+z0;)* |
                            (lower=lower+aux1; aux1=aux1+z0;)*
```

Listing 3.1. The abbreviations for initializations of variables and definitions of bodies that are used in the grammar

A description of a supported structure of a simple loop segment is given in Listing 3.2. The simple loop segment contains initialization statements, one *while* loop, and a loop body. We initialize only those variables that are used in the loop body. The body is a sequence of statements. We define four different body types, denoted with $body_1$, $body_2$, $body_3$, and $body_4$ (Listing 3.1). Each type contains its corresponding statements listed in

the table. Simple loops with $body_1$ and $body_2$ have complexity $O(n)$, loops with $body_3$ have complexity $O(log(n))$ and loops with $body_4$ have complexity $O(sqrt(n))$.

```
simple_loop :=  init(a0,b0)
                init(v0,v1,v2)
                while(a0<b0){
                    body(a0,b0,v0,v1,v2)
                }
```

Listing 3.2. Grammar rule that describes a simple loop segment

Two source code segments are data dependent if the first segment produces a result used in the second segment and if that result impacts the time complexity of the combination of these two segments. Otherwise, these two source code segments are mutually independent. Figure 1 also shows an example of the data dependency where two segments are put into a sequence. Template T8, with $O(n)$ complexity, produces a value for variable s. After the loop, the value stored in s is a sum of the first n numbers. The variable s is forwarded as an upper loop bound of the T9. Therefore, the time complexity of the sequence is $O(n^2)$.

Each template has ports defined by the system. Ports describe data entry points (input ports) and exit points (output ports). A port is a variable that satisfies one of the following conditions: the variable value is gathered from the other template (input port), or the variable value is forwarded to the other template (output port). Template T1 from Figure 1 has $O(1)$ complexity, input port b, and output ports a and b. The output port of the template can be connected to the input port on another template. Port connections represent data flow between templates and exist only if that data flow impacts the resulting complexity calculation. This happens when templates have data dependencies. Port connections are used to mark particular data dependencies.

During the assembling, the variables inside one template can be renamed or specialized to constants. In this case, renaming increases the possibility of template reuse inside different rules. Renaming is also used in the creation of port connections between two templates. In this case, all appearances of the output port names are replaced with the input port names.

An operation is a template merging method. It can be *sequencing*, *nesting*, and *selecting*. Sequencing operation puts one template after another. The nesting operation puts one template inside another template. Selecting operation puts one template into the *then* branch of an *if* statement or two templates into the *then* and *else* branches of the *if* statement. *If* statement additionally has a selection condition. We use conditions that do not trivially evaluate to *true* or *false*. Therefore, the code in existing branches must be considered when calculating time complexity.

The rule collects all the information needed to assemble a new segment. This information contains used templates, the data dependencies between them (denoted as port connections), and the operation used for template merging. Also, the rule defines variable renaming. Figure 1 contains three rule examples. Rule 1 puts templates T1 and T2 into the sequence, renames variable a to p, specializes variable b to constant 1, and produces template T4. This combining does not have data dependencies (and therefore does not use port connections). Rule 3 nests template T5 in T6. In this case, data dependency between T5 and T6 is created (and therefore port a is connected to port n), variable b is renamed to n, and variable c is specialized to constant 2. The assembled segment is a temporary

result. To be accepted as a result, its time complexity must be calculated. The following section describes the time complexity calculation.

3.2. Time complexity calculation

Time complexity calculation differs in two cases: when assembling is done using two templates with and without data dependencies between them. The following sections describe the time complexity calculation in both cases.

Assembling without data dependencies If data dependencies are not present, the time complexity calculation is straightforward. For the sequencing and selecting operations, the resulting complexity is a sum of the two complexity classes. In the case of nesting, the resulting complexity is a product of two complexity classes. Therefore, starting with simple loops and assignment statements, we can assemble segments with the complexity classes described with Equation 5.

$$O(n^i \cdot log^j(n) \cdot sqrt(n)^k), \text{where } i, j, k \geq 0\,. \tag{5}$$

Additionally, complexity classes calculated with data dependencies can be mutually multiplied or multiplied with complexity classes described with Equation 5. All these possibilities provide significant coverage of important complexity classes (denoted in Section 2.1) and other non-trivial complexity classes.

If the assignment statements or the if statements (with constant complexity) are nested inside the simple loop and data dependencies are not present, the result is an enhanced loop segment. Templates T6 and T8 (Figure 1) are examples of enhanced loops.

Assembling with data dependencies If a segment is assembled from two templates with data dependencies between them, time complexity calculation is challenging, and we introduce some constraints to make the calculation possible. When data dependencies are present, the algorithm for time complexity calculation requires a defined segment structure such that the time complexity can be calculated analytically. This is not a limiting factor because support for more data dependencies can be easily added when necessary. We support the following three cases of template assembling with data dependencies:

1. One or more assignment statements are nested inside one empty *while* loop. This case is used for assembling the simple loop segments when the system is started. Empty *while* loop and assignment statements are predefined.
2. One or more assignment statements are nested inside one simple loop.
3. Two enhanced loops are sequenced or nested.

In the first two cases, the newly created loop body must comply with one of the four supported simple loop bodies (Listing 3.1). This implies that the assembled loop is a simple loop segment, and its time complexity can be determined, as described in Section 3.1. Figure 1 illustrates this situation. Template T5, having the O(1) complexity, is nested into template T6, width $O(n)$ complexity. Template T5 changes the loop bound of the T6. Consequently, the resulting segment has $O(log(n))$ complexity. The third case is described in the following paragraphs.

Sequencing with data dependencies Listing 3.3 describes the supported structure of the sequence of the two enhanced loop segments.

```
loop_sequencing   : =
    (init(a0,b0) init(v0,v1,v2) (s0=0;|s0=1;) v3=c0;)*
    while(a0<b0){
      (body(a0,b0,v0,v1,v2)
      (s0=s0+o0*v3;|s0=s0*o0*v3;|s0=s0+z0;|s0=s0*o0;))*
    }
    (init(a1,b1) init(v4,v5,v6) b1=s0;)*
    while(a1<b1){
      body(a1,b1,v4,v5,v6)
    }
```

Listing 3.3. Grammar rule that describes two enhanced loop segments put into a sequence

The left-side variable from that statement (marked as s_0) is used as an upper loop bound in the second segment, which creates a data dependency. Let the $O(f(n))$ be the complexity of the first segment in sequence and $O(g(n))$ be the complexity of the second segment in sequence. The resulting complexity calculation is given in Table 3. There are four ways to calculate the overall complexity, depending on the chosen statement.

Table 3. Asymptotic value of the variable s_0 after the first segment is executed and the resulting complexity of the sequence of two enhanced loops

Statement	Variable s_0	Resulting complexity
$s_0 = s_0 + o_0 * v_3$	$\approx f(n^2)$	$max(O(f(n), O(g(f(n^2)))))$
$s_0 = s_0 * o_0 * v_3$	$\approx f(n!)$	$max(O(f(n), O(g(f(n!)))))$
$s_0 = s_0 + z_0$	$\approx f(n)$	$max(O(f(n), O(g(f(n)))))$
$s_0 = s_0 * o_0$	$\approx f(o^n)$	$max(O(f(n), O(g(f(o^n)))))$

Nesting with data dependencies We describe three supported cases when two enhanced loop segments are nested, and data dependency is present between them. These three cases lead to three different ways of resulting complexity calculation. The data dependencies are created with the assignment statements placed inside the outer loop and before the inner loop. These statements modify the bounds of the inner loop in a controlled manner. Grammar in Listing 3.4 describes these three cases.

Let the $O(f(n))$ be the outer loop complexity and $O(g(n))$ be the inner loop complexity. In the first case, inner loop bounds are constants, which implies that the inner loop does not impact the complexity calculation anymore, and the resulting complexity is $O(f(n))$.

In the second case, we use multiplication and division to modify the inner loop bounds, and the resulting complexity is described with Equation 6.

$$\sum_{i=1}^{f(n)} g(o_0^i) \tag{6}$$

In the third case, we use addition and subtraction for the same purpose. The resulting complexity calculation for this case is described with Equation 7.

$$\sum_{i=1}^{f(n)} g(i \cdot z_0) \tag{7}$$

These formulas cannot be analytically calculated for the arbitrary functions $f(n)$ and $g(n)$. Therefore, we use the following functions for $f(n)$ and $g(n)$: n, $log(n)$ and $sqrt(n)$. These functions are supported complexity classes of used enhanced simple loop segments (as described in Section 3.1). The resulting complexity is given in Table 4 and Table 5.

```
loop_nesting  :=  loop_nesting1 | loop_nesting2 | loop_nesting3
loop_nesting1 :=  init(a0,b0) init(a1,b1) init(v0,v1,v2) v6=c0
        while(a0<b0){
          body(a0,b0,v0,v1,v2)
          init(v3,v4,v5)
          ((a1=z1; b1=z0;) | (a1=v6; b1=v6;))
          while(a1<b1){
            body(a1,b1,v3,v4,v5)
          }
        }
loop_nesting2 :=  init(a0,b0) init(a1,b1) init(v0,v1,v2)
        while(a0<b0){
          body(a0,b0,v0,v1,v2) init(v3,v4,v5)
          ((a1=a1*o0; b1=n;) | (b1=b1/o0; a1=c1;))*
          while(a1<b1){
            body(a1,b1,v3,v4,v5)
          }
        }
loop_nesting3 :=  init(a0,b0) init(a1,b1) init(v0,v1,v2)
        while(a0<b0){
          body(a0,b0,v0,v1,v2)
          init(v3,v4,v5)
          ((a1=a1+z0; b1=n;) | (b1=b1-z0; a1=c2;))*
          while(a1<b1){
            body(a1,b1,v3,v4,v5)
          }
        }
```

Listing 3.4. Grammar rules that describe two nested enhanced loops

Table 4. Lookup table for the complexity calculation by Equation 6

f(n) / *g(n)*	n	$log(n)$	$sqrt(n)$
n	o^n	n	$o^{sqrt(n)}$
$log(n)$	n^2	$log(n)^2$	n
$sqrt(n)$	$sqrt(o^n)$	$sqrt(n)$	$sqrt(o^{sqrt(n)})$

Table 5. Lookup table for the complexity calculation by Equation 7

f(n) / *g(n)*	n	$log(n)$	$sqrt(n)$
n	n^2	$log(n)^2$	n
$log(n)$	n	$log(n) \cdot log(log(n))$	$sqrt(n) \cdot log(sqrt(n))$

4. Implementation details

We present the architecture of the proposed system for source code assembling based on the defined templates and rules. The system can produce many resulting code segments from a limited set containing initial templates and rules. This is important for the simplicity of usage and for controlling the correctness of each output.

4.1. System overview

The system generates various source code segments using the set of available segments and combining operations. It has a modular structure. Figure 2 presents system modules. These modules are the graphical user interface (GUI), template assembling module, rule assembling module, template database, rule database, rule execution, complexity calculation module, result segment processor, and segment-assembling manager.

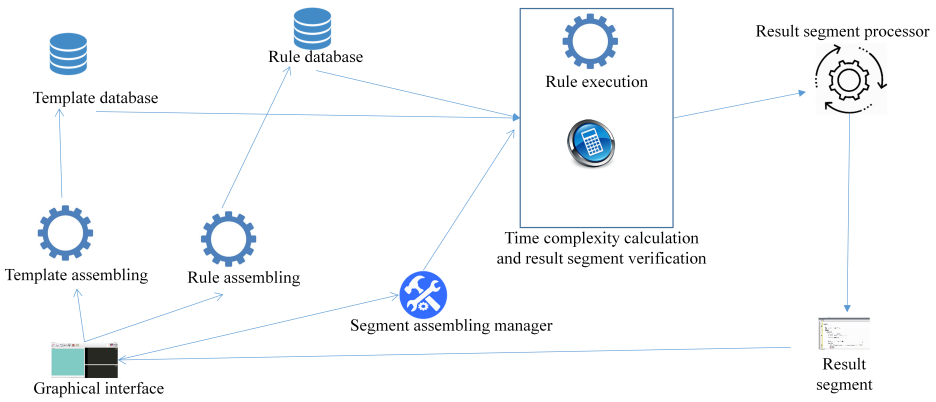


Fig. 2. System overview

Rule and template assembling modules are directly connected to the databases and the GUI. The template assembling module collects information from the GUI and then builds the template intermediate representation, which is further downloaded to the database. Similarly, the rule assembling module gets information from the GUI, creates the rule representation, and saves it within the corresponding database.

4.2. Graphical user interface

Graphical user interface is introduced in [40], and has been constantly improved. It serves for creating custom user templates. These templates can be used to enrich the template database or to test the system. Figure 3 displays a template assembling GUI after starting the tool.

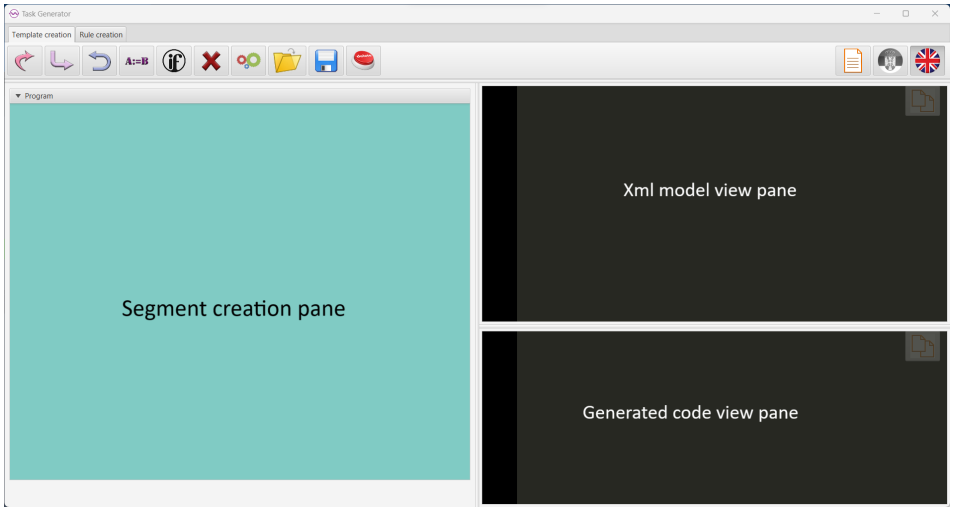


Fig. 3. GUI for template assembling

The template assembling GUI has the following parts: the toolbar, segment creating pane, XML model view pane, and generated code view pane. The toolbar contains commands for adding segment building blocks (loops, expressions, and selections), deleting segments, generating code from graphical representation, saving to a file and copying the generated code, and selecting an interface language (English and Serbian are currently supported). The click on the creation pane adds or removes a segment, and depending on the position of the click, we can achieve nesting or sequencing. The XML model view pane displays the intermediate implementation of the generated segment, while the generated code view pane displays the final source code. Rule assembling GUI is implemented similarly. It has a similar role as the template assembling GUI and enables creating custom rules and testing the system.

4.3. Rule execution

The rule execution module is the core of the system. Figure 4 presents the steps of the rule execution process.

Two operands used during the rule execution process can be templates, rules, or their combination. If an operand is a template, it is forwarded to the next step. If the operand is a rule, it is executed, and the output is delivered to the next step. Operation execution

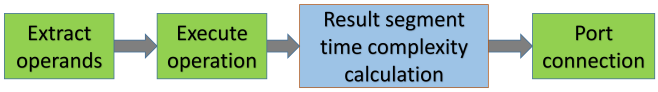


Fig. 4. Rule execution diagram

is straightforward. The *nesting* operation nests the second operand into the first. The *sequencing* operation creates a sequence where the second operand follows the first. The *selecting* operation puts the first operand into the *then* branch and the second operand into the *else* branch. *Else* branch may not exist.

The most complex step is the time complexity calculation of the resulting segment, verified according to the theory and grammar described in Section 3. The last step is establishing data dependencies using port connection, when needed. The rule definition contains mappings between the first segment's output ports and the second segment's input ports. These mappings are executed via variable renaming.

The output of rule execution is an intermediate program segment representation, which can be stored in the template database and used as an input to another rule. The result segment processor takes the intermediate segment and converts it to the source code of the target language. Currently, we support C and C++ programming languages, but other programming languages can be easily added to the system. During the processing, the resulting segment is obfuscated [44]. Obfuscation changes the resulting source code, without changing its time complexity, by modifying variable names not included in data dependencies. As a result, the number of generated segments is additionally increased.

The segment assembling manager module is responsible for implementing higher-level assembling strategies, which use the current template and rule mechanism to generate many segments. There are two envisioned scenarios for higher-level assembling strategies. The first is to assemble a set of segments of some desired (target) time complexity. The other one is different: starting from the seed set of basic templates and rules, the system should assemble a set of segments and let the system calculate their time complexity. A large number of segments can be further used in online examinations and CAT (computer-aided testing).

5. Evaluation and discussion

We evaluated our proposed solution in several contexts. First, similarly to some recent studies from the open literature, we used a machine learning approach to verify code segment complexity. Second, we presented our experiences of using assembled segments in exams and discussed them with the help of field experts in the form of an interview on the topic.

5.1. Evaluated code segments

To evaluate briefly and concisely, we focused on the four most significant complexity classes: $O(log(n))$, linear $O(n)$, linear logarithmic $O(n \cdot log(n))$, and quadratic $O(n^2)$. We chose two representative code segments for each class: one hand-crafted (Listings 5.1,

5.3, 5.5, and 5.7), typically given as an exam question or presented in lectures, and one assembled with our system (Listings 5.2, 5.4, 5.6, and 5.8).

It can be seen that some of the hand-crafted code segments represent well-known algorithms, such as binary search and insertion sort, which are often taught as examples in programming and algorithms and data structure courses. Those examples are shown in Listing 5.1 and Listing 5.7. The other two hand-crafted code segments do not represent any particular algorithm but were assembled with target complexity in mind. In Listing 5.3, it can be noted that a "dead" loop has been inserted into the code with no iterations to confuse the less knowledgeable students.

```
d = 0, g = n-1;
while (d <= g) {
  s = (d + g) / 2;
  if (arr[s] == k)
    break;
  else if (k < arr[s])
    g = s - 1;
  else
    l = s + 1;
}
```

Listing 5.1. Hand-crafted code segment with time complexity $O(log(n))$ which represent binary search algorithm

```
f = 0;
a = 1;
b = n;
while (a<b)
{
  a = a*2;
  f = f+1;
}
d = f;
c = 1;
while (c<d)
{
  d = d-1;
  c = c*2;
}
```

Listing 5.2. Assembled code segment with time complexity $O(log(n))$

Assembled segments with our tool follow the rules presented in Section 3.2. The segments are slightly longer than those hand-crafted, mainly because the processor uses more intermediate variables to assure correctness. Compilers-based techniques (such as constant folding, constant propagation, strength reduction, copy propagation, common subexpression elimination, and arithmetic simplification [1]) could further simplify the code.

```
i = s = 0;
while (i < n) {
  i += 1;
  j = m = 0;
  for (; j <= m; ++j) {
    m *= i;
      s++;
  }
  s--;
}
```

Listing 5.3. Hand-crafted code segment with time complexity $O(n)$. Note the dead inner loop

```
a = 1;
b = n;
do
{
  a = a+4;
}
while (a<b);
a = 1;
b = n;
while (a<b)
{
  b = b-24;
}
```

Listing 5.4. Assembled code segment with time complexity $O(n)$

```
s = 0;
for (i = 0; i < n; i++) {
  j = 1;
  while (j < i) {
    s += i * j;
    j *= 2;
  }
}
```

Listing 5.5. Hand-crafted code segment with time complexity $O(n \cdot log(n))$

```
a = 1;
b = n;
do
{
  b = b/2;
  c = 1;
  d = n;
  do
  {
    d = d-1;
  }
  while (c<d);
}
while (a<b);
```

Listing 5.6. Assembled code segment with time complexity $O(n \cdot log(n))$

```
for (int i = 1; i < n; ++i ) {
  int k = a[i], j = i - 1;
  while ( ( j > 0 ) && a[j] > k ) {
    a[j + 1] = a[j];
    j = j - 1;
  }
  a[j + 1] = k;
}
```

Listing 5.7. Hand-crafted code segment with time complexity $O(n^2)$ which represents insertion sort algorithm

```
a = 1;
b = n;
while (a<b)
{
  d = n;
  c = 1;
  while (c<d)
  {
    c = c+13;
    d = d-64;
  }
  b = b-24;
}
```

Listing 5.8. Assembled code segment with time complexity $O(n^2)$

5.2. ChatGPT-based evaluation

We used ChatGPT [38], a language model-based chatbot, to assess several hand-crafted and assembled code segments. Although the validity of such an evaluation in different areas is widely discussed [47], [2], several recent studies [55], [9], [29] suggest the ability of AI tools to provide explanations, examples, and guidance on complex programming tasks. In addition, machine learning techniques can be used for time and space complexity prediction [53], [34], [52].

We asked ChatGPT to answer the following two questions for each of the code segments given in Section 5.1:

- What is the time complexity of the following piece of code?
- Is this piece of code written by humans or generated by an algorithm?

The snapshot of the whole conversation with ChatGPT can be found at the following link, but also on the GitHub project [42]. Conclusions drawn from the ChatGPT experiment are given in the following paragraphs.

Regarding the first question, ChatGPT successfully calculated the time complexity for all given segments, except for the hand-crafted code segment from Listing 5.3 and the computer-assembled code segment from Listing 5.8, where it firstly gave wrong answers. However, with additional inputs and hints, ChatGPT eventually gave correct answers. As mentioned, the code from Listing 5.3 has a "trick" in the form of a dead loop, which confused the chatbot. The piece of code given in Listing 5.8 indirectly uses the dimension of the problem n through intermediate variables b and d. It confuses ChatGPT, as it could also deceive an inattentive human reader.

Based on the answers to the second question, we conclude that ChatGPT does not imply that any code segment was machine-assembled. A typical answer resembles: *"This code appears to be something a human might write. It follows typical programming patterns using constructs like loops and variable manipulation. There are no obvious signs of unusual patterns or styles that would suggest an algorithm generated it"*. It shows that from this perspective, we managed to imitate the humans in the way they assemble similar code segments. ChatGPT also recognized well-known binary search and insertion sort algorithms as human-written.

5.3. Real-life experiences

Instead of the hand-crafted counterparts, code segments generated by our system have been occasionally used as exam questions in the Programming 2 course at the University of Belgrade, School of Electrical Engineering, for the past three years. Typically, they were used two or three times out of five exam periods in which the exam can be taken. There was no observable change in students' success on these questions. Therefore, students find code segments generated by our system equally difficult to solve.

To assess our solution from an instructor's perspective, we organized a focus group of four field experts engaged in courses that deal with code complexity calculation. We described our approach and presented the participants with the same code segments given in Section 5.1. We asked them for comments on the following general questions:

- How much time is needed to assemble a code segment of a target complexity as an exam question?
- How useful is the presented solution in the context of preparation of exam questions?
- For the presented pieces of code, to distinguish whether they were written by humans or generated by an algorithm?

The participants agreed that the effort needed to assemble a code segment as an exam question greatly depends on its time complexity and the desired structure of the code. On average, it varies from 15 to 30 minutes, together with final verification. Taking a code segment from some predefined repository or a book significantly reduces that time, as only final verification and some code obfuscation are needed. In that context, a presented software solution can be beneficial to produce a large code base and avoid question repeating. They also pointed out that the offered approach guarantees target complexity, which reduces the time needed for verification and guarantees correctness.

The participants were more successful in determining whether the code segment was machine-assembled or hand-crafted. They were correct for all hand-crafted cases and three out of four machine-assembled cases. However, they mostly noted that this is due to the unusual naming conventions for the variables used in assembled segments and the generally excessive use of variables, which can be corrected in the further developments of the system.

6. Related work

As the most fundamental concepts of programming, analysis of algorithms and computational complexity have been studied from the early days of computer science. On the

other hand, recently, automated test assembling and evaluation have attracted significant research interest because introductory programming courses have become massive. We provide information on the related research and compare it with our approach.

Algorithm time complexity is a topic of various studies. There are two different approaches: empirical (profiling) and analytical. In the first approach, a program is instrumented and executed, possibly several times and with various inputs. Instrumentation is achieved by inserting additional instructions into the source code or binary file, and these instructions observe program behavior during execution. *Valgrind* is a tool that does extensive dynamic binary instrumentation [35]. *Gprof* is a tool that does a combination of compile-time instrumentation and runtime sampling [20]. Based on the collected empirical data, time complexity can be estimated. Further, profiling can help with program optimization. McGeoch, Precup, and Cohen developed a heuristic algorithm for finding upper bounds among datasets [31]. A dataset is represented by the two vectors X and Y, such that $Y = f(X)$, where f is an unknown function. The problem is to analyze X and Y and estimate complexity bound $O(f(X))$.

S. F. Goldsmith, A. S. Aiken, and D. S. Wilkerson [19] developed a profiling software called *trend-prof*. It constructs empirical computational complexity models. The authors showed that for real-world programs, linear and exponential complexity models are enough for roughly modeling execution time. We decided to use a formal approach because empirical complexity calculation involves compiling, linking, and executing the generated segment numerous times with extensive input data. In addition, our software should generate many segments, which would take too long if empirical complexity calculation is used.

In the analytical approach, the time complexity of the program is analyzed statically, i.e., without executing the program. This approach might give more precise results than profiling but is generally undecidable as it requires calculating loop summaries symbolically. J. Gustafson, A. Ermedahl, C. Sandberg, and B. Lisper [22] presented a method for deriving a static worst-case execution time using symbolic execution. To estimate the time, they estimate the upper bound on the number of loop iterations. The estimated time is a number, while our work uses symbolical values for complexity functions. Research on symbolic execution [54] is based on the computation of loop summaries for loops along acyclic paths leading to the target location. It demonstrates that the usage of the symbolic execution can lead to a very complex calculation that requires SMT [6] solvers usage. S. Gulwani, K. K. Mehra, and T. Chilimbi [21] developed a technique for computing symbolic bounds on the number of statements as a function of scalar input and user-defined functions that describe input data structures. They add counters to the source code, and the modified code is symbolically executed and generates symbolic bounds. However, their algorithm may not calculate the result in some cases. The mentioned approaches use extensive symbolic computation, where entire programs are subject to execution. We use a symbolic execution-based technique only when we combine two program segments with data dependencies. Instead of using SMT solvers, we constrained our software to the cases when symbolic calculation can be done analytically, and then precalculated results are incorporated into the software as lookup tables.

There are various approaches to automated test assembling and evaluation. One approach is to choose the questions from the database considering various parameters, such as the examination area, question difficulty, the appearance of the previous exams, etc.

That kind of software system often uses artificial intelligence, logic programming, and genetic algorithms [10], [23], [59]. Another approach is the question parameterization [16], [17], [18]. If the questions are used for programming exams, parameterization can be done with the change of the input data, which results in a change in the output. Parameterized parts can be marked or chosen randomly. Our software also uses question parameterization during the assembling process.

The idea for the source code segments assembling process is based on A. W. Bierman's work [8]. He presented program segments as a sequence of abstract expressions derived from the rules applied to the specific source code. Our approach is the reverse: starting from the set of defined templates and rules, templates are combined with the rules, and as a result, the segments are assembled.

The evaluation can be automated when the examination is done using pencil-and-paper tests or online. In the case of pencil-and-paper tests, answers are written on particular paper forms, which are then scanned and further processed by specialized software [33]. Online examination is usually performed by popular learning management systems such as Blackboard, Canvas, and Moodle [15]. Different techniques are used for automated evaluation [3], [24]. New approaches combine several techniques [57]. For example, M. V. Janicic, M. Nikolic, D. Tosic, and V. Kuncak developed the software tool for automated grading of students' assignments, using testing, software verification, and control flow graph similarity measurement [58]. Our system automatically generates correct answers. Therefore, it can evaluate generated questions.

The difficulty levels of generated questions are similar, although the questions look different to the students. In our earlier study [43], we have shown that there is a noticeable difference between textual similarity of the generated segments that was additionally verified using the JPlag tool [46] for source code similarity detection. On the other hand, the semantic similarity of the generated segments was high.

The cheating problem can be present no matter how exams are organized. With the help of software tools, teachers can prevent cheating in different ways. One way is to check if the solutions offered by students are a product of cheating. This check can be done with various tools for software comparison [49], [46], which detect similar solutions. Another way is trying to prevent cheating before the examination even starts. This prevention can be achieved by assembling the exam questions such that each student is given a different set of questions with similar characteristics. This is especially adequate for computer-aided tests but can also be used for paper tests. Our approach supports this solution.

Finally, the approach to source code assembly with defined time complexity presented in this study has several limitations. Although a broad scope of segments of different time complexities can be produced, they are still limited to a closed set based on the defined grammar. However, all practically important examples of time complexities found in the most important scholarly books from the field [12], [14], [39], [25] can be produced. The only notable exception is the support for code segments that contain recursion, which is not supported by our current approach and will be the topic for future research.

7. Conclusion

This paper presents a novel approach to source code assembly for defined time complexity. A formal, grammar-based method models a broad class of code segments with different time complexities. Code templates and rules guide the assembly process. In contrast with machine learning approaches, our approach guarantees the time complexity correctness of the generated code segment.

The implemented system can be used both as an educational tool during the teaching process and as a helper tool for the automated assembly of exam questions. The assembling algorithm makes it possible to follow the assembling process so students can easily understand the presented time complexity.

There are several directions for future work. The core of the system can be further expanded to include additional, rarely used classes of time complexities that are not present in the system. Produced segments can be assigned weights based on defined metrics, which would facilitate the final choice of segment by the instructor.

From our experience, students find time complexity calculation difficult. In that sense, any effort that can help students learn it step by step can be beneficial. One of the apparent directions for future work is to make our tool suitable for educational purposes, which follows the recommendations for new approaches in learning code complexity analysis [28] and a general call for innovative technologies in education [37].

References

1. Aho, A.V., Lam, M.S., Sethi, R., Ullman, J.D.: Compilers: Principles, Techniques, and Tools (2nd Edition). Addison-Wesley Longman Publishing Co., Inc., USA (2006)
2. Aiyappa, R., An, J., Kwak, H., Ahn, Y.Y.: Can we trust the evaluation on chatgpt? arXiv preprint arXiv:2303.12767 (2023)
3. Ala-Mutka, K.M.: A survey of automated assessment approaches for programming assignments. Computer science education 15(2), 83–102 (2005), https://doi.org/10.1080/08993400500150747
4. Alon, U., Zilberstein, M., Levy, O., Yahav, E.: code2vec: Learning distributed representations of code. Proceedings of the ACM on Programming Languages 3(POPL), 1–29 (2019), https://doi.org/10.1145/3290353
5. Arora, S., Barak, B.: Computational complexity: a modern approach. Cambridge University Press (2009)
6. Barrett, C., Tinelli, C.: Satisfiability modulo theories, Handbook of Model Checking. Springer, Cham (2018)
7. Bentley, J.L., Haken, D., Saxe, J.B.: A general method for solving divide-and-conquer recurrences. ACM SIGACT News 12(3), 36–44 (1980), https://doi.org/10.1145/1008861.1008865
8. Biermann, A.W.: A simple methodology for studying program time complexity. Computer Science Education 1(4), 281–292 (1990), https://doi.org/10.1080/0899340900010402
9. Biswas, S.: Role of chatgpt in computer programming. Mesopotamian Journal of Computer Science 2023, 8–16 (2023)
10. Bošnjaković, A., Protic, J., Bojić, D., Tartalja, I.: Automating the knowledge assessment workflow for large student groups: A development experience. The International journal of engineering education 31(4), 1058–1070 (2015)

11. Conway, J.: Unpredictable iterations. The Ultimate Challenge: The 3x+ 1 Problem pp. 49–52 (1972)
12. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to algorithms. MIT press (2022)
13. Dagienė, V., Jevsikova, T., Stupurienė, G., Juškevičienė, A.: Teaching computational thinking in primary schools: Worldwide trends and teachers' attitudes. Computer Science and Information Systems 19(1), 1–24 (2022)
14. Dasgupta, S., Papadimitriou, C.H., Vazirani, U.: Algorithms. McGraw-Hill Education (2006)
15. Draskovic, D., Misic, M., Stanisavljevic, Z.: Transition from traditional to lms supported examining: A case study in computer engineering. Computer Applications in Engineering Education 24(5), 775–786 (2016), https://doi.org/10.1002/cae.21750
16. Geerlings, H., van der Linden, W.J., Glas, C.A.: Optimal test design with rule-based item generation. Applied Psychological Measurement 37(2), 140–161 (2013), https://doi.org/10.1177/0146621612468313
17. Glas, C.A., van der Linden, W.J.: Computerized adaptive testing with item cloning. Applied Psychological Measurement 27(4), 247–261 (2003), https://doi.org/10.1177/0146621603027004001
18. Glas, C.A., van der Linden, W.J., Geerlings, H.: Estimation of the parameters in an item-cloning model for adaptive testing. In: Elements of adaptive testing, pp. 289–314. Springer (2009)
19. Goldsmith, S.F., Aiken, A.S., Wilkerson, D.S.: Measuring empirical computational complexity. In: Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering. pp. 395–404 (2007), https://doi.org/10.1145/1287624.1287681
20. Graham, S.L., Kessler, P.B., McKusick, M.K.: Gprof: A call graph execution profiler. ACM Sigplan Notices 17(6), 120–126 (1982), https://doi.org/10.1145/872726.806987
21. Gulwani, S., Mehra, K.K., Chilimbi, T.: Speed: precise and efficient static estimation of program computational complexity. ACM Sigplan Notices 44(1), 127–139 (2009), https://doi.org/10.1145/1594834.1480898
22. Gustafsson, J., Ermedahl, A., Sandberg, C., Lisper, B.: Automatic derivation of loop bounds and infeasible paths for wcet analysis using abstract execution. In: 2006 27th IEEE International Real-Time Systems Symposium (RTSS'06). pp. 57–66. IEEE (2006), https://doi.org/10.1109/RTSS.2006.12
23. Hernández-Del-Olmo, F., Gaudioso, E.: Autotest: An educational software application to support teachers in creating tests. Computer Applications in Engineering Education 21(4), 636–640 (2013), https://doi.org/10.1002/cae.20508
24. Ihantola, P., Ahoniemi, T., Karavirta, V., Seppälä, O.: Review of recent systems for automatic assessment of programming assignments. In: Proceedings of the 10th Koli calling international conference on computing education research. pp. 86–93 (2010), https://doi.org/10.1145/1930464.1930480
25. Knuth, D.E.: The art of computer programming. Vol. 1: Fundamental algorithms. Addison-Wesley (1968)
26. Knuth, D.E.: The analysis of algorithms. In: Actes du Congres International des Mathématiciens (Nice, 1970). vol. 3, pp. 269–274 (1970)
27. Knuth, D.E.: Big omicron and big omega and big theta. ACM Sigact News 8(2), 18–24 (1976), https://doi.org/10.1145/1008328.1008329
28. Licht, B.: Obstacles in learning algorithm run-time complexity analysis. University of Nebraska at Omaha (2022), theses/Capstones/Creative Projects. 193. https://digitalcommons.unomaha.edu/university_honors_program/193
29. Lo, C.K.: What is the impact of chatgpt on education? a rapid review of the literature. Education Sciences 13(4), 410 (2023)
30. McCabe, T.J.: A complexity measure. IEEE Transactions on software Engineering SE-2(4), 308–320 (1976), https://doi.org/10.1109/TSE.1976.233837

31. McGeoch, C.C., Cohen, P.: How to find big-oh in your data set (and how not to). In: Sixth International Workshop on Artificial Intelligence and Statistics. pp. 347–354. PMLR (1997)
32. MIDAS Research Laboratory, I.D.: Corcod: Code runtime complexity dataset (2019), [Online]. Available: https://github.com/midas-research/corcod-dataset (current October 2023)
33. Mišić, M., Lazić, M., Protić, J.: A software tool that helps teachers in handling, processing and understanding the results of massive exams. In: Proceedings of the Fifth Balkan Conference in Informatics. pp. 259–262 (2012), https://doi.org/10.1145/2371316.2371370
34. Moudgalya, K., Ramakrishnan, A., Chemudupati, V., Lu, X.H.: Tasty: A transformer based approach to space and time complexity. arXiv preprint arXiv:2305.05379 (2023)
35. Nethercote, N., Seward, J.: Valgrind: a framework for heavyweight dynamic binary instrumentation. ACM Sigplan notices 42(6), 89–100 (2007), https://doi.org/10.1145/1273442.1250746
36. Nešković, R.: Predicting time complexity of algorithms on pascal programming language using machine learning techniques. University of Belgrade, School of Electrical Engineering (2021), master thesis (in Serbian)
37. Omonayajo, B., Al-Turjman, F., Cavus, N.: Interactive and innovative technologies for smart education. Computer science and information systems 19(3), 1549–1564 (2022)
38. OpenAI: ChatGPT, https://chat.openai.com/
39. Papadimitriou, C.H.: Computational complexity. In: Encyclopedia of computer science, pp. 260–265. John Wiley and Sons (2003)
40. Pešić, Đ., Mišić, M., Protić, J., Vujošević Janičić, M.: Sistem za generisanje programskih segmenata za ispitivanje u oblasti vremenske slozenosti algoritama. In: ETRAN 2017, Kladovo, Serbia (2017), (in Serbian)
41. Pešić, Đ., Mišić, M., Protić, J., Vujošević Janičić, M.: Prototype implementation of segment assembling software. SJEE 15(1), 71–83 (2018)
42. Pešić, Đ., Mišić, M., Vujošević Janičić, M., Protić, J.: Task generator project, https://github.com/djpesic/TaskGenerator
43. Pešić, Đ., Protić, J., Vujošević Janičić, M., Mišić, M.: Ispitivanje kvaliteta softverski generisanih segmenata u oblasti vremenske složenosti algoritama za automatizovano sastavljanje ispita. In: XXV Skup Trendovi razvoja: Kvalitet visokog obrazovanja, Kopaonik (2019), (in Serbian)
44. Pešić, Đ., Purić, S., Mišić, M., Protić, J.: Softversko generisanje pitanja i odgovora iz oblasti analize složenosti algoritama za testove na kursevima programianja. In: XXII Skup Trendovi razvoja: Nove tehnologije u nastavi, Zlatibor (2016), (in Serbian)
45. Pešić, Đ., Purić, S., Mišić, M., Protić, J.: Softversko generisanje programskih segmenata baziranih na strategijama modeliranim pomoću xml-a. In: ETRAN 2016, Zlatibor, Serbia (2016), (in Serbian)
46. Prechelt, L., Malpohl, G., Philippsen, M., et al.: Finding plagiarisms among a set of programs with jplag. Journal of Universal Computer Science 8(11), 1016 (2002), http://dx.doi.org/10.3217/jucs-008-11-1016
47. Rudolph, J., Tan, S., Tan, S.: Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? Journal of Applied Learning and Teaching 6(1) (2023)
48. Savic, M., Ivanovic, M., Lukovic, I., Delibašic, B., Protic, J., Jankovic, D.: Students' preferences in selection of computer science and informatics studies–a comprehensive empirical case study. Computer Science and Information Systems 18(1), 251–283 (2021)
49. Schleimer, S., Wilkerson, D.S., Aiken, A.: Winnowing: local algorithms for document fingerprinting. In: Proceedings of the 2003 ACM SIGMOD international conference on Management of data. pp. 76–85 (2003), https://doi.org/10.1145/872757.872770
50. Sedgewick, R., Flajolet, P.: An introduction to the analysis of algorithms. Pearson Education India (2013)

51. ShanghaiRanking: Computer science & engineering. Global Ranking of Academic Subjects (2022), [Online]. Available: http://www.shanghairanking.com/rankings/gras/2022/RS0210 (current July 2023)
52. Siddiq, M.L., Samee, A., Azgor, S.R., Haider, M.A., Sawraz, S.I., Santos, J.C.: Zero-shot prompting for code complexity prediction using github copilot. In: 2023 IEEE/ACM 2nd International Workshop on Natural Language-Based Software Engineering (NLBSE). pp. 56–59. IEEE Computer Society (2023)
53. Sikka, J., Satya, K., Kumar, Y., Uppal, S., Shah, R.R., Zimmermann, R.: Learning based methods for code runtime complexity prediction. In: Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42. pp. 313–325. Springer (2020)
54. Strejček, J., Trtík, M.: Abstracting path conditions. In: Proceedings of the 2012 International Symposium on Software Testing and Analysis. pp. 155–165 (2012), https://doi.org/10.1145/2338965.2336772
55. Surameery, N.M.S., Shakor, M.Y.: Use chat gpt to solve programming bugs. International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290 3(01), 17–22 (2023)
56. Turing, A.M., et al.: On computable numbers, with an application to the entscheidungsproblem. J. of Math 58(345-363), 5 (1936)
57. Vujosevic-Janicic, M., Maric, F.: Regression verification for automated evaluation of students programs. Comput. Sci. Inf. Syst. 17(1), 205–227 (2020), https://doi.org/10.2298/CSIS181220019V
58. Vujošević-Janičić, M., Nikolić, M., Tošić, D., Kuncak, V.: Software verification and graph similarity for automated evaluation of students' assignments. Information and Software Technology 55(6), 1004–1016 (2013), https://doi.org/10.1016/j.infsof.2012.12.005
59. Yildirim, M.: A genetic algorithm for generating test from a question bank. Computer Applications in Engineering Education 18(2), 298–305 (2010), https://doi.org/10.1002/cae.20260
60. Zheng, G., Zhang, X., Wang, R., Zhao, L., Wang, C., Wang, C.: Construction of innovative thinking training system for computer majors under the background of new engineering subject. Computer Science and Information Systems 19(3), 1499–1516 (2022)

Đorđe Pešić is an experienced software developer, with extensive industrial experience in system software development in various domains: compilers, IoT, blockchain, and accessibility software. He received his BSc in 2014, MSc in 2015, and enrolled PhD studies. His focus now is on accessibility and educational software development, with a focus on Rust programming language. He is currently employed at Grid Dynamics.

Milena Vujošević Janičić received a PhD degree in Computer Science from the University of Belgrade in 2013. She is currently an Associate Professor at the Department of Computer Science and Head of the Computer Science Study Program at the Faculty of Mathematics, University of Belgrade. She also works as a principal researcher at Oracle Labs. Her main research interests are in compiler optimizations, automated bug finding, and the application of software verification techniques in different fields.

Marko Mišić received his BSc degree in 2007, his MSc degree in 2010, and his PhD in 2017 in the field of computer engineering and informatics from the University of Belgrade, Serbia. He is an associate professor at the School of Electrical Engineering of

the University of Belgrade. His main fields of interest are parallel programming with an emphasis on GPU computing, image processing, machine learning, social and complex network analysis, as well as educational tools and applications.

Jelica Protić received the PhD degree in electrical engineering from the University of Belgrade in 1999. She is currently a Full Professor and the Head of Department of Computer Engineering and Informatics with University of Belgrade, the School of Electrical Engineering. With Milo Tomasevic and Veljko Milutinovic, she co-authored Distributed Shared Memory: Concepts and Systems (IEEE CS Press, 1997) and presented numerous pre-conference tutorials on this subject. She has long term experience in teaching a diversity of courses in programming languages, as well as the development of various educational software tools. Her research interests include distributed systems, consistency models, complex networks, and all aspects of computer-based quantitative performance analysis and modeling.

Received: July 30, 2023; Accepted: January 23, 2024.

Computer Science and Information Systems 21(3):807–830 https://doi.org/10.2298/CSIS231002019B

Medical Record Information Storage Scheme based on Blockchain and Attribute Role-Based Access Control

Aoao Bian, Dezhi Han, Mingming Cui, and Dun Li

College of Information Engineering at Shanghai Maritime University
Shanghai China
aobian2022@163.com
dezhihan88@sina.com
mmcui@stu.shmtu.edu.cn
li.dun@telecom-sudparis.eu

Abstract. The integration of the Internet with the healthcare industry has ushered in a transformative revolution in the medical field, where efficient and prompt electronic medical records can save precious resources.However, incomplete or tampered electronic medical records have serious implications on patient. Ensuring the complete, accurate storage of medical record information is an urgent problem that needs to be addressed.Therefore, this paper proposes a blockchain-based solution for secure storage and sharing of medical record information, combining the role-based access control (RBAC) with the attribute-based access control (ABAC).Firstly, by utilizing RBAC and ABAC, dynamic fine-grained access control for medical information is achieved based on role differences.Then, the medical record information is stored in the blockchain through chaincode .In addition, Advanced Encryption Standard and Feature-Aware Stateful Routing technology is applied to further enhance security and storage efficiency of the scheme. Experimental results demonstrate that the proposed solution ensures the security and integrity of medical record, while providing efficient information storage and access.

Keywords: Medical record, blockchain, role-based access control (RBAC), attribute-based access control (ABAC), Feature-Aware Stateful Routing (FASR).

1. Introduction

Medical record information is a valuable treasure.Medical records store all the medical records of patients. These medical records are not only convenient for them to grasp their basic physical conditions, but also for doctors to understand the patient's past medical history, so that they can make further treatment plans according to the actual situation of the patient. With the rapid development of the Internet industry, the medical industry is also moving towards informatization, and medical record information has also been digitized, thus evolving into electronic medical record[1],[2].

Compared with traditional paper medical records, electronic medical records are more convenient to store and query, but the current information storage technology still cannot ensure the security of information, so there are still risks in the transmission process of medical records information. More and more medical institutions use centralized storage to ensure the security of medical record information[3]. However, centralized access storage is easy to cause data theft and loss because it stores data in a central location. If there

is a single point of failure in the centralized storage system, all the medical records in the system will be affected, which is likely to lead to the collapse of the medical system, causing immeasurable losses[4]. Another problem with centralized access storage is the inability to share information between different medical institutions. Depending on the hospital, the patient's medical record information is stored in separate storage servers, which are usually only accessible to people with specific permissions internally, and thus do not have any form of interoperability[5]. Most healthcare organizations use security policies to restrict access to and sharing of medical record information. There is a lack of safe and effective medical record information sharing mechanism among institutions[6],[7]. Obviously, this management mode lacks the necessary flexibility and efficiency, which can neither guarantee the security and reliability of medical record information, nor realize the sharing and utilization of medical record information.In fact, medical record information is not only a treasure, but also an important material to understand the patient's physical condition. In order to facilitate the transfer of electronic medical records between systems, most medical institutions adopt different methods to reduce the redundancy of data.

Fortunately, the development of blockchain technology has brought new solutions to address the security storage issue in medical data management[8].However, most of the projects are still in the conceptual stage. To address this issue, according to the characteristics of the information supply chain, this paper uses ABAC and RBAC to divide the access roles in a fine granularity, and combines FASR de-duplication technology to propose a security storage scheme of medical record information based on the blockchain of ABAC, RBAC and Hyperledger Fabric consortium blockchain[9],[10].

The main contributions of this paper are as follows:

1) ABAC, RBAC, and blockchain enables dynamic fine-grained access control for medical information based on role differences. Medical record information is stored in blockchain blocks, and its security is ensured through the formulation of different smart contracts[11].
2) Selecting the appropriate encryption technology to encrypt the medical record information further improves the security of data.
3) FASR deduplication technology is employed to eliminate duplicate medical record information, reducing data redundancy and improving storage efficiency.
4) Simulation experiment results demonstrate that the proposed solution not only ensures the security and data integrity in the processes of medical record storage and sharing, but also improves the efficiency of information storage and access.

The remaining sections of this paper are organized as follows. In Section 2, we summarize the literature related to the program.In Section 3, we provide the necessary background knowledge. In Section 4, we describe the main architecture of the proposed solution and its functional design.And we present the results of the simulation experiments and analyze them in detail in Section 5. Finally, in Section 6, we make concluding remarks on the overall solution and suggest future directions for improvement.

2. Related Work

With the advent of the digital era, the traditional centralized system has been difficult to meet people's needs, and distributed ledger technology, as a decentralized emerging

technology, is gradually recognized and applied by people. The emergence of blockchain technology has allowed people to see the future of decentralization and new opportunities in the digital economy. Blockchain is used in many fields and is a trusted data storage technology[12]. Due to its decentralized and tamper-proof characteristics, this technology is often used for information security storage and information traceability. For this reason, a lot of research is currently being done into this technology. In addition, the technology is expanding into more and more applications[13],[14].

Blockchain is essentially a highly scalable distributed database and its tamper-resistant nature can prevent data from being tampered with, deleted or leaked, etc., thus ensuring the security of medical data. Blockchain has been applied in various fields and is a trusted data storage technology[15],[16]. In terms of copyright information, Wang et al.[17]proposed the use of blockchain to prevent the loss of image data. Han et al.[18] designed a blockchain-based copyright certificate storage and transaction system. Jing et al. [19] presented a blockchain-based code copyright management system. Garilli et al.[20] analyzed the pros and cons of using blockchain and smart contracts in a copyright context. Xiao et al.[21] proposed a blockchain-based algorithm for intellectual property copyright protection.

In the supply chain field,Gao and Tan et al.[22,23] explored the application of blockchain technology in logistics supply chain.Musamih et al.[24] achieved efficient product traceability in healthcare supply chains using smart contracts and decentralization. Saberi et al.[25]discussed the potential applications of blockchain technology and smart contracts in supply chain management, providing insights to overcome obstacles in supply chain management.

In the education field, Chen et al.[26] discussed the prospects of blockchain technology in education. Bhaskar et al.[27] summarized the applications of blockchain technology in the education field. Rii et al.[28] constructed a system using blockchain methods to collect information through digital observation and questionnaire surveys.Li et al.[29] proposed a blockchain-based education record storage and sharing solution called EduRSS.

Blockchain has expanded its application scenarios and produced unexpected effects in the medical field. Radanović and Chen et al.[30,31] provided insights into the application of blockchain in the healthcare domain. Li et al.[10] applied blockchain to the supply chain of medical devices and combine RBAC and ABAC for access control management. Qu et al. [32]proposed a blockchain-based medical information system solution. Liu et al.[33] presented a mechanism for sharing medical resources under the conditions of blockchain technology.Bakro et al.[34]analyzed the security aspects of data stored in the cloud through hybrid security system enabled with blockchain technology using cryptographic algorithms ECC and AES.Yassine et al.[35] discussed a new approach that will allow a smart deletion of data stored in the file system as well as its reference in the Blockchain.

However, in the medical field, most of the above projects are still in the conceptual stage. In this paper, we describe the study of this scheme in detail.

3. Preliminaries

In this section, we provide an introduction to several essential concepts, including Hyperledger Fabric, $ABAC$ and $RBAC$ models, AES encryption algorithm, and $FASR$.

3.1. Hyperledger Fabric

Compared to traditional databases, data in a blockchain is stored across multiple nodes in the entire network, with each node holding a complete copy of the ledger[36]. Hyperledger Fabric is a consortium blockchain that allows only authorized users to join the network[37]. Additionally, Hyperledger Fabric utilizes a channel mechanism, where users with the same channel can access each other's account ledgers, thus ensuring the privacy of sensitive information for those users[38]. Fig.1 illustrates the operational architecture of Hyperledger Fabric.

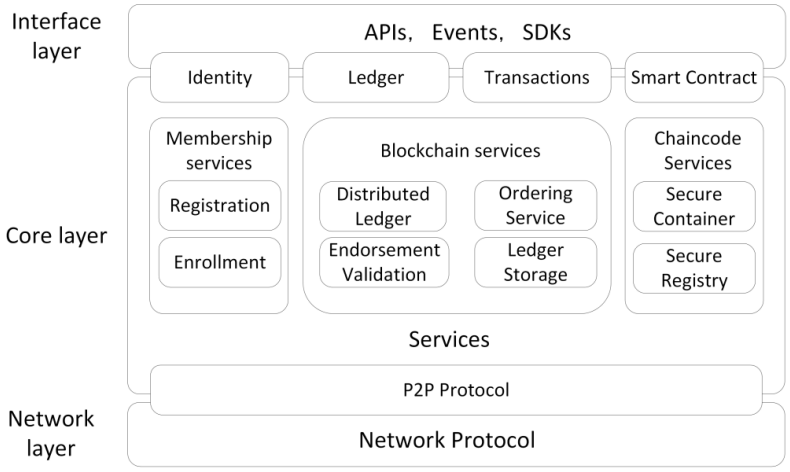


Fig. 1. Hyperledger Fabric's Operational Architecture Diagram

3.2. ABAC and RBAC Models

ABAC

$ABAC$ is an attribute-based access control model that uses attributes to determine whether a user has the right to access resources [39]. Unlike traditional access control models such as Discretionary Access Control (DAC) and Mandatory Access Control (MAC), $ABAC$ emphasizes defining and enforcing policies based on attributes rather than just user identity or resource classification. It utilizes multiple attributes to determine access permissions, making access control more flexible and fine-grained. In healthcare, $ABAC$ can simplify the management of medical records by defining and managing access control policies in a unified location, reducing administrative workload and the

chance of errors. In order to achieve dynamic management of medical records, this paper instantiates the $ABAC$ model as an access control policy called Policy, composed of $AS, AO, AP, andAE$. Table 1 below compares $ABAC$ with MAC and DAC.

Table 1. Comparison of ABAC with MAC and DAC Access Control Models

Categories	Flexibility	Security	Extensibility
ABAC	higher	higher	higher
MAC	lower	higher	lower
DAC	higher	lower	lower

RBAC

In $RBAC$, permissions are granted to specific roles and users are then assigned to corresponding roles. This allows for more precise management of access rights, ensuring that users can only access resources within the scope of their job functions. $RBAC$ offers several advantages compared to other access control models. Firstly, $RBAC$ is more flexible and can better adapt to different organizational structures and security requirements. Secondly, $RBAC$ simplifies the process of managing permissions, making it easier and more secure to allocate and modify access rights. Besides, it enhances system security and scalability[40].

3.3. AES

AES is a symmetric encryption algorithm widely used in data protection and security domains[41],[42]. It is currently recognized as one of the most secure encryption algorithms and is also adopted as one of the confidentiality standards by the U.S. government[43].AES encryption and decryption operations are both performed based on blocks, offering high security and efficiency[44].

The AES algorithm's encryption and decryption operations involve the following four steps:

1) SubBytes:uses a substitution table called the S-Box to replace each byte in the input block with its corresponding S-Box value.
2) ShiftRows: cyclically shifts each row to the left, where the number of shifts depends on the block size and the row number.
3) MixColumns:maps each input column vector in the block to an output column vector, enhancing data irregularity and interactivity.
4) AddRoundKey:performs bitwise XOR between the round key used in the encryption process and the current state matrix.

AES algorithm offers advantages in both security and efficiency and finds extensive applications in various fields such as network communication, data storage, financial transactions, etc. Its key benefits include high security, efficiency, and reliability.

3.4. FASR

$FASR$ is a novel Feature-Aware Stateful Routing method designed to reduce system overhead in distributed environments and maintain a high deduplication rate[45]. The $FASR$ system primarily consists of three working nodes, $Client$, $Master$, and $Data\ nodes$. The $Client$ is responsible for receiving data streams and sending requests for deduplication and recovery. The $Master\ node$ maintains the mapping between data blocks and file fingerprints to provide the metadata required for deduplication and recovery. While the $Data\ nodes$ primarily handle data retrieval and storage.

4. System Model Design

This section mainly introduces the architecture of the medical record information management system based on blockchain and access control. Section 4.1 presents the fundamental framework of the system, while Section 4.2 provides a detailed description of the system design process. The proposed solution's system flow is illustrated in Section 4.3.

4.1. Basic Architecture

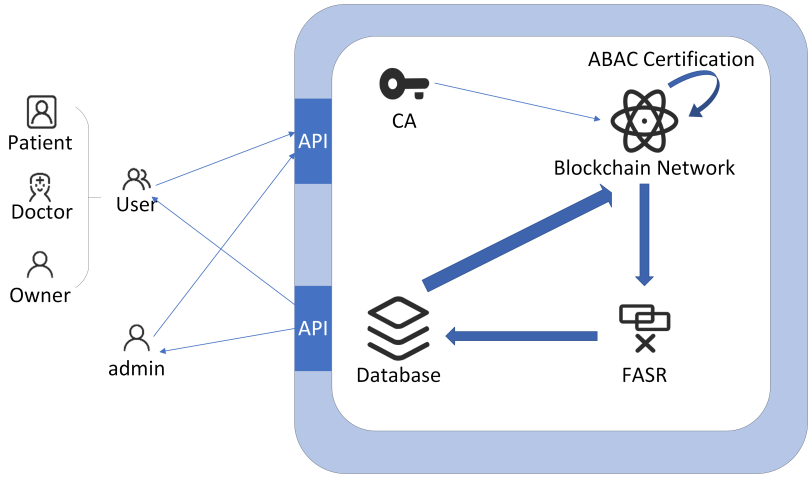


Fig. 2. System Architecture Diagram

The system architecture consists of three essential components, users, the blockchain network, and $FASR\ nodes$. After completing identity authentication, each user is issued a digital certificate by the Certificate Authority (CA). Only users with digital certificates have access privileges to the blockchain system, which assures the fundamental security of the blockchain. After receiving a user's access information, the blockchain utilizes smart contracts to determine the legitimacy of the access information and checks if the user has

the necessary permissions to execute the operation. And then it transmits a corresponding feedback to the user. The detailed architecture of the system is illustrated in Figure2.

The system users mainly consist of backend institutions, administrators, doctors, and patients. Different users have different access privileges. Backend institutions undertakes to package and upload medical record information. Additionally, they play a supervisory role to prevent malicious attacks or illegal operations by users. Administrators are only in charge of creating doctor and patient accounts, and the access permissions for doctor and patient users are system-defined and independent of the administrators. Doctors have permissions to view medical records, add patient information and medical record details, and upload modifications to medical records. As for patients, they can only access their own medical records. Each role has its own responsibilities and cannot exceed their authorized boundaries.

The blockchain network is built on Hyperledger Fabric and primarily implements the following functionalities.

1) The CA issues digital certificates to participating nodes in the blockchain network for identity authentication and communication encryption.
2) Smart contracts automatically upload data to the database after each operation and supervise user access requests based on access control rules.

In this paper, $FASR$ is used to compress medical record data in a lean format and store them in the blockchain database, so as to reduce redundant operations on duplicate data during interactions with the blockchain system.

4.2. Access Control Model Design

In the supply chain of medical record information, a finer-grained distinction of role permissions and dynamic management of different access rights are required. Therefore, in this paper, we adopt the $ABAC$ model as the overall system architecture and define User Attribute (UA), Medical Record Attribute (MA), Permission Attribute (PA), and Environment Attribute (EA) separately. Considering the potential overlapping operations among different users in UA, we introduce the $RBAC$ policy to bundle all relevant operations into Roles and take Role as an attribute in the $ABAC$ model. This design not only achieves dynamic fine-grained access to complex permissions in the traceability supply chain but also resolves the problem of redundant policy storage by utilizing the $RBAC$ model.

RBAC Design

This subsection primarily introduces all user entities in the system. Three identity roles, Administrator, Doctor, and Patient, are designed. Table2 displays the access control permissions for different users.

1) The Administrator is the manager of the entire system, and every user who wants to access the system needs to be granted different user roles by the Administrator. Therefore, the Administrator can assign new users either the Doctor role or the Patient role and provide each user with a unique identification ID.

Table 2. Access Control Permissions for Different Users

	FindByDid	CreDoc	CrePat	AddUit	AddAit	AddTit
Owner		✓	✓			
Doctor	✓			✓	✓	✓
Patient	✓					

2) The Doctor plays a crucial and pivotal role in the entire system, as only doctors have the authority to modify and add patient information. Doctors diagnose patients' conditions, determine their medical record information, and promptly upload the medical records to the system.
3) The Patient is a vital component of the entire system. Patients can access the system and inquire about their medical record information using their own ID information. To ensure system security, doctors can only operate on medical records with the patient's permission, using the medical record identification number.

ABAC Design

Based on the actual application scenario of the medical traceability system, we define each attribute UA, MA, EA, and PA in the ABAC model as follows.

UA denotes the attributes of user entities and its specific attributes are shown in Table3. $UserId$ serves as the unique identifier for each user, and a user has only one $UserId$. $Role$ represents the user's role, which can be Administrator, Doctor, or Patient. Different user roles correspond to different system access permissions. The definition of roles has been elaborated in the $RBAC$ section.

Table 3. UA Attribute Definition Table

Name	Definition
UserId	Unique number for each user
Role	Identification of user identity
Group	The group to which the user belongs

MA represents the entity attributes of medical record information and its detailed attributes are shown in Table4. Did serves as the unique identifier for each medical record and a medical record has only one Did. $Disease$ represents the medical record form in which specific information about the patient's medical history, medication, etc., is recorded.

EA is the environmental information. In our paper, EA's attribute denotes $AllowedTime$, a reasonable amount of time for temporary users to access. If the requested time by a temporary user exceeds the validity period of the permission, the user's operation will be denied.

Table 4. MA Attribute Definition Table

Name	Definition
Did	Unique number for each medical record
OwnerId	Patients with medical records
Name	The patient's name
Age	The patient's age
Disease	medical records
HisUpdate	A list of medical record modifications
HisAdd	A list of medical records is added

PA represents the system's access control policy. If it is true, access is allowed; otherwise, access is denied. Eq1precisely describes the access permission.

$$PA = \begin{cases} true & allow \\ false & deny \end{cases} \tag{1}$$

$ASAO$ shows the mapping relationship between users and medical records. To achieve fine-grained permission allocation, we utilize $ASAO$ as the index for storing access control policies in the blockchain. $ASAO$ is comprised of $AS.UserID$, $AS.Role$, and $AO.Did$, and establishes a one-to-one correspondence between different user identities and medical records. Furthermore, the AES algorithm is employed to encrypt $ASAO$ and the encrypted index is stored in the blockchain as the index for access control policies, so as to achieve the uniqueness of the index. $ASAO$ is calculated based on eq2.

$$index \overset{AES}{\longleftarrow} ASAO \leftarrow \{UA.UserId, MA.Did, MA.UserRole\} \tag{2}$$

The access control policy, Policy, consists of four main components: UA, MA, PA, and EA.

$$Policy = \{UA, MA, PA, EA\} \tag{3}$$

To enhance the efficiency of access control queries and validations, Policy is stored in the blockchain ledger in the form of $\langle Index, Policy \rangle$.

FASR Design

$FASR$ technology is employed to perform deduplication on data that needs to be stored in the blockchain, reducing the storage burden on the blockchain, minimizing data redundancy, and improving storage efficiency. $FASR$ mainly comprises three working nodes, namely the $Client$, the $Master\ node$, and the $Data\ nodes$. The primary workflow in this paper includes the following steps.

1) The $Client$ transfers data to the $Master\ node$.

$$Client \overset{data}{\longrightarrow} Master \longrightarrow node\ feature\ list \tag{4}$$

2) The $Master\ node$ filters out nodes that do not contain the same type of data, based on the node feature table, and selects the node with the highest data type storage capacity.

$$Master \overset{node\ feature\ list}{\longrightarrow} Nodes \tag{5}$$

3) $FASR$ calculates the similarity between the data and the selected node. Based on the similarity and node storage capacity, the target node is chosen.

$$Nodes \overset{similarity}{\longrightarrow} Target\ Node \tag{6}$$

4) Using data similarity and locality, local duplicate data removal is performed, and the processed data is stored in the blockchain.

$$Target\ Node \overset{deduplicate}{\longrightarrow} Data \longrightarrow Blockchain \tag{7}$$

Smart Contract Design

In addition to implementing access control, smart contracts are also related to the storage of medical record information. To achieve a medical record traceability system based on access control, three types of smart contracts, Access Contract (AC), Device Contract (DC), and Policy Contract (PC), are formulated.

1) AC: It implements access control for user requests within the blockchain network. AC verifies user permissions based on the requests submitted by users and their roles. AC serves as an interface for user access. Upon receiving users' request, it performs the necessary authentication and then calls other relevant contracts for specific operations. The AC mainly includes the following methods.
$CheckAccess()$: This method is the core function of AC, responsible for validating the access information sent by users and executing corresponding operations based on specific requests. $CheckAccess()$ can be divided into three steps, confirming input parameters, authenticating access control policies, and executing operation instructions. Firstly, users input the access parameters to submit the operation request to the blockchain. As shown in eq8, $AccessAvg$ is composed of $ASAO$,$Operation$ and $OperationContent$. and are name and content of the operation, respectively.

$$AccessAvg \longleftarrow (ASAO, Operation, OperationContent) \tag{8}$$

Among them, there are six types of operation, namely $FindByDid$, $CrePat$, $CreDoc$, $AddTit$, $AddUit$, and $AddAit$, representing the operations to find medical record information, create a patient user, create a doctor user, add medical record details, add medical record modification information, and add medical record additional information, respectively.

$$AccessAvg \longleftarrow (ASAO, FindByDid) \tag{9}$$

$$AccessAvg \longleftarrow (ASAO, Credoc, UseId) \tag{10}$$

$$AccessAvg \longleftarrow (ASAO, AddUit, Iid, Result) \tag{11}$$

$$AccessAvg \longleftarrow (ASAO, AddTit, Iid, Disease, Medicine) \tag{12}$$

Algorithm 1 Part of CheckAccess(): The Specific Process of Handling User Operations

Require: *Args*
Ensure: *Response or Error*

if $Operations == FindByDid$ **then**
  if $len(Args)! = 2$ **then**
    return $Error : Args\ Invalid$
  end if
  if AS.Role $\notin \langle Docter, Patient \rangle$ **then**
    return $Error : Identity\ Invalid$
  end if
  $MaDid \leftarrow InvokeChaincode(DC, FindByDid, AO.Did)$
  $\langle Resp, Error \rangle \leftarrow UnMamarshal(MaDid)$
  if $Did == NullorError! = Null$ **then**
    return $Error : NofindDid$
  end if
  return $Did.HisAdd$
end if
if $Operations == CreDoc$ **then**
  if $len(Args)! = 3$ **then**
    return $Error : Args\ Invalid$
  end if
  if $AS.Role! = Owner$ **then**
    return $Error : Identity\ Invalid$
  end if
  $NewABACPolicy \leftarrow \langle NewAS, AO, AP, AE \rangle$
  $Reponse \leftarrow InvokeChaincode(PC, AddPolicy, NewABACPolicy)$
  return Reponse
end if
if Operations == AddTit **then**
  if len(Args)!=5 **then**
    return Error:Args Invalid
  end if
  if $AS.Role \notin \langle Docter \rangle$ **then**
    return Error: Identity Invalid
  end if
  $Reponse \leftarrow InvokeChaincode(DC, AddMedicalRe, AO.Did, ArgsAdd)$
  return $Reponse$
end if

As shown in Eq 9,10,11,12, the number of parameters in $AccessAvg$ during the user request process can only be 2, 3, 4, or 5. Therefore, before validating the access request policy, it is essential to check the number of parameters in $AccessAvg$ and reject access with invalid parameters, which significantly improves the execution speed of AC policies.

Next, the value of AP in the AC policy needs to be determined. If it is 1, access is permitted; otherwise, access is denied.

$$Policy.PA = \begin{cases} 1 & allow \\ 0 & deny \end{cases} \tag{13}$$

In the blockchain, different operation permissions are assigned to different roles. After the user are verified, the system further validates the remaining $ABAC$ attributes based on the user's identity and the operation for request. Upon successful verification, the relevant smart contract is automatically invoked to execute the corresponding operations. Different operations require different $ABAC$ attribute to validate and different methods to invoke. The detailed smart contract design is shown in Algorithm 1.

2) DC: It primarily handles the management of medical record information on the blockchain. Its main functionalities include uploading medical record information and adding medical record entries. DC does not directly open operations to patient or doctor users. AC verifies users' queries or modification requests to devices based on the access control rules and then automatically calls DC to update or return the modified data. According to the main functions of the data center, the following methods are designed for it:

   $FindByDid()$: The main function of this method is to query medical record information through the unique index, Did, and return detailed data of the medical record.

   $AddMedRecord()$: Adding medical record information, including the patient's condition, medication, etc.

   $AddUitem()$: Adding modifications to the medical record, including the reason for the modification and the modified record entries.

3) PC: It takes charge of managing the entire system's access control policies. When users employ AC to initiate operations on devices, the system automatically calls PC to query and return the user's access control policies for AC authentication. Besides, when users perform AC-related functions, PC is invoked to update the access control policies if any changes occur in the access control policies. Similar to DC, PC is not directly accessible to users, but only provides interfaces for communication between users.

Algorithm 2 Part of AddU():Add Policy

Require: *Args*
Ensure: *True or Error*
1: $\langle AS, AO, AE, AP\rangle \leftarrow$ Args
2: $Policyld \leftarrow AES(AS.Userd + AO.Did + AS.Role)$
3: $ExistPolicy \leftarrow InvokeChaincode(Pc, QueryPolicy, Policyld)$
4: **if** $ExistPolicy! = False$ **then**
5: **return** $Error : Policy\ HaveAlreadyExisted$
6: **end if**
7: $\langle MaPolicy, err\rangle \leftarrow Marshal(AS, AO, AP, AE)$
8: **if** $err! = True$ **then**
9: **return** $Error$
10: **end if**
11: $resp \leftarrow APIstub.PutState(Policyld, MaPolicy)$
12: **if** $err! = True$ **then**
13: **return** $Error : Failed\ To\ Add\ Policy$
14: **end if**
15: **return** $true$

$AddPolicy()$: It aims to upload permission policy information. The policy is stored on the blockchain with the index being the hashed value of encrypted ASAO. When uploading a policy, the blockchain is first queried for the existence of the policy index. If it already exists, the policy cannot be added. Algorithm 2outlines the detailed process of $AddPolicy()$.
$QueryPolicy()$: Before performing any operation in the system, users must undergo permission verification. This function returns the user's access policy on the blockchain for AC to verify the user's access permissions. Based on the input policy index, it searches the blockchain and returns the corresponding search result.
$UpdatePolicy()$: This method does not support nor provide interfaces for external client invocation. When some of the users' operations involve changes to their access control policies, the smart contract automatically calls this method to update and maintain the access control policy ledger on the blockchain.

4.3. Execution Steps

As illustrated in Figure3, the specific workflow of this proposed solution mainly consists of four parts. The symbols used are shown in Table5.

Table 5. Used Symbols

Notations	Definition
CA	Certificate Authority
Cert	Certificate file
Key	Key in Hyperledger Fabric
Ledger	Ledger in Hyperledger Fabric
CC	Chaincode in Hyperledger Fabric
Channel	Passage in Hyperledger Fabric
f_x	Functions defined in source code
Ledger	Ledger in Hyperledger Fabric
SDB	State Database in Hyperledger Fabric
ABAC	Attribute based Access Control
ABAC	Feature-aware de-duplication
Cli	Blockchain system client

Part 1: This process involves building the blockchain network and deploying the chaincode. These fundamental steps are carried out by the backend institution. Part 1 is further divided into the following steps.
Step 1: CA tools are used to generate certificates and keys for identity authentication and encryption within the Fabric network.

$$CA = \{Cert, Key\} \tag{14}$$

Step 2: Using configtxgen and $PeerChannel$ tools to create channels and join peer nodes.

$$\{blockchain, ledger\} \xrightarrow{jion} Channel \tag{15}$$

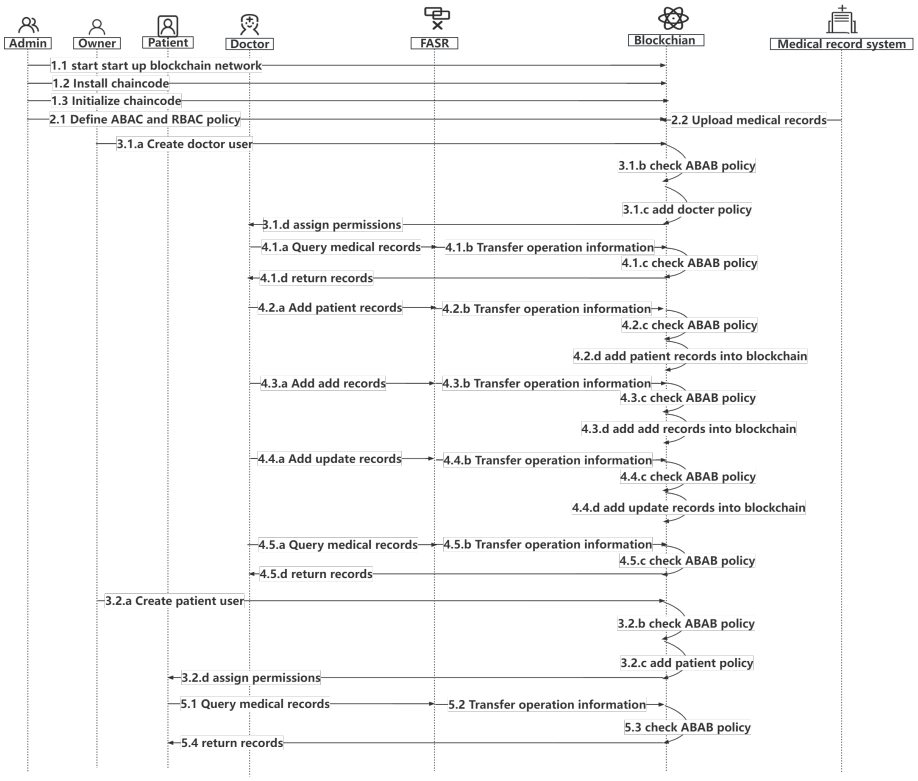


Fig. 3. System Architecture Diagram

Step 3: Deploy and execute smart contract using $Chaincode$ tool

$$ChainCode(f_x) \longrightarrow CC \tag{16}$$

Part 2: After creating the policies, the administrator needs to save them to the blockchain.
Step 1: The administrator and users set access control policies based on UA, MA, PA, and EA.

$$Set\,(UA, MA, PA, EA) \longrightarrow ABAC \tag{17}$$

Step 2: The administrator uploads the developed access control policies to the blockchain network.

$$deploy\,(ABAC) \longrightarrow Constract \tag{18}$$

Step 3:The administrator runs PC to execute operations such as adding and modifying policies, and saves the final policy values to SDB and the $ledge$r.

$$PC\,(ABAC) \longrightarrow \{SDB, Ledger\} \tag{19}$$

Part 3:In this part, users' access requests are first uploaded to $FASR$ to reduce data redundancy before storing the processed data in the blockchain.
Step 1: Users upload medical record information to $FASR$.

$$upload\,(MedicalRecords) \longrightarrow FASR \tag{20}$$

Step 2:$FASR$ deduplicates the medical record information and stores it in the blockchain.

$$FASR\,(MedicalRecords) \longrightarrow blockchain \tag{21}$$

Part 4: This part is the process of accessing medical information based on attribute-based access control and can be divided into four specific steps.
Step 1: The user initiates a request to access medical data.

$$Request \longrightarrow blockchain \tag{22}$$

Step 2: Upon receiving the user's request, AC policies are invoked to verify if the user can access the data.

$$AC\,(request) \longrightarrow \begin{cases} 1 & allow \\ 0 & deny \end{cases} \tag{23}$$

Step 3: The user's request is stored after deduplication through $FASR$.

$$blockchain\,(Request) \longrightarrow FASR \tag{24}$$

Step 4: The result of the user's request is returned.

$$Response\,(MedicialRecord) \longrightarrow Cli \tag{25}$$

5. Experimental Results and Analysis

This section mainly describes the entire experimental process. In Section5.1, the results of the simulation experiment are presented, and in Section5.2, we analyze the experimental data. To highlight the innovation of this solution, we conducted a comparison with relevant literature in Section5.3. The simulation experiment environment for this solution is presented in Table6.

Table 6. Experimental Environment

Experimental environment	
CPU	Intel core i5-7300
Memory	8 GB
Docker	v19.03.2
Docker-compose	v1.24.1
Node	v12.12.0
Hyperledger Fabric	v1.4.3
Golang	v1.13.5

5.1. Experimental Results

To validate the data security, decentralization of permission management, dynamic nature of permission updates, and the immutability of permission transfer records in the proposed system, three users are designed for the experiment, namely Owner, Doctor, and Patient, with IDs ”20001123”, ”19991123”, and ”20001028”, respectively. A medical record information P1 with a unique index Did of ”P100010001” is used. Firstly, the medical institution adds the policies to the blockchain network and the experimental results for policy upload are shown in Figure4. The main purpose of uploading policies is to configure the administrator's basic information on the blockchain network.

In addition to uploading administrator information, the medical institution's backend also needs to package and upload P1 to the blockchain system. In this experiment, a set of medical record information is simulated. Figure5 shows the process of of requesting the upload of medical record information and returning parameters.

```
root@yjzhu-ubuntu:/home/gopath/src/github.com/bian-fabric/network# ./cc.sh invoke pc 1.0 go/pc AddPolicy '"{\"AS\":{\"userId\":\"20001123\",\"role\":\"Owner\",\"group\":\"g1\"},\"AO\":{\"Did\":\"P100010001\",\"ownerId\":\"20001028\"},\"AP\":1}"'
invoke pc:1.0 github.com/newham/fabric-iot/chaincode/go/pc AddPolicy "{\"AS\":{\"userId\":\"20001123\",\"role\":\"Owner\",\"group\":\"g1\"},\"AO\":{\"Did\":\"P100010001\",\"ownerId\":\"20001028\"},\"AP\":1}"
Submitting transaction:AddPolicy to smart contract on iot-channel
ARGS:{"function":"AddPolicy","Args":["{\"AS\":{\"userId\":\"20001123\",\"role\":\"Owner\",\"group\":\"g1\"},\"AO\":{\"Did\":\"P100010001\",\"ownerId\":\"20001028\"},\"AP\":1}"]}
The transaction is sent to all of the peers so that chaincode is built before receiving the following requests
2023-04-18 01:53:51.142 UTC [chaincodeCmd] ClientWait -> INFO 001 txid [682bf83c7bd5636914a05b23d27272113fb587b37d162f84102
78fd0c346368d] committed with status (VALID) at peer0.org1.fabric-iot.edu:7051
2023-04-18 01:53:51.191 UTC [chaincodeCmd] ClientWait -> INFO 003 txid [682bf83c7bd5636914a05b23d27272113fb587b37d162f84102
78fd0c346368d] committed with status (VALID) at peer1.org2.fabric-iot.edu:10051
2023-04-18 01:53:51.191 UTC [chaincodeCmd] ClientWait -> INFO 002 txid [682bf83c7bd5636914a05b23d27272113fb587b37d162f84102
78fd0c346368d] committed with status (VALID) at peer1.org1.fabric-iot.edu:8051
2023-04-18 01:53:51.224 UTC [chaincodeCmd] ClientWait -> INFO 004 txid [682bf83c7bd5636914a05b23d27272113fb587b37d162f84102
78fd0c346368d] committed with status (VALID) at peer0.org2.fabric-iot.edu:9051
2023-04-18 01:53:51.224 UTC [chaincodeCmd] chaincodeInvokeOrQuery -> INFO 005 Chaincode invoke successful. result: status:200 payload:"ok"
done
```

Fig. 4. Upload Policy

In the system design presented in Section 4, we have provided a detailed overview of the access control permissions for each role. The medical institution is only required to upload blockchain policies and package medical record information. Subsequent simulated operations are carried out by the administrator, patient, and doctor. The Owner creates Doctor and Patient users by calling $CheckAccess()$ from AC, the system automatically verifies the access control permissions using $QueryPolicy()$ from PC. After successful verification, the $AddUser()$ is invoked, as shown in Figure6. In Figure7, the Owner creates the Patient role.

```
root@yjzhu-ubuntu:/home/gopath/src/github.com/bian-fabric/network# ./cc.sh invoke dc 1.0 go/dc AddDP '"{\"docTy
pe\":\"APL\",\"did\":\"P100010001\",\"name\":\"Bian\",\"age\":\"21\",\"ownerId\":\"20001028\"}"'
invoke dc:1.0 github.com/newham/fabric-iot/chaincode/go/dc AddDP "{\"docType\":\"APL\",\"did\":\"P100010001\",\
"name\":\"Bian\",\"age\":\"21\",\"ownerId\":\"20001028\"}"
Submitting transaction:AddDP to smart contract on iot-channel
ARGS:{"function":"AddDP","Args":["{\"docType\":\"APL\",\"did\":\"P100010001\",\"name\":\"Bian\",\"age\":\"21\",
\"ownerId\":\"20001028\"}"]}
The transaction is sent to all of the peers so that chaincode is built before receiving the following requests
2023-04-19 01:19:02.065 UTC [chaincodeCmd] ClientWait -> INFO 001 txid [f9d35e517f8e435a49d1ef967b077d8724fd6c1
daa9a875bd7ee060344ad26e1] committed with status (VALID) at peer0.org1.fabric-iot.edu:7051
2023-04-19 01:19:02.065 UTC [chaincodeCmd] ClientWait -> INFO 002 txid [f9d35e517f8e435a49d1ef967b077d8724fd6c1
daa9a875bd7ee060344ad26e1] committed with status (VALID) at peer1.org2.fabric-iot.edu:10051
2023-04-19 01:19:02.091 UTC [chaincodeCmd] ClientWait -> INFO 003 txid [f9d35e517f8e435a49d1ef967b077d8724fd6c1
daa9a875bd7ee060344ad26e1] committed with status (VALID) at peer1.org1.fabric-iot.edu:8051
2023-04-19 01:19:02.108 UTC [chaincodeCmd] ClientWait -> INFO 004 txid [f9d35e517f8e435a49d1ef967b077d8724fd6c1
daa9a875bd7ee060344ad26e1] committed with status (VALID) at peer0.org2.fabric-iot.edu:9051
2023-04-19 01:19:02.109 UTC [chaincodeCmd] chaincodeInvokeOrQuery -> INFO 005 Chaincode invoke successful. resu
lt: status:200 payload:"ok"
done
```

Fig. 5. Upload Medical Record Information

```
root@yjzhu-ubuntu:/home/gopath/src/github.com/bian-fabric/network# ./cc.sh invoke ac 1.0 go/ac CheckAccess '"{\"AS\":{\"use
rId\":\"20001123\",\"role\":\"Owner\"},\"AO\":{\"did\":\"P100010001\"}}","CreDoc","19991123"'
invoke ac:1.0 github.com/newham/fabric-iot/chaincode/go/ac CheckAccess "{\"AS\":{\"userId\":\"20001123\",\"role\":\"Owner\"
},\"AO\":{\"did\":\"P100010001\"}}","CreDoc","19991123"
Submitting transaction:CheckAccess to smart contract on iot-channel
ARGS:{"function":"CheckAccess","Args":["{\"AS\":{\"userId\":\"20001123\",\"role\":\"Owner\"},\"AO\":{\"did\":\"P100010001\"
}}","CreDoc","19991123"]}
The transaction is sent to all of the peers so that chaincode is built before receiving the following requests
2023-04-18 01:56:06.284 UTC [chaincodeCmd] ClientWait -> INFO 001 txid [6f6b8798d0154f1b64efe9a836a5fff87f9ce4fb0c3c3af8e96
635cf9577d5c8] committed with status (VALID) at peer1.org2.fabric-iot.edu:10051
2023-04-18 01:56:06.318 UTC [chaincodeCmd] ClientWait -> INFO 002 txid [6f6b8798d0154f1b64efe9a836a5fff87f9ce4fb0c3c3af8e96
635cf9577d5c8] committed with status (VALID) at peer0.org2.fabric-iot.edu:9051
2023-04-18 01:56:06.318 UTC [chaincodeCmd] ClientWait -> INFO 003 txid [6f6b8798d0154f1b64efe9a836a5fff87f9ce4fb0c3c3af8e96
635cf9577d5c8] committed with status (VALID) at peer0.org1.fabric-iot.edu:7051
2023-04-18 01:56:06.341 UTC [chaincodeCmd] ClientWait -> INFO 004 txid [6f6b8798d0154f1b64efe9a836a5fff87f9ce4fb0c3c3af8e96
635cf9577d5c8] committed with status (VALID) at peer1.org1.fabric-iot.edu:8051
2023-04-18 01:56:06.341 UTC [chaincodeCmd] chaincodeInvokeOrQuery -> INFO 005 Chaincode invoke successful. result: status:2
00 payload:"ok"
done
```

Fig. 6. Create Doctor User

```
root@yjzhu-ubuntu:/home/gopath/src/github.com/bian-fabric/network# ./cc.sh invoke ac 1.0 go/ac CheckAccess '"{\
"AS\":{\"userId\":\"20001123\",\"role\":\"Owner\"},\"AO\":{\"did\":\"P100010001\"}}","CrePat","20001028"'
invoke ac:1.0 github.com/newham/fabric-iot/chaincode/go/ac CheckAccess "{\"AS\":{\"userId\":\"20001123\",\"role
\":\"Owner\"},\"AO\":{\"did\":\"P100010001\"}}","CrePat","20001028"
Submitting transaction:CheckAccess to smart contract on iot-channel
ARGS:{"function":"CheckAccess","Args":["{\"AS\":{\"userId\":\"20001123\",\"role\":\"Owner\"},\"AO\":{\"did\":\"
P100010001\"}}","CrePat","20001028"]}
The transaction is sent to all of the peers so that chaincode is built before receiving the following requests
2023-04-19 01:20:57.744 UTC [chaincodeCmd] ClientWait -> INFO 001 txid [80d83fca366c8910d87162bead8043b26705320
a6016a27dd29c9844f0bb1104] committed with status (VALID) at peer0.org1.fabric-iot.edu:7051
2023-04-19 01:20:57.780 UTC [chaincodeCmd] ClientWait -> INFO 002 txid [80d83fca366c8910d87162bead8043b26705320
a6016a27dd29c9844f0bb1104] committed with status (VALID) at peer0.org2.fabric-iot.edu:9051
2023-04-19 01:20:57.780 UTC [chaincodeCmd] ClientWait -> INFO 003 txid [80d83fca366c8910d87162bead8043b26705320
a6016a27dd29c9844f0bb1104] committed with status (VALID) at peer1.org1.fabric-iot.edu:8051
2023-04-19 01:20:57.800 UTC [chaincodeCmd] ClientWait -> INFO 004 txid [80d83fca366c8910d87162bead8043b26705320
a6016a27dd29c9844f0bb1104] committed with status (VALID) at peer1.org2.fabric-iot.edu:10051
2023-04-19 01:20:57.800 UTC [chaincodeCmd] chaincodeInvokeOrQuery -> INFO 005 Chaincode invoke successful. resu
lt: status:200 payload:"ok"
done
```

Fig. 7. Create Patient User

```
root@yjzhu-ubuntu:/home/gopath/src/github.com/bian-fabric/network# ./cc.sh invoke ac 1.0 go/ac CheckAccess '"{\"AS\":{\"use
rId\":\"19991123\",\"role\":\"Docter\"},\"AO\":{\"did\":\"P100010001\"}}","AddTit","I100010001","cold","Penicillin"'
invoke ac:1.0 github.com/newham/fabric-iot/chaincode/go/ac CheckAccess "{\"AS\":{\"userId\":\"19991123\",\"role\":\"Docter\
"},\"AO\":{\"did\":\"P100010001\"}}","AddTit","I100010001","cold","Penicillin"
Submitting transaction:CheckAccess to smart contract on iot-channel
ARGS:{"function":"CheckAccess","Args":["{\"AS\":{\"userId\":\"19991123\",\"role\":\"Docter\"},\"AO\":{\"did\":\"P100010001\
"}}","AddTit","I100010001","cold","Penicillin"]}
The transactin is sent to all of the peers so that chaincode is built before receiving the following requests
2023-04-19 01:25:28.004 UTC [chaincodeCmd] ClientWait -> INFO 001 txid [5110dda568447562a9cc88b1d75a90ed11ff90e744df2570e85
9faa5966ca605] committed with status (VALID) at peer0.org1.fabric-iot.edu:7051
2023-04-19 01:25:28.040 UTC [chaincodeCmd] ClientWait -> INFO 002 txid [5110dda568447562a9cc88b1d75a90ed11ff90e744df2570e85
9faa5966ca605] committed with status (VALID) at peer0.org2.fabric-iot.edu:9051
2023-04-19 01:25:28.040 UTC [chaincodeCmd] ClientWait -> INFO 003 txid [5110dda568447562a9cc88b1d75a90ed11ff90e744df2570e85
9faa5966ca605] committed with status (VALID) at peer1.org1.fabric-iot.edu:8051
2023-04-19 01:25:28.061 UTC [chaincodeCmd] ClientWait -> INFO 004 txid [5110dda568447562a9cc88b1d75a90ed11ff90e744df2570e85
9faa5966ca605] committed with status (VALID) at peer1.org2.fabric-iot.edu:10051
2023-04-19 01:25:28.061 UTC [chaincodeCmd] chaincodeInvokeOrQuery -> INFO 005 Chaincode invoke successful. result: status:2
00 payload:"ok"
done
```

Fig. 8. Add Medical Record Information

```
root@yjzhu-ubuntu:/home/gopath/src/github.com/bian-fabric/network# ./cc.sh invoke ac 1.0 go/ac CheckAccess '"{\"AS\":{\"use
rId\":\"19991123\",\"role\":\"Docter\"},\"AO\":{\"did\":\"P100010001\"}}","AddUit","I100010001","operate"'
invoke ac:1.0 github.com/newham/fabric-iot/chaincode/go/ac CheckAccess "{\"AS\":{\"userId\":\"19991123\",\"role\":\"Docter\
"},\"AO\":{\"did\":\"P100010001\"}}","AddUit","I100010001","operate"
Submitting transaction:CheckAccess to smart contract on iot-channel
ARGS:{"function":"CheckAccess","Args":["{\"AS\":{\"userId\":\"19991123\",\"role\":\"Docter\"},\"AO\":{\"did\":\"P100010001\
"}}","AddUit","I100010001","operate"]}
The transaction is sent to all of the peers so that chaincode is built before receiving the following requests
Error: endorsement failure during invoke. response: status:500 message:"no illnessid"
root@yjzhu-ubuntu:/home/gopath/src/github.com/bian-fabric/network# ./cc.sh invoke dc 1.0 go/dc AddAitem '"P100010001","Chen
","AnHui"'
invoke dc:1.0 github.com/newham/fabric-iot/chaincode/go/dc AddAitem "P100010001","Chen","AnHui"
Submitting transaction:AddAitem to smart contract on iot-channel
ARGS:{"function":"AddAitem","Args":["P100010001","Chen","AnHui"]}
The transaction is sent to all of the peers so that chaincode is built before receiving the following requests
2023-04-19 01:26:37.777 UTC [chaincodeCmd] ClientWait -> INFO 001 txid [c448be19ba1120d4363c7f6d602a11e87580d8c3063316c67da
3d051c3b60a4b] committed with status (VALID) at peer1.org2.fabric-iot.edu:10051
2023-04-19 01:26:37.842 UTC [chaincodeCmd] ClientWait -> INFO 002 txid [c448be19ba1120d4363c7f6d602a11e87580d8c3063316c67da
3d051c3b60a4b] committed with status (VALID) at peer0.org1.fabric-iot.edu:7051
2023-04-19 01:26:37.873 UTC [chaincodeCmd] ClientWait -> INFO 003 txid [c448be19ba1120d4363c7f6d602a11e87580d8c3063316c67da
3d051c3b60a4b] committed with status (VALID) at peer0.org2.fabric-iot.edu:9051
2023-04-19 01:26:37.887 UTC [chaincodeCmd] ClientWait -> INFO 004 txid [c448be19ba1120d4363c7f6d602a11e87580d8c3063316c67da
3d051c3b60a4b] committed with status (VALID) at peer1.org1.fabric-iot.edu:8051
2023-04-19 01:26:37.887 UTC [chaincodeCmd] chaincodeInvokeOrQuery -> INFO 005 Chaincode invoke successful. result: status:2
00 payload:"ok"
done
```

Fig. 9. Add Medical Record Modification Information

```
root@yjzhu-ubuntu:/home/gopath/src/github.com/bian-fabric/network# ./cc.sh invoke ac 1.0 go/ac CheckAccess '"{\"AS\":{
\"userId\":\"19991123\",\"role\":\"Docter\"},\"AO\":{\"did\":\"P100010001\"}}","FindByDid"'
invoke ac:1.0 github.com/newham/fabric-iot/chaincode/go/ac CheckAccess "{\"AS\":{\"userId\":\"19991123\",\"role\":\"Do
cter\"},\"AO\":{\"did\":\"P100010001\"}}","FindByDid"
Submitting transaction:CheckAccess to smart contract on iot-channel
ARGS:{"function":"CheckAccess","Args":["{\"AS\":{\"userId\":\"19991123\",\"role\":\"Docter\"},\"AO\":{\"did\":\"P10001
0001\"}}","FindByDid"]}
The transaction is sent to all of the peers so that chaincode is built before receiving the following requests
2023-04-19 01:27:21.121 UTC [chaincodeCmd] ClientWait -> INFO 001 txid [791b149af36d41dfbe3dc36226228e602e0b6f45916d17
12d5e0ca1c60770a94] committed with status (VALID) at peer0.org1.fabric-iot.edu:7051
2023-04-19 01:27:21.163 UTC [chaincodeCmd] ClientWait -> INFO 002 txid [791b149af36d41dfbe3dc36226228e602e0b6f45916d17
12d5e0ca1c60770a94] committed with status (VALID) at peer1.org2.fabric-iot.edu:10051
2023-04-19 01:27:21.180 UTC [chaincodeCmd] ClientWait -> INFO 004 txid [791b149af36d41dfbe3dc36226228e602e0b6f45916d17
12d5e0ca1c60770a94] committed with status (VALID) at peer1.org1.fabric-iot.edu:8051
2023-04-19 01:27:21.180 UTC [chaincodeCmd] ClientWait -> INFO 003 txid [791b149af36d41dfbe3dc36226228e602e0b6f45916d17
12d5e0ca1c60770a94] committed with status (VALID) at peer0.org2.fabric-iot.edu:9051
2023-04-19 01:27:21.181 UTC [chaincodeCmd] chaincodeInvokeOrQuery -> INFO 005 Chaincode invoke successful. result: sta
tus:200 payload:"Disease is :{\"Illness\":\"cold\",\"IllnessId\":\"I100010001\",\"Medicine\":\"Penicillin\",\"NowTime\
":1681867525}\nHisAdd is :{\"DocterName\":\"Chen\",\"Local\":\"AnHui\"}\nHisUpdate is :{\"IllnessId\":\"I100010001\",\
"Result\":\"operate\"}"
done
```

Fig. 10. Query Medical Record Information

The Doctor adds medical record information P1. Here, the Doctor user uses Check-Access(). However, when the system uses $CheckAccess()$, it also automatically calls $QueryPolicy()$ from PC and $AddMedRecords()$ from DC to perform access control authorization verification and medical record information addition, respectively. As shown in Figure8, the Doctor user successfully adds the medical record information.

For adding modification and additional information to the medical record, as shown in Figure9, the Doctor user adds modification information. The process of adding additional information is similar and also utilizes $CheckAccess()$. Similarly, the blockchain system will call the $QueryPolicy()$ method from PC and the $FindByDid()$ and $AddUitem()$ methods from DC. After adding the information, the Doctor can query the complete medical record information, as shown in Figure10.

5.2. Performance Testing

To validate the actual performance of the proposed system, three sets of simulated experiments are conducted. In the first and second sets of experiments, the number of virtual concurrent clients is set to 100, 200, 400, 600, 800, and 1000, and the time consumption and throughput of PC and AC are measured, as shown in Figs.11 and 12, respectively. In the second set of experiments, the number of virtual concurrent clients is set to 100, 200, 300, 400, 500, 600, 800, and 1000, and the time consumption and throughput of DC are measured, as shown in Fig.13.

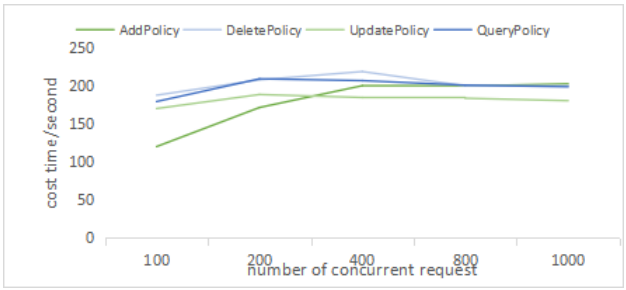


Fig. 11. Throughput of PC under different concurrent requests.

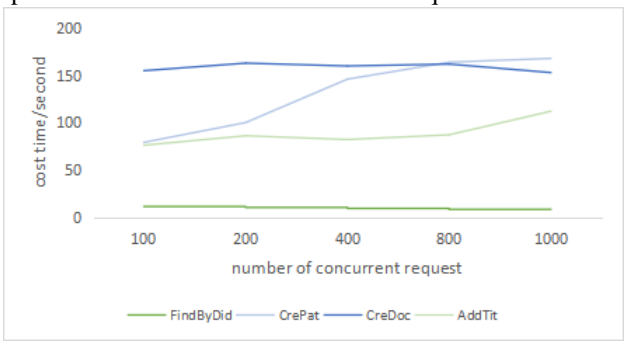


Fig. 12. Throughput of AC under different concurrent requests.

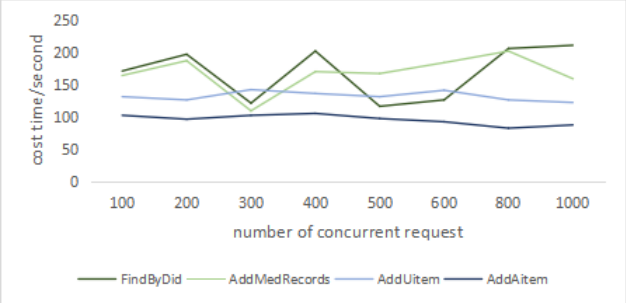


Fig. 13. Throughput of DC under different concurrent requests.

Based on the above experimental results, the following conclusions can be drawn. The throughput of adding and updating operations is lower than that of querying and deleting operations. When the number of concurrent requests reaches a certain value, the throughput does not significantly decrease. This is because when the number of connections in the blockchain network's connection pool reaches its limit, the system's throughput tends to stabilize.

We compare the throughput of the AC method by setting different numbers of nodes, ranging from 100 to 1000. We observe that the throughput of $FindByDid()$is significantly lower than the other three methods.

The throughput of operations in AC is generally lower than that of operations in DC and PC. In fact, operations in AC often require calling operations in DC and PC, making the complexity of operations in AC higher than in DC and PC.

5.3. Comparison with Existing Works

To highlight the innovation of our proposed system, we compare it with other relevant systems, as shown in Table7. In terms of data security, protecting privacy information requires enhanced data security measures. But Literature[46] overlooks data security issues. Literatures[47]and[48] utilize consortium blockchain to store private data, which indeed improves data security to some extent. However, solely relying on blockchain features is not sufficient to ensure data security. Storing data in the blockchain does not necessarily achieve dynamic and fine-grained access to medical records. Compared to the method in[47], the work in[48] proposes a blockchain-based medical information sharing platform, enabling medical information to be stored, shared, and trusted across distributed networks. Regarding decentralized access control, accessing medical records is often closely related to user roles in practical applications. Literatures[46],[47],and [48]merely store medical records in the blockchain without distinguishing fine-grained access to medical information. To overcome the limitations mentioned above, we combine the strengths of ABAC and RBAC models and utilize the features of AES algorithm to achieve fine-grained and dynamic permission management,and use FASR deduplication technology to reduce data redundancy. And the access control model is implemented on the basis of smart contracts, ensuring real-time monitoring of on-chain data and enhancing the overall system's security.

Table 7. Comparison with Existing Solutions

Paper	[46]	[47]	[48]	Ours
Access control	ABAC	No	ABAC	ABACandRBAC
Security	No	No	No	AES
Smart contract	No	Yes	Yes	Yes
Blockchain platform	No	Hyperledger Fabric	Hyperledger Fabric	Hyperledger Fabric
Storage strategy	No	No	No	FASR

6. Conclusions

This paper combines blockchain technology with ABAC and RBAC models to effectively break the information silos in medical data, and ensure the security of medical records while enabling data sharing. The FASR technology is used to handle medical records, reducing data redundancy and alleviating blockchain storage pressure. Moreover, the proposed solution utilizes a distributed architecture to achieve dynamic and fine-grained role division, integrating blockchain with real-world applications for efficient medical record management. The deployment and invocation process of chaincode are described in detail, and the feasibility of the system is demonstrated through experiments.

This model is expected to be applied to practical medical scenarios, and with the rise of the concept of meta-universe, traditional medical treatment should also keep pace with The Times. Led by relevant authorities, the medical record information is uploaded to the blockchain network, and everyone can access the record through an account with specific permissions. Of course, there are still many problems to be solved.

Future work can focus on the following areas for further improvements. Our experiments can be performed on clusters instead of a single PC to optimize the performance of the distributed system. Addressing data storage issues at the fundamental level in blockchain rather than relying solely on FASR technology to mitigate storage pressure. Besides, additional features should be introduced into the framework, such as supporting external institutions to access medical records securely and implementing encrypted retrieval.

Acknowledgments. This work is supported in part by the Top-notch Innovative Talent Training Program for Graduate students of Shanghai Maritime University under Grant 2021YBR008.

References

1. Edward H Shortliffe. The evolution of electronic medical records. *Academic Medicine*, 74(4):414–9, 1999.
2. AL Rector, WA Nowlan, and Shazia Kay. Foundations for an electronic medical record. *Methods of information in medicine*, 30(03):179–186, 1991.
3. Jiahong Cai, Wei Liang, Xiong Li, Kuanching Li, Zhenwen Gui, and Muhammad Khurram Khan. Gtxchain: A secure iot smart blockchain architecture based on graph neural network. *IEEE Internet of Things Journal*, 2023.

4. Dezhi Han, Yujie Zhu, Dun Li, Wei Liang, Alireza Souri, and Kuan-Ching Li. A blockchain-based auditable access control system for private data in service-centric iot environments. *IEEE Transactions on Industrial Informatics*, 18(5):3530–3540, 2021.
5. Elias Drakopoulos and Matt Merges. Performance study of client-server storage systems. In *[1991] Digest of Papers Eleventh IEEE Symposium on Mass Storage Systems*, pages 67–68. IEEE Computer Society, 1991.
6. Shivansh Kumar, Aman Kumar Bharti, and Ruhul Amin. Decentralized secure storage of medical records using blockchain and ipfs: A comparative analysis with future directions. *Security and Privacy*, 4(5):e162, 2021.
7. Wei Liang, Yang Yang, Ce Yang, Yonghua Hu, Songyou Xie, Kuan-Ching Li, and Jiannong Cao. Pdpchain: A consortium blockchain-based privacy protection scheme for personal data. *IEEE Transactions on Reliability*, 2022.
8. Nazanin Zahed Benisi, Mehdi Aminian, and Bahman Javadi. Blockchain-based decentralized storage networks: A survey. *Journal of Network and Computer Applications*, 162:102656, 2020.
9. Hongzhi Li, Dezhi Han, and Mingdong Tang. A privacy-preserving storage scheme for logistics data with assistance of blockchain. *IEEE Internet of Things Journal*, 9(6):4704–4720, 2021.
10. Jiatao Li, Dezhi Han, Zhongdai Wu, Junxiang Wang, Kuan-Ching Li, and Arcangelo Castiglione. A novel system for medical equipment supply chain traceability based on alliance chain and attribute and role access control. *Future Generation Computer Systems*, 142:195–211, 2023.
11. Ed Coyne and Timothy R Weil. Abac and rbac: scalable, flexible, and auditable access management. *IT professional*, 15(03):14–16, 2013.
12. Elli Androulaki, Artem Barger, Vita Bortnikov, Christian Cachin, Konstantinos Christidis, Angelo De Caro, David Enyeart, Christopher Ferris, Gennady Laventman, Yacov Manevich, et al. Hyperledger fabric: a distributed operating system for permissioned blockchains. In *Proceedings of the thirteenth EuroSys conference*, pages 1–15, 2018.
13. Chongqing Chen, Dezhi Han, and Xiang Shen. Clvin: Complete language-vision interaction network for visual question answering. *Knowledge-Based Systems*, page 110706, 2023.
14. Chongqing Chen, Dezhi Han, and Chin-Chen Chang. Caan: Context-aware attention network for visual question answering. *Pattern Recognition*, 132:108980, 2022.
15. Dun Li, Dezhi Han, Zibin Zheng, Tien-Hsiung Weng, Hongzhi Li, Han Liu, Arcangelo Castiglione, and Kuan-Ching Li. Moocschain: A blockchain-based secure storage and sharing scheme for moocs learning. *Computer Standards & Interfaces*, 81:103597, 2022.
16. Dun Li, Dezhi Han, Tien-Hsiung Weng, Zibin Zheng, Hongzhi Li, Han Liu, Arcangelo Castiglione, and Kuan-Ching Li. Blockchain for federated learning toward secure distributed machine learning systems: a systemic survey. *Soft Computing*, 26(9):4423–4440, 2022.
17. Baowei Wang, Shi Jiawei, Weishen Wang, and Peng Zhao. Image copyright protection based on blockchain and zero-watermark. *IEEE Transactions on Network Science and Engineering*, 9(4):2188–2199, 2022.
18. Pengbin Han, Aina Sui, Tao Jiang, and Chaonan Gu. Copyright certificate storage and trading system based on blockchain. In *2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, pages 611–615. IEEE, 2020.
19. Nan Jing, Qi Liu, and Vijayan Sugumaran. A blockchain-based code copyright management system. *Information Processing & Management*, 58(3):102518, 2021.
20. Chiara Garilli. Blockchain and smart contracts: New perspectives on copyright protection in the digital single market. In *Handbook of Research on Applying Emerging Technologies Across Multiple Disciplines*, pages 159–175. IGI Global, 2022.
21. Lijun Xiao, Weihong Huang, Yong Xie, Weidong Xiao, and Kuan-Ching Li. A blockchain-based traceable ip copyright protection algorithm. *IEEE Access*, 8:49532–49542, 2020.

22. Na Gao, Dezhi Han, Tien-Hsiung Weng, Benhui Xia, Dun Li, Arcangelo Castiglione, and Kuan-Ching Li. Modeling and analysis of port supply chain system based on fabric blockchain. *Computers & Industrial Engineering*, 172:108527, 2022.
23. Bing Qing Tan, Fangfang Wang, Jia Liu, Kai Kang, and Federica Costa. A blockchain-based framework for green logistics in supply chains. *Sustainability*, 12(11):4656, 2020.
24. Ahmad Musamih, Khaled Salah, Raja Jayaraman, Junaid Arshad, Mazin Debe, Yousof Al-Hammadi, and Samer Ellahham. A blockchain-based approach for drug traceability in healthcare supply chain. *IEEE access*, 9:9728–9743, 2021.
25. Sara Saberi, Mahtab Kouhizadeh, Joseph Sarkis, and Lejia Shen. Blockchain technology and its relationships to sustainable supply chain management. *International journal of production research*, 57(7):2117–2135, 2019.
26. Guang Chen, Bing Xu, Manli Lu, and Nian-Shing Chen. Exploring blockchain technology and its potential applications for education. *Smart Learning Environments*, 5(1):1–10, 2018.
27. Preeti Bhaskar, Chandan Kumar Tiwari, and Amit Joshi. Blockchain in education management: present and future applications. *Interactive Technology and Smart Education*, 18(1):1–17, 2021.
28. Kim Beom Rii. Digital ilearning chain scheme in education blockchain based. *Aptisi Transactions on Technopreneurship (ATT)*, 4(2):174–183, 2022.
29. Hongzhi Li and Dezhi Han. Edurss: A blockchain-based educational records secure storage and sharing scheme. *IEEE access*, 7:179273–179289, 2019.
30. Yi Chen, Shuai Ding, Zheng Xu, Handong Zheng, and Shanlin Yang. Blockchain-based medical records secure storage and medical service framework. *Journal of medical systems*, 43:1–9, 2019.
31. Igor Radanović and Robert Likić. Opportunities for use of blockchain technology in medicine. *Applied health economics and health policy*, 16:583–590, 2018.
32. Jia Qu. Blockchain in medical informatics. *Journal of Industrial Information Integration*, 25:100258, 2022.
33. Hu Liu and Yuxuan Liu. Construction of a medical resource sharing mechanism based on blockchain technology: evidence from the medical resource imbalance of china. In *Healthcare*, volume 9, page 52. MDPI, 2021.
34. Mhamad Bakro, Sukant K Bisoy, Ashok K Patel, and M Adib Naal. Hybrid blockchain-enabled security in cloud storage infrastructure using ecc and aes algorithms. In *Blockchain based Internet of Things*, pages 139–170. Springer, 2022.
35. Yassine El Khanboubi, Mostafa Hanoune, and Mohamed El Ghazouani. A new data deletion scheme for a blockchain-based de-duplication system in the cloud. *Int. J. Commun. Netw. Inf. Secur. IJCNIS*, 13:331–339, 2021.
36. Sisi Zhou, Kuanching Li, Lijun Xiao, Jiahong Cai, Wei Liang, and Arcangelo Castiglione. A systematic review of consensus mechanisms in blockchain. *Mathematics*, 11(10):2248, 2023.
37. Chris Dannen. *Introducing Ethereum and solidity*, volume 1. Springer, 2017.
38. Dejan Vujičić, Dijana Jagodić, and Siniša Ranić. Blockchain technology, bitcoin, and ethereum: A brief overview. In *2018 17th international symposium infoteh-jahorina (infoteh)*, pages 1–6. IEEE, 2018.
39. Eric Yuan and Jin Tong. Attributed based access control (abac) for web services. In *IEEE International Conference on Web Services (ICWS'05)*. IEEE, 2005.
40. David Ferraiolo, Janet Cugini, D Richard Kuhn, et al. Role-based access control (rbac): Features and motivations. In *Proceedings of 11th annual computer security application conference*, pages 241–48, 1995.
41. Alex Biryukov and Dmitry Khovratovich. Related-key cryptanalysis of the full aes-192 and aes-256. In *Advances in Cryptology–ASIACRYPT 2009: 15th International Conference on the Theory and Application of Cryptology and Information Security, Tokyo, Japan, December 6-10, 2009. Proceedings 15*, pages 1–18. Springer, 2009.

42. Shiwen Zhang, Biao Hu, Wei Liang, Kuan-Ching Li, and Brij B Gupta. A caching-based dual k-anonymous location privacy-preserving scheme for edge computing. *IEEE Internet of Things Journal*, 2023.
43. Han Liu, Dezhi Han, Mingming Cui, Kuan-Ching Li, Alireza Souri, and Mohammad Shojafar. Idenmultisig: identity-based decentralized multi-signature in internet of things. *IEEE Transactions on Computational Social Systems*, 2023.
44. Dezhi Han, Nannan Pan, and Kuan-Ching Li. A traceable and revocable ciphertext-policy attribute-based encryption scheme based on privacy protection. *IEEE Transactions on Dependable and Secure Computing*, 19(1):316–327, 2020.
45. Wenbin Yao, Mengyao Hao, Yingying Hou, and Xiaoyong Li. Fasr: An efficient feature-aware deduplication method in distributed storage systems. *IEEE Access*, 10:15311–15321, 2022.
46. Marcela T de Oliveira, Yiannis Verginadis, Lúcio HA Reis, Evgenia Psarra, Ioannis Patiniotakis, and Sílvia D Olabarriaga. Ac-abac: Attribute-based access control for electronic medical records during acute care. *Expert Systems with Applications*, 213:119271, 2023.
47. Mingxiao Du, Qijun Chen, Jieying Chen, and Xiaofeng Ma. An optimized consortium blockchain for medical information sharing. *IEEE Transactions on Engineering Management*, 68(6):1677–1689, 2020.
48. Zhijie Sun, Dezhi Han, Dun Li, Xiangsheng Wang, Chin-Chen Chang, and Zhongdai Wu. A blockchain-based secure storage scheme for medical information. *EURASIP Journal on Wireless Communications and Networking*, 2022(1):40, 2022.

Aoao Bian received the B.S. degree in Computer science and Technology from the Anqing Normal University , Anqing, China, in 2022,and he is currently pursuing the M.S. degree in Shanghai Maritime University, Shanghai, China, in 2024..Her main research interests include blockchain technology and its applications and cryptography.

Dezhi Han received the B.S. degree from Hefei University of Technology, Hefei, China, the MS degree and Ph.D. degree from Huazhong University of Science and Technology, Wuhan, China. He is currently a professor of computer science and engineering at Shanghai Maritime University. His specific interests include storagearchitecture, blockchain technology, cloud computing security and cloud storage security technology.

Mingming Cui received a B.S. degree in Computer Science and Technology from the Anhui University of Finance and Economics, China. She is currently pursuing a Ph.D. degree from Shanghai Maritime University, China, and a Visiting Ph.D. student at the Nanyang Technological University, Singapore. Her research interests include cryptology, blockchain, data privacy protection, network security, VANETS security, and the Internet of things.

Dun Li received the B.S. degree in Human Resource Management from the Huaqiao University, Quanzhou, China, in 2013, and the M.S. degree in Finance from the Macau University of Science and Technology, Macau, China, in 2015. He is currently doing his Ph.D. degree in Information Management and Information Systems at Shanghai Maritime University. His main research interests include smart finance, big data, machine learning, IoT, and Blockchain.

Received: October 02, 2023; Accepted: January 25, 2024.

Computer Science and Information Systems 21(3):831–850 https://doi.org/10.2298/CSIS231116020X

CTA-Net: A Gaze Estimation network based on Dual Feature Aggregation and Attention Cross Fusion

Chenxing Xia[1,2,3], Zhanpeng Tao[1,⋆], Wei Wang[4], Wenjun Zhao[1], Bin Ge[1], Xiuju Gao[4], Kuan-Ching Li[6], and Yan Zhang[7]

[1] College of Computer Science and Engineering, Anhui University of Science and Technology
232001 Huainan, China
847990008@qq.com
[2] Institute of Energy, Hefei Comprehensive National Science Center
230031 Hefei, China
cxxia@aust.edu.cn
[3] Anhui Purvar Bigdata Technology Co. Ltd
232001 Huainan, China
[4] Anyang Cigarette Factory, China Tobacco Henan Industrial Co
Anyang, CHINA
[5] College of Electrical and Information Engineering, Anhui University of Science and Technology
232001 Huainan, China
[6] Department of Computer Science and Information Engineering, Providence University
43301 Taichung City, Taiwan
[7] The School of Electronics and Information Engineering, Anhui University
Hefei, China

Abstract. Recent work has demonstrated the Transformer model is effective for computer vision tasks. However, the global self-attention mechanism utilized in Transformer models does not adequately consider the local structure and details of images, which may result in the loss of information and local details, causing decreased estimation accuracy in gaze estimation tasks when compared to convolution or sequential stacking methods. To address this issue, we propose a parallel CNNs-Transformer aggregation network (CTA-Net) for gaze estimation, which fully leverages the advantages of the Transformer model in modeling global context while the convolutional neural networks (CNNs) model in retaining local details. Specifically, Transformer and ResNet are deployed to extract facial and eye information, respectively. Additionally, an attention cross fusion (ACFusion) Block is embedded with CNN branch, which decomposes features in space and channels to supplement lost features, suppress noise, and help extract eye features more effectively. Finally, a dual-feature aggregation (DFA) module is proposed to effectively fuse the output features of both branches with the help feature a selection mechanism and a residual structure. Experimental results on the MPIIGaze and Gaze360 datasets demonstrate that our CTA-Net achieves state-of-the-art results.

Keywords: Appearance-based gaze estimation, Deep neural networks, Dilated convolution, Fusion, Transformer

⋆ Corresponding author

1. Introduction

Estimating gaze from a single low-cost RGB sensor is an important research topic in computer vision, where eye or facial images are typically used as inputs to estimate the real gaze direction and locate gaze points. Gaze estimation has important applications in fields such as human-computer interaction [32], education [10], and medical diagnosis [36] [30].

Existing gaze estimation methods can be roughly categorized into two categories: model-based methods [9] [23] [25] and appearance-based methods [13] [26] [22]. Model-based methods focus on learning the geometric model of the entire eye and perform gaze estimation through manual calibration of features. However, this approach heavily relies on complex experimental equipment, which limits its effectiveness in harsh and unconstrained environments. In recent years, appearance-based methods have attracted much attention due to the development and application of deep learning, which only requires a regular RGB camera to capture images and directly learns the mapping function from facial appearance to human gaze [1]. CNNs have the ability to learn highly complex mapping functions, making them suitable for gaze regression. For example, Dilated-Net [4] , RT-Gene-Net [12], CA-Net [5]. Although these CNNs-based gaze estimation methods have achieved good performance, they always rely on using dilation convolution operations with different padding rates to extract contextual information, which may cause the loss of local information related to gaze and make the contextual information unrelated. Therefore, developing more effective gaze estimation models is critical for achieving better performance, robustness, and generalization in gaze estimation tasks.

Along with the Transformer model [28], thanks to its self-attention mechanism that selectively captures long-term dependencies between all tokens, has demonstrated outstanding performance in natural language processing tasks. In recent years, researchers have explored the use of Transformers in visual tasks, including gaze estimation. For instance, Cheng *et al.* [6] designed a Hybrid Transformer, which combines CNN and Transformer to extract low-level features and model global interactions, respectively. Cai *et al.* [2] used a linear combination of different Transformer frameworks for prediction and achieved a high ranking in the ETH-XGaze competition leaderboard. These methods effectively address the limitations of the Transformer architecture in modeling fine-grained details by using convolutional networks to introduce spatial biases in modeling local information. However, given that the operation of flattening image patches in the Transformer architecture may negatively impact the internal structural information of low-resolution images . Additionally, Transformer-based approaches often suffer from a quadratic growth of computational complexity with spatial size, which can result in cumbersome network architectures. Therefore, there is a need for a new network framework that can effectively address these issues while maintaining the integrity of structural information in low-resolution images.

In this paper, a parallel CNNs-Transformer aggregation network is proposed for gaze estimation (CTA-Net), which explicitly embeds global context and local information. Different from most existing Gaze estimation methods using a single feature encoder to extract feature, our CTA-Net adopts a parallel CNNs-Transformer structure to extract local and global cues via CNNs and Transformer networks from facial and eyes images, respectively. Moreover, since existing methods either ignore the correlation between the two eyes or handle the eye images separately for final output, we propose an eye image

Attention Cross Fusion (ACFusion) Block by using different combinations and interaction mechanisms. Specifically, we generate multiple attention feature maps through an attention mechanism to effectively filter out redundant or noisy information in feature channels, and further enhance the internal information of features and improve the performance of gaze estimation and target detection by interacting with binocular attention features. Finally, a dual-feature aggregation (DFA) module is introduced to merge the output features of different encoders for joint gaze regression.

Our contributions are as follows:

• We propose a parallel CNNs-Transformer aggregation network for gaze estimation (CTA-Net), which simultaneously considers the local details of features and global high-level semantic information.

• We design a dual-feature aggregation (DFA) module to fully integrate feature information from different encoders, promoting global information from the Transformer branch and reducing the noise that CNN features may contain.

• We present an eye image Attention Cross Fusion Block(ACFusion), which generates multiple attention feature maps to aggregate attention feature information for binocular image interaction.

2. RELATEDWORK

In the realm of gaze estimation, our exploration of related work focuses on two pivotal aspects: Gaze Estimation and Transformer-based approaches. These choices are motivated by the historical evolution of gaze estimation methodologies and the recent transformative impact of Transformer architectures in computer vision.

2.1. Gaze Estimation

Before the advent of CNN-based approaches to gaze estimation, regression functions were commonly used to create specific gaze mapping functions. These methods, such as neural networks [31], adaptive linear regression [21], Gaussian process regression [33], and dimension reduction [20], showed reasonable accuracy in constrained settings, but they were significantly less accurate in unconstrained settings due to the highly non-linear nature of the mapping function. However, with the rapid development of deep learning in recent years, Zhang *et al.* [38] proposed the first CNN-based gaze estimation method that far exceeded the performance of function regression methods using only a simple CNN. Since then, many improved and extended CNN-based gaze estimation methods [7] have emerged. For instance, Yu *et al.* [19] designed a multi-task gaze estimation model with landmark constraints, and Fischer *et al.* [12] used VGG-16 to extract features from binocular images to estimate gaze. Zhang *et al.* [38] proposed GazeNet, which inputed eye images into a 16-layer architecture where the head pose information was connected to the first fully connected layer after the convolutional layer. Cheng et al. [4] introduced dilated convolution in their gaze estimation method, which was combined with joint inference for head and eye avatars. Moreover, Cheng *et al.* [8] developed the FAR-Net to estimate the 3D gaze points of both eyes by combining the asymmetric properties of both eyes, which has the best performance in several public data sets. Krafka *et al.* [18] presented a multi-channel architecture to takes as input a left-eye image, a right-eye image, and a

cropped image of the face and face grid information. Cheng *et al.* [5] later proposed a coarse-to-fine network to integrate face and eye images by estimating a basic gaze from the face image and then refining the basic gaze with the eye image. To better investigate CNN-based gaze estimation, many large-scale gaze dataset tasks have been mentioned [16] [40] [37]. Despite the great success of these methods in gaze estimation, more accurate models are always needed and gaze estimation still faces challenges.

2.2. Transformer

The Transformer was initially introduced by Vaswani *et al.* [28] in the realm of natural language processing (NLP), as a unique self-attention mechanism to replace convolutional and recurrent networks. Compared with recurrent networks, the global computation of self-attention layers can selectively capture long-term dependencies between all tokens, and efficiently capturing continuity between semantic information and exhibiting outstanding performance. The success of the Transformer in NLP has inspired research in the field of computer vision, and it has been applied to various visual tasks. For instance, Carion *et al.* [3] used the encoder-decoder structure of the Transformer as the detection head (DETR) by extracting image features through a CNN and inputting the features into the detection head for prediction. Dosovitskiy *et al.* [11] proposed a purely self-attention-based visual Transformer (ViT), which solved the image classification problem by directly predicting possible categories using the Transformer encoder after dividing the image into non-overlapping 16×16 patches. Zheng *et al.* [41] applied the Transformer to replace the encoder in the natural image segmentation task to achieve state-of-the-art results. Cheng *et al.* [6] first introduced the Transformer into the gaze estimation task and proposed a Hybrid Transformer, which extracts bottom-level features using a CNN and models global interaction through the Transformer. Cai *et al.* [2] combined various Transformer frameworks to make predictions, and Huang *et al.* [15] proposed a lightweight transformer network for gaze estimation by using self-attention mechanisms. However, when replacing convolution with the Transformer or stacking them sequentially, it may lead to the loss of semantic information and features, affecting gaze estimation accuracy. Therefore, new methods are needed to address this problem.

3. Methods

3.1. Overall architecture

The architecture of CTA-NET is shown in Fig. 1, which mainly consists of a Gaze Transformer-ResNet encoder, an attention cross fusion (ACFusion), and a dual-feature aggregation (DFA) module. We first adopt a parallel Transformer-ResNet network to extract facial and eye information, respectively. Specifically, the Transformer branch is utilized to process facial images and starts with global self-attention, which is then followed by the restoration of detailed local features. Meanwhile, the CNN branch processes eye images, and the ACFusion module is employed to enhance the expression ability of local features by increasing the receptive field through two layers. The features extracted from both branches are fed into our proposed DFA module, which selectively fuses the information using attention and Hadamard addition. The multi-level fused feature maps are combined using residual connections, and then passed through a fully connected layer for gaze estimation.

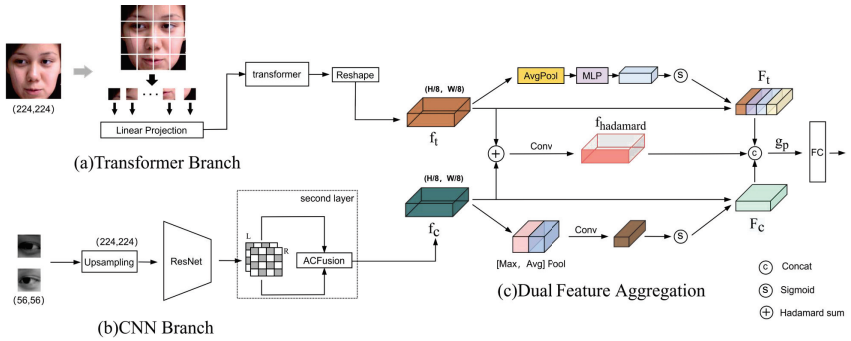


Fig. 1. An overview of the CNNs and Transformer aggregation network framework(CTA), which includes two parallel branches(Transformer and CNN), and our proposed Dual Feature Aggregation module

3.2. Gaze Transformer-ResNet Encoder

Unlike previous methods[6][2], we do not send the facial and eye images together to the Transformer for gaze regression after processing with the CNN. This decision stems from our consideration that the Transformer relies on the concept of position encoding, and the amalgamation of features from distinct facial regions may compromise the inherent positional information of the subject. Such compromise has the potential to diminish the accuracy of gaze estimation. In lieu of this approach, our paper introduces a two-stream framework, meticulously designed to concurrently extract features from diverse image types. The merit of this parallel processing framework lies in its ability to not only capture comprehensive global information but also retain sensitivity to local details. This dual focus proves advantageous in achieving more precise and accurate gaze estimation.

Transformer Branch The Transformer branch adopts a traditional encoder-decoder architecture, starting with the input image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ being divided into $N = \frac{H}{S} \times \frac{W}{S}$ blocks with S typically set to 16. These blocks are flattened and linearly projected into a D_0 -dimensional output, generating the original embedding sequence $\mathbf{e} \in \mathbb{R}^{N\times D_0}$. To incorporate prior knowledge, a learnable position embedding is added to $\mathbf{e}$. Then, the generated embedding $\mathbf{z}^0 \in \mathbb{R}^{N\times D_0}$ is used as the input to the Transformer encoder, which consists of L layers of multi-head self-attention (MSA) and multi-layer perceptron (MLP) [27]. The self-attention module is the core of the Transformer encoder, and updates the state of each embedded patch by globally aggregating information across all layers. By applying the softmax function, the inner product between the query(q) vector and the key(k) vector is normalized, resulting in a relative weight assigned to each elemenAt of the global information. This facilitates the aggregation and updating of the global information, as expressed through the following formula:

$$\mathrm{SA}\left(\mathbf{z}_i\right) = \mathrm{softmax}\left(\frac{\mathbf{q_i}\mathbf{k}^T}{\sqrt{D_h}}\right)\mathbf{v}, \tag{1}$$

where $[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z}\mathbf{W}_{qkv}, \mathbf{W}_{qkv} \in \mathbb{R}^{D_0 \times 3D_h}$is the projection matrix, $\mathbf{z}_i \in \mathbb{R}^{1 \times D_0}$ and $\mathbf{q}_i \in \mathbb{R}^{1 \times D_h}$ are the i-th rows of $\mathbf{z}$ and $\mathbf{q}$, respectively. MSA is extended by concatenating multiple self-attention modules SA and projecting the latent dimension to $\mathbb{R}^{D_0}$. The MLP is a dense layer stack consisting of fully connected layers with GELU activation and Dropout, where the first fully connected layer increases the number of input nodes by four times, and the second fully connected layer recovers the original number of nodes. The output of the last transformer layer is layer-normalized to obtain the encoded sequence $\mathbf{z}^L \in \mathbb{R}^{N \times D_0}$. The encoded sequence $\mathbf{z}^L$ is reshaped to produce the final output $\mathbf{f}_{\mathrm{t}} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times D_0}$, which is then jointly inputted into the DFA module with the feature map outputted by the CNN branch for sufficient fusion.

CNN Branch Our CNN branch is designed based on the ResNet network architecture. The traditional ResNet contains 4 layers to obtain enough low-level information while avoiding excessive consumption of model resources from overly deep networks, which can increase the model's sensitivity to noise, interference, and adversarial attacks. In our approach, we specifically chose the output of the second layer $CONV \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$ as the feature information for subsequent fusion. This choice aims to enhance the overall generalization ability. To better preserve the details of the image and reduce information loss, we introduce an upsampling operation before feature fusion. Specifically, we first use bicubic interpolation to upsample the left and right eye images (56×56→224×224). Compared with traditional bilinear interpolation, bicubic interpolation has better edge-preserving ability and performs better in visual tasks that require accuracy. Next, the ResNet network is utilized to extract features to generate the original feature images $\mathbf{f}_L \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$ and $\mathbf{f}_R \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$. These paired features are then input into our proposed feature-enhanced attention module ACFusion to fuse the left and right eye features. Furthermore, The feature $\mathbf{f}_{\mathrm{c}} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_1}$ obtained by the CNN branch is then combined with the feature map output by the Transformer branch and input into the DFA module for comprehensive fusion.

3.3. Attention Cross Fusion

This article proposes a module named ACFusion Fig. 2 to address various issues in the task of gaze estimation caused by the fusion of information from the left and right eyes, such as feature noise, information loss, and region imbalance. The ACFusion module combines residual connections with attention mechanisms to effectively fuse the feature information from the left and right eyes. Specifically, the ACFusion module employs a channel attention mechanism to assign different weights to feature maps, emphasizing important features while suppressing noise. For the left eye feature $\mathbf{t}_{\mathrm{L}}$, the ACFusion module first calculates the weight of each channel and adjusts the feature map based on these weights to retain valuable features. The channel attention mechanism is then used to further enhance the left eye feature map. The attention-enhanced left eye feature map is multiplied by the right eye feature $\mathbf{t}_{\mathrm{R}}$, and their spatial information is encoded and mapped using a 3x3 convolutional operation to highlight or suppress the necessary regions. The resulting convolutional features are calculated using the ReLU function to retain useful feature information. To further improve the quality of the fused feature representation, the ACFusion module employs a spatial attention mechanism to enhance the

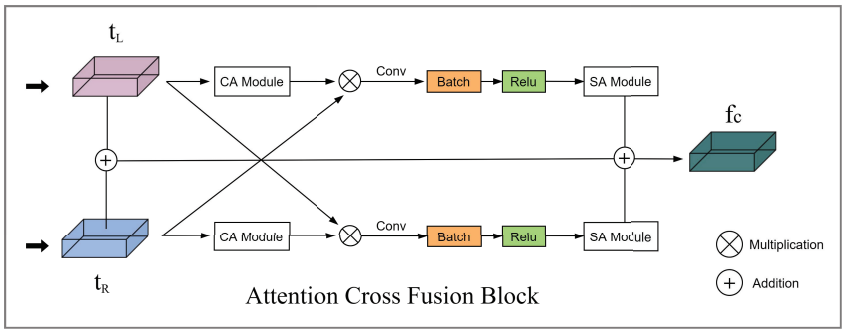


Fig. 2. Schematic diagram of the attention cross-fusion module. As shown in the figure, it includes a channel attention module(CA Module), convolution module, and spatial attention module(SA Module). The feature map is converged by each module in turn and cross-fused with the original feature map to get the final output. The fusion operation of the eye image is carried out simultaneously

global contextual information of the feature map. In this process, the refined left and right eye feature maps are progressively combined with the original feature map to obtain an information-enhanced feature map. Finally, these feature maps are concatenated and output as the fused feature $\mathbf{f}_c$.

The ACFusion module effectively utilizes the left and right eye feature information. Compared with previous single extraction methods, the ACFusion module can more effectively capture the left and right eye feature information while reducing feature noise and information loss. By introducing attention mechanisms, the ACFusion module can also better adapt to the imbalance of image regions, further improving the quality of the fused feature representation.

3.4. Dual Feature Aggregation

Due to the issues of dimension mismatch, context integration, and imbalanced feature representation in the fusion of encoding features from CNN and Transformer, in order to balance the feature expression capabilities of both and ensure comprehensive representation and integration of global and local contextual information, we propose a novel DFA Block (see Fig. 1(c)) that combines attention and multi-layer feature fusion mechanisms. Specifically, we leverage the SE-Block introduced in [14] to incorporate channel attention with residual connections to process the global information $\mathbf{f}_t$ from the Transformer branch, resulting in the featured representation $\mathbf{F}_t$:

$$\mathbf{F}_t = \text{Residual}\left([\mathbf{f}_t,\ \text{ChannelAttn}\ (\mathbf{f}_t)]\right). \tag{2}$$

Here, $Residual()$ denotes the residual connection operation, while $ChannelAttn()$ denotes the complete channel attention module, and $SpatialAttn()$ denotes the complete spatial attention module. Considering that high-level features in CNNs typically possess a larger receptive field and stronger semantic information, we adopt spatial attention from CBAM [34] as a spatial filter combined with residual connections to process the feature

information $\mathbf{f}_{\mathrm{c}}$ from the CNN branch, enhancing local details and suppressing irrelevant regions to obtain the feature $\mathbf{F}_{\mathrm{c}}$:

$$\mathbf{F}_{\mathrm{c}} = \mathrm{Residual}\left([\mathbf{f}_{\mathrm{c}}, \mathrm{SpatialAttn}\left(\mathbf{f}_{\mathrm{c}}\right)]\right). \tag{3}$$

Next, Hadamard addition is applied to element-wise sum the outputs from both branches to alleviate the issue of gradient vanishing, preserve the positional information of the original feature maps, and enhance the expressive capacity of the model:

$$\mathbf{F}_{\text{hadamard}} = \mathrm{Conv}\left(\mathbf{f}_{\mathrm{t}} \oplus \mathbf{f}_{\mathrm{c}}\right).$$

Here, $\mid \oplus \mid$ denotes Hadamard addition, and Conv refers to a 3x3 convolutional layer. Finally, the interactive feature $\mathbf{F}_{\text{hadamard}}$ is concatenated with the attention features $\mathbf{F}_{\mathrm{t}}$ and $\mathbf{F}_{\mathrm{c}}$, followed by a final fully connected layer. This step allows for effective integration of global and local contextual information, resulting in the feature representation $\mathbf{g}_p$ that effectively captures the global and local context at the current spatial resolution:

$$\mathbf{g}_{\mathrm{p}} = \left(\mathbf{F}_{\mathrm{t}}\right) \mathrm{Concat}\left(\mathbf{F}_{\mathrm{c}}\right) \mathrm{Concat}\left(\mathbf{F}_{\text{hadamard}}\right). \tag{4}$$

Here, the $Concat$ operation entails merging the outputs of multiple independent self-attention modules along the feature dimension. The proposed block combines two attention mechanisms to enhance the learning of global and local contextual information.Firstly, the SE block introduced in [14] is employed to fuse channel attention with residual connections, aiding the flow of global information in the Transformer branch. Secondly, spatial attention is utilized as a spatial filter combined with residual connections to selectively emphasize local details and suppress irrelevant regions. This approach leverages the fact that high-level features in neural networks typically possess larger receptive fields and stronger semantic information. By combining these attention mechanisms, our goal is to effectively capture both global and local contextual information. To further enhance the model's expressive power, we utilize the Hadamard addition operator to element-wise combine the outputs of the two branches, effectively integrating both global and local contextual information. Finally, the interactive feature $\mathbf{F}_{\text{hadamard}}$ and the attention features $\mathbf{F}_{\mathrm{t}}$ and $\mathbf{F}_{\mathrm{c}}$ are concatenated and passed through a final fully connected layer. This step enables the model to effectively integrate global and local contextual information, resulting in the feature representation $\mathbf{g}_{\mathrm{p}}$ that captures contextual information at the current spatial resolution.

3.5. Loss Function

Deep supervision is a technique that introduces intermediate supervision signals at multiple stages of the network. In our approach, we employ deep supervision to enhance the training process[24]. Specifically, we incorporate the Transformer branch and fusion branch into our architecture, serving as additional sources of supervision. These branches aim to capture different aspects and features of the input data.By adding these intermediate supervision signals, our goal is to improve gradients and facilitate convergence during network training. Deep supervision signals provide valuable information at different levels of abstraction, enabling the network to learn more effectively and efficiently. This

approach not only helps optimize the overall loss function but also enhances the interpretability and generalization capability of the network. We formulate a comprehensive objective function as follows:

$$\mathcal{L} = \alpha \mathrm{L}\left(\mathrm{G}, \mathbf{f}_{\mathrm{t}}\right) + \beta \mathrm{L}\left(\mathrm{G}, \mathbf{f}_{\mathrm{c}}\right) + \gamma \mathrm{L}\left(\mathrm{G}, \mathbf{g}_{\mathrm{p}}\right), \tag{5}$$

where α, β, and γ are adjustable hyperparameters that control the weights of each component. $\mathrm{L}\left(\mathrm{G}, \mathbf{f}_{\mathrm{t}}\right)$ represents the weighted IoU loss between the predicted facial gaze values $\mathbf{f}_{\mathrm{t}}$ and the ground truth (G). This term encourages accurate localization of facial pixels and precise depiction of facial regions. $\mathrm{L}\left(\mathrm{G}, \mathbf{f}_{\mathrm{c}}\right)$ represents the binary cross-entropy loss between the predicted facial gaze values and the ground truth, promoting pixel-wise classification accuracy of facial regions. Finally, represents a specific loss term that captures additional information related to facial geometry. By combining these different loss terms, the network is guided to learn robust and accurate facial gaze values. Through experimentation and iterative optimization of hyperparameters, we can find the optimal balance between the different components of the loss function, thereby improving the overall performance of the network. In summary, adopting weighted IoU loss, binary cross-entropy loss, and deep supervision techniques contributes to enhancing the effectiveness and robustness of our gaze prediction network. These strategies enable the network to leverage both pixel-level and high-level cues to accurately depict and localize facial regions, making it suitable for various practical applications such as facial analysis, recognition, and virtual reality.

4. Experiments

4.1. Setups

Dataset Our experiments are conducted on two main-stream datasets: MPIIGaze [40] and Gaze360 [17]. MPIIGaze is a kind of data set of gaze estimation based on appearance. It consists of MPIIGaze and MPIIFaceGaze, with MPIIGaze containing 15 subjects and 3,000 eye images per subject. The 3000 eye images are made up of 1500 left-eye images and 1500 right-eye images. MPIIFaceGaze, as an extension of MPIIGaze, contains the face image corresponding to each eye image in MPIIGaze. Note that MPIIGaze provides a standard evaluation method that selects 3000 images for each subject to make up the evaluation set. With leave-one-out evaluation, we do the experiment in the evaluation set, not in the whole evaluation set.

Gaze360 contains part of the back image without facial features, so we need to delete the image without face detection result according to the face detection annotation provided. Gaze360 contains a training set with 84K images for 54 subjects and a test set with 16K images for 15 subjects.

Data preprocessing We normalized both datasets in the same way. Specifically, we used virtual camera rotation and translation to eliminate the roll angle of the head and maintain the same distance between the virtual camera and the reference point (the center of the face). Additionally, we cropped eye images of size 56×56 from the normalized facial images, which were automatically detected using a face detection algorithm. The eye images were then histogram equalized and converted to grayscale to eliminate lighting effects.

Implementation details CTA-Net was built using the PyTorch framework and trained using an NVIDIA-A40 GPU. We used the Adam optimizer to train the entire model and used a linear learning rate warm-up with a warm-up phase of 5. For training on MPIIGaze, we use a batch size of 512 and 80 iterations, with a learning rate set to 0.0005 and a decay rate of 0.5. On Gaze360, we use a 256 batch size and 80 iterations, with a learning rate set to 0.0005 and a decay rate of 0.5. It is worth noting that we trained MPIIGaze using the leave-one-out method, taking an average of 15 sessions.

Evaluation metrics We use gaze Angle error as the evaluation measure of most gaze estimation methods. Assuming that the ground truth gaze direction is $\mathrm{G} \in \mathbb{R}^3$ and the predicted gaze vector is $\mathrm{g}' \in \mathbb{R}^3$, the gaze Angle error can be calculated as follows:

$$\mathcal{L}_{\text{angular}} = \frac{\mathrm{G} \cdot \mathrm{g}'}{\|\mathrm{G}\| \, \|\mathrm{g}'\|}. \tag{6}$$

The smaller the Angle error value is, the closer the model result is to the real value.

4.2. Comparison with appearance-based methods

To demonstrate the performance of our proposed method compared to other appearance-based gaze estimation methods on two datasets, MPIIFaceGaze and Gaze 360, we conducted a comprehensive evaluation. Considering the unique characteristics of the CTA-Net network architecture, we compared it with a total of nine convolutional network-based gaze estimation methods, including Full-face [39], Dilated-Net [4], RT-Gene [12], and GazeTR-Hybrid [6](the transformer-based gaze estimation method) to showcase its design advantages. Due to the unavailability of source code for some recent appearance-based methods, we referenced Cheng's survey [7] for some of our data.

Table 1. Performance in MPIIFaceGaze dataset. Due to the improved accuracy of RT-Gene through ensemble of four models, we also provided the results of model ensemble and referred to it as RT-Gene (4 models) to distinguish it from RT-Gene

Methods	MPIIFaceGaze	Gaze360 (Front 180°)	Realtime
iTracker (AlexNet)[18]	5.6°	None	$37ms$
MeNets[35]	4.9°	None	$36ms$
FullFace[39]	4.8°	14.99°	$37ms$
Dilated-Net[4]	4.8°	13.73°	$34ms$
RT-Gene[12]	4.8°	None	$34ms$
RT-Gene(4 ensemble)	4.3°	12.26°	$36ms$
Bayesian Approach[29]	4.3°	None	$32ms$
FAR-Net[8]	4.3°	None	$33ms$
CA-Net[5]	4.27°	12.26°	$34ms$
Gaze360[17]	None	11.40°	$33ms$
GazeTR-Hybrid[6]	4.00°	10.62°	$28ms$
CTA-Net(ours)	3.91°	10.44°	$30ms$

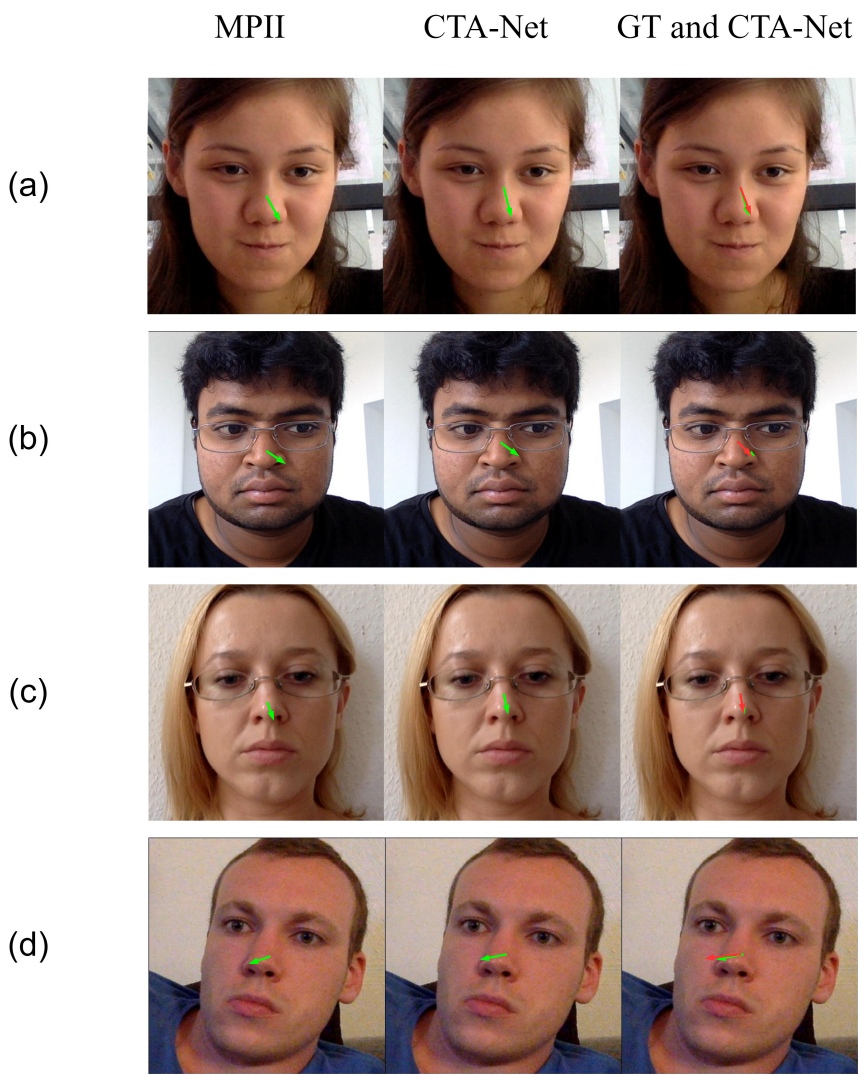


Fig. 3. The dataset is presented in a visualized format, where the ground truth is represented by a red arrow

Table 1 presents the results on the MPIIFaceGaze dataset. Our CTA-Net achieves lower angular error compared to the other methods. Specifically, our CTA-Net outperforms the highly accurate GazeTR-Hybrid with an angular error of 3.91°, representing an improvement of nearly 0.1° on the MPIIFaceGaze dataset. Table 1 displays the results on the Gaze 360 dataset. Following the division of the Gaze 360 dataset into train-val-test sets and evaluation ranges by Kellnhofer *et al.* [17], we adopt the same evaluation criteria, focusing on the frontal 180° range. This allows for a fair comparison with all relevant methods trained and evaluated on datasets within the 180° range. The proposed CTA-Net

achieves state-of-the-art gaze estimation performance with an average angular error of 10.44° in the frontal 180° range.

Fig. 3 illustrates the visualization results of the proposed method on the MPIIGaze dataset. Panels (a)-(d) show visualizations of eye directions for different subjects. Each group presents the visualizations for the same subject under different methods. We also compared the visualizations from the MPII model and the ground truth. The experimental results demonstrate that our method can adapt to different eye appearances while maintaining high accuracy and robustness.In summary, our research extends beyond theoretical contributions to underscore the practical efficacy of the proposed model. Through rigorous visual analyses in real-world application scenarios, depicted in Fig. 4, we confirm the model's robust generalizability. The positive outcomes observed in diverse contexts not only enhance theoretical understanding but also position the model as a versatile and effective solution for addressing real-world challenges. This dual validation, both in theory and practical application, establishes our model as a promising and impactful advancement in the field.



Fig. 4. The Citation Representation of Models in the Real World

4.3. Ablation study

We propose ablation experiments for the network structure and the proposed module respectively, and evaluate the effectiveness of the parallel branch design and the fusion module by changing the serial-parallel structure of the backbone network and the design choice of the fusion module.

Ablation experiments of structures. In order to better reflect the effectiveness of the parallel structure of CNN and Transformer, we design two simple network frameworks respectively, as shown in the figure, and recorde our experimental results on MPIIFaceGaze and Gaze360.

In this study, we present two architectures for gaze estimation: (a) a unified approach where both facial and eye images are processed by the same convolutional network before being fed into a transformer decoder and a fully connected layer for the final output, and

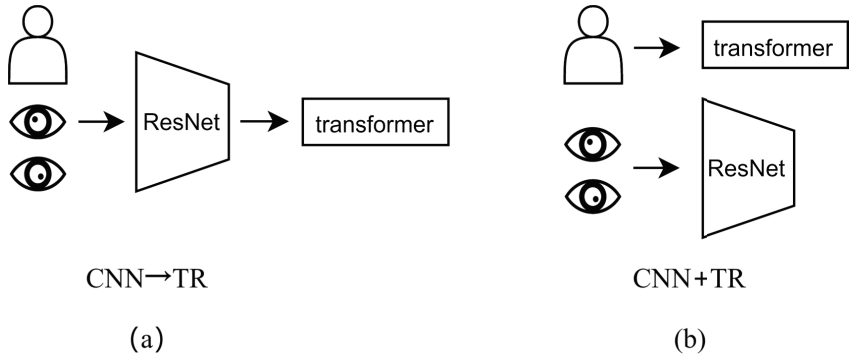


Fig. 5. The structure of CNN is mainly a pre-trained backbone network, including a pooling layer and fully connected layer, while the structure of the transformer is consistent with that in the text. Figure (a) is a serial structure, and Figure (b) is a parallel structure

(b) a parallel architecture inspired by CTA-Net, which utilizes separate convolutional and transformer networks that are connected via a fully connected layer, shown in Fig. 5. To ensure experimental accuracy, we do not incorporate any additional modules and relied solely on ImageNet pre-trained networks for our CNN backbone. Our results (Table 2) demonstrate that the parallel architecture outperforms the unified approach by a margin of 0.13° and 0.25° on MPIIFaceGaze and Gaze360 datasets, respectively.

Table 2. Ablation experiments of structures. 'CNN→TR'and 'CNN+TR'indicate a serial structure and a parallel structure, respectively

Methods	Backbones	Pre-train	MPIIFaceGaze	Gaze360
CNN→TR	ResNet18	ImageNet	4.47°	12.31°
CNN→TR	ResNet50	ImageNet	4.50°	13.16°
CNN+TR	ResNet18	ImageNet	4.44°	12.57°
CNN+TR	ResNet50	ImageNet	4.37°	12.21°

In addition, we conduct a thorough analysis of the rationality of the CNN branch design. Specifically, we perform comparative experiments to evaluate different upsampling methods and output layers of the CNN branch network. The tested upsampling methods include linear and bicubic interpolation, while the four different levels of the ResNet50 network are used as output layers. The overall architecture of the model remain unchanged, with the transformer branch generating an encoding sequence, which is reshaped to match the output channels of the CNN branch. The results, as shown in Table 3, indicates that the bicubic interpolation method exhibited higher accuracy compared to the linear interpolation method. In addition, our performance analysis of the output layer selection of CTA-Net demonstrates that it represented the optimal choice. These findings provide compelling evidence for the benefits of our proposed CNN branch design strategy.

Table 3. Ablation experiments of CNN branch

CNN-ResNet50		MPIIFaceGaze		Gaze360	
		bilinear	bicubic	bilinear	bicubic
Layer L	*L=1*	3.94°	3.94°	10.60°	10.59°
	L =2	3.93°	3.91°	10.51°	10.44°
	L =3	3.97°	3.95°	10.77°	10.71°
	L =4	4.02°	4.01°	10.61°	10.61°

Ablation experiments of modules.

In order to evaluate the impact of ACFusion and DFA modules on gaze estimation accuracy, we design CT-Net for module comparison, which replaces the module details in CTA-Net with fully connected layers for direct output. Results on both datasets are shown in Table 4(‘CTF-Net’stands for CT-Net+ACFusion network.), where our proposed method demonstrates a significant improvement in accuracy when incorporating the AC-Fusion module, with an increase of 0.16° and 1.92° on MPIIFaceGaze and Gaze360, respectively. At the same time, both datasets exhibit worse performance compared to CTA-Net, which proves the advantages of the proposed attention components. In addition, we present the gaze estimation accuracy of each participant in the MPIIFaceGaze dataset and compare different methods. Out of 15 participants, our proposed method achieves better gaze estimation accuracy for 9 participants, as shown in Fig. 6.

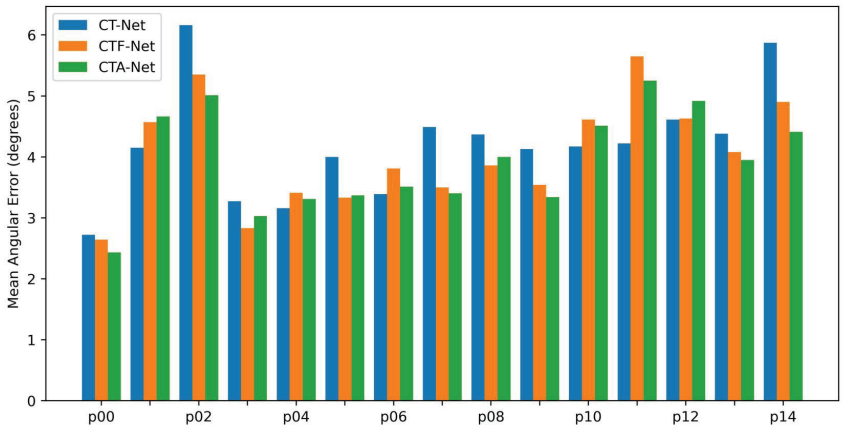


Fig. 6. The structure of CNN is mainly a pre-trained backbone network, including a pooling layer and fully connected layer, while the structure of the transformer is consistent with that in the text. Figure (a) is a serial structure, and Figure (b) is a parallel structure

In addition, we conduct ablation experiments to investigate the impact of the internal design of the modules.

Attention ablation:

Table 4. Ablation experiments of modules

Methods	MPIIFaceGaze	Gaze360
CT-Net	4.18°	12.51°
CTF-Net	4.02°	10.59°
CTA-Net(Single)	4.41°	14.01°
CTA-Net	3.91°	10.44°

We specifically focus on the ACFusion module and evaluate its influence on gaze estimation performance. The ACFusion module incorporates an attention mechanism to assign different weights to the left and right eyes, enabling the estimation of gaze residuals. To examine the role of the attention module, we propose an experimental configuration called "Without-Att", where the weights for the left and right eyes were fixed at 0.5, effectively removing the attention module for generating eye features. Furthermore, to explore the effect of the stacking order within the ACFusion module, we introduce a different stacking order called "Reverse-Att", which interchanges the spatial attention and channel attention. The results of these ablation experiments are presented in Table 5.

Compared to the complete ACFusion model, the attention ablation experiment shows a decrease in performance by 1.58° on the Gaze360 dataset. Similarly, the performance of the reverse attention order experiment is reduced by 2.27° compared to the CTA-Net. These results demonstrate the advantages of the attention component and the importance of the stacking order. By analyzing the findings from these ablation experiments, we can draw conclusions about the internal design of the ACFusion module and further validate its significance in gaze estimation tasks. This deeper understanding will contribute to unraveling the functionality and benefits of the ACFusion module.Furthermore, attention ablation experiments were conducted on the extraction of monocular feature information (CTA Net Single). As depicted in Figure 4, the accuracy of independently processing eye images decreased by nearly 0.5, rendering it ineffective in guiding gaze estimation. The fundamental reason for this phenomenon is that monocular feature information may lack sufficient context and comprehensiveness, posing a challenge in capturing the complexity of the visual system. Although eyes play a crucial role in visual perception, processing eye images independently may not fully capture the interaction and synergy between the eyes and the entire visual process. Therefore, our experimental results underscore the importance of considering binocular or more extensive visual information in gaze estimation tasks to achieve more accurate and comprehensive gaze estimation results.

Hadamard ablation:

To evaluate the impact of the DFA module on gaze estimation performance, we conduct an ablation experiment called "Without-Had". The DFA module is a critical component in our approach, utilizing Hadamard addition to generate feature maps. We design an ablation experiment specifically targeting the Hadamard addition operation within the DFA module. In the Hadamard ablation experiment, we retain the DFA module but remove the Hadamard addition operation, implying that we no longer employ Hadamard addition for generating feature maps. This allows us to assess the importance of the Hadamard addition operation within the DFA module. By comparing the performance

of these two ablation experiments with the complete model, we can evaluate the influence of the DFA module and the Hadamard addition operation on gaze estimation.

Based on the results from Table 5 and the ablation experiments, we observe a decrease in accuracy of 0.97° in the DFA module when the Hadamard addition operation is removed. This highlights the significance of the Hadamard addition operation in the DFA module for accurate gaze estimation

Table 5. Attention ablation and Hadamard ablation

Methods		Gaze360
Attention ablation	Without- Att	12.02°
Attention ablation	Reverse- Att	12.71°
Hadamard ablation	Without-Had	11.41°
CTA-Net		10.44°

4.4. Limitation and Future Work

Our model currently faces challenges due to a large number of overall parameters, resulting in prolonged training times. Furthermore, the complexity of the model structure introduces the risk of overfitting, especially when dealing with limited training data. This complexity may compromise the model's ability to generalize to new data. In the future, we are committed to thoroughly exploring these identified limitations and proposing practical solutions. Additionally, we will carefully consider the challenges posed by extensive training times and the risk of overfitting in our work. This is aimed at enhancing the efficiency and robustness of the model.

5. Conclusion

Our proposed dual-stream framework for accurate gaze estimation demonstrates exceptional functionality and meticulous internal design techniques. Through image fusion and the utilization of attention mechanisms, our approach preserves crucial positional information and effectively combines features from the left and right eyes. The attention cross-fusion (ACFusion) module addresses challenges such as feature noise and information loss, while the Dual Feature Aggregation (DFA) block integrates global and local contextual information. These designs yield superior accuracy in gaze estimation.

Although extensive experimental results on two mainstream datasets demonstrate the effectiveness of our proposed approach, it is important to note that the design based on this framework requires high-quality datasets and is inevitably influenced by environmental factors. Future research can focus on refining feature extraction methods to handle more complex scenarios and further improve accuracy and robustness.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (62102003), Anhui Postdoctoral Science Foundation (2022B623),Medical Special Cultivation Project

of Anhui University of Science and Technology (YZ2023H2B003),Huainan City Science and Technology Plan (2023A316), Natural Science Foundation of Anhui Province (2108085QF258), the University Synergy Innovation Program of Anhui Province (GXXT-2022-038), Central guiding local technology development special funds (202107d06020001), University-level general projects of Anhui University of science and technology (xjyb2020-04).

References

1. Biswas, P., et al.: Appearance-based gaze estimation using attention and difference mechanism. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3143–3152 (2021)
2. Cai, X., Chen, B., Zeng, J., Zhang, J., Sun, Y., Wang, X., Ji, Z., Liu, X., Chen, X., Shan, S.: Gaze estimation with an ensemble of four architectures. arXiv preprint arXiv:2107.01980 (2021)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision. pp. 213–229. Springer (2020)
4. Chen, Z., Shi, B.E.: Appearance-based gaze estimation using dilated-convolutions. In: Proceedings of the Asian Conference on Computer Vision. pp. 309–324. Springer (2018)
5. Cheng, Y., Huang, S., Wang, F., Qian, C., Lu, F.: A coarse-to-fine adaptive network for appearance-based gaze estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10623–10630 (2020)
6. Cheng, Y., Lu, F.: Gaze estimation using transformer. arXiv preprint arXiv:2105.14424 (2021)
7. Cheng, Y., Wang, H., Bao, Y., Lu, F.: Appearance-based gaze estimation with deep learning: A review and benchmark. arXiv preprint arXiv:2104.12668 (2021)
8. Cheng, Y., Zhang, X., Lu, F., Sato, Y.: Gaze estimation by exploring two-eye asymmetry. IEEE Transactions on Image Processing 29, 5259–5272 (2020)
9. Cristina, S., Camilleri, K.P.: Model-based head pose-free gaze estimation for assistive communication. Computer Vision and Image Understanding 149, 157–170 (2016)
10. Dari, S., Kadrileev, N., Hüllermeier, E.: A neural network-based driver gaze classification system with vehicle signals. In: Proceedings of the International Joint Conference on Neural Networks. pp. 1–7. IEEE (2020)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Fischer, T., Chang, H.J., Demiris, Y.: Rt-gene: Real-time eye gaze estimation in natural environments. In: Proceedings of the European Conference on Computer Vision. pp. 334–352 (2018)
13. Hansen, D.W., Ji, Q.: In the eye of the beholder: A survey of models for eyes and gaze. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(3), 478–500 (2009)
14. Huang, C.H., Wu, H.Y., Lin, Y.L.: Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. arXiv preprint arXiv:2101.07172 (2021)
15. Huang, H., Ren, L., Yang, Z., Zhan, Y., Zhang, Q., Lv, J.: Gazeattentionnet: Gaze estimation with attentions. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 2435–2439. IEEE (2022)
16. Huang, Q., Veeraraghavan, A., Sabharwal, A.: Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. Machine Vision and Applications 28, 445–461 (2017)
17. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6912–6921 (2019)

18. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2176–2184 (2016)
19. Liu, G., Yu, Y., Mora, K.A.F., Odobez, J.M.: A differential approach for gaze estimation with calibration. In: Proceedings of the British Machine Vision Conference. vol. 2, p. 6 (2018)
20. Lu, F., Chen, X., Sato, Y.: Appearance-based gaze estimation via uncalibrated gaze pattern recovery. IEEE Transactions on Image Processing 26(4), 1543–1553 (2017)
21. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Adaptive linear regression for appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(10), 2033–2046 (2014)
22. Mora, K.A.F., Odobez, J.M.: Person independent 3d gaze estimation from remote rgb-d cameras. In: Proceedings of the IEEE International Conference on Image Processing. pp. 2787–2791. IEEE (2013)
23. Morimoto, C.H., Mimica, M.R.: Eye gaze tracking techniques for interactive applications. Computer Vision and Image Understanding 98(1), 4–24 (2005)
24. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7479–7489 (2019)
25. Stampe, D.M.: Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. Behavior Research Methods, Instruments, & Computers 25, 137–142 (1993)
26. Sugano, Y., Matsushita, Y., Sato, Y.: Appearance-based gaze estimation using visual saliency. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(2), 329–341 (2012)
27. Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., et al.: Resmlp: Feedforward networks for image classification with data-efficient training. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Proceedings of the Advances in Neural Information Processing Systems 30 (2017)
29. Wang, K., Zhao, R., Su, H., Ji, Q.: Generalizing eye tracking with bayesian adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11907–11916 (2019)
30. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: using gaze to supervise computer-aided diagnosis. IEEE Transactions on Medical Imaging 41(7), 1688–1698 (2022)
31. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018)
32. Wang, Z., Wang, H., Yu, H., Lu, F.: Interaction with gaze, gesture, and speech in a flexibly configurable augmented reality system. IEEE Transactions on Human-Machine Systems 51(5), 524–534 (2021)
33. Williams, O., Blake, A., Cipolla, R.: Sparse and semi-supervised visual mapping with the sˆ 3gp. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 230–237. IEEE (2006)
34. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. pp. 3–19 (2018)
35. Xiong, Y., Kim, H.J., Singh, V.: Mixed effects neural networks (menets) with applications to gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7743–7752 (2019)
36. Yang, B., Cui, J., Tong, Y., Wang, L., Zha, H.: Recognition of infants' gaze behaviors and emotions. In: Proceedings of the International Conference on Pattern Recognition. pp. 3204–3209. IEEE (2018)

37. Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In: Proceedings of the European Conference on Computer Vision. pp. 365–381. Springer (2020)
38. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4511–4520 (2015)
39. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 51–60 (2017)
40. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(1), 162–175 (2017)
41. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6881–6890 (2021)

Chenxing Xia received the Ph.D. degree from Hunan University in 2019. He is currently an Associate Professor at the College of Computer Science and Engineering, Anhui University of Science and Technology. His research interests include saliency detection, computer vision, image processing, and depth prediction.

Zhanpeng Tao received the Anhui University of Science and Technology. He is currently pursuing the master degree with Anhui University of Science and Technology. His research interests include computer vision, machine learning, monocular 3D object detection,gaze estimation and depth estimation.

Wenjun Zhao received the bachelor degree from Wanjiang University of Technology. He is currently pursuing the master degree with Anhui University of Science and Technology. His research interests include computer vision, machine learning, monocular 3D object detection, and depth estimation.

Wei Wang received his bachelor's degree from the City College of Wuhan University of Science and Technology, He is currently a technician at Anyang Cigarette Factory, China Tobacco Henan Industrial Co., Ltd.,and his job title is an engineer. his research interests include automatic control and tobacco machinery.

Bin Ge received the Ph.D. degree in computer application technology from Hefei university of technology, Hefei China, in 2016. He began to teach at Anhui University of Science and Technology in 1999. Currently, he is a professor of information security at Anhui University of Science and Technology. His research interests include information security, Internet of things technology and artificial intelligence.

Xiuju Gao received the M.S. degree in Computer Science and Electronic Engineering from Hunan University, Changsha China, in 2016. She is currently an assistant at the College of Electrical and Information Engineering, Anhui University of Science and Technology. Her research interests include image/video processing and computer vision.

Kuan-Ching Li is a Distinguished Professor in the Dept of Computer Science and Information Engineering (CSIE) at Providence University, Taiwan, where he also serves as the Director of the High-Performance Computing and Networking Center. He received the Licenciatura in Mathematics, and MS and Ph.D. degrees in electrical engineering from the University of Sao Paulo (USP), Brazil, in 1994, 1996, and 2001, respectively. He published more than 250 scientific papers and articles and is author, co-author or editor of more than 25 books published by Taylor & Francis, Springer, and McGraw-Hill. Professor Li is the Editor in Chief of the Connection Science and serves as an associate editor for several leading journals. Also, he has been actively involved in many major conferences and workshops in program/general/steering conference chairman positions and has organized numerous conferences related to computational science and engineering. He is a Fellow of IET and a Senior Member of the IEEE. His research interests include parallel and distributed computing, Big Data, and emerging technologies.

Yan Zhang received the Ph.D. degree in computer application technology from Anhui University in 2010. She is a professor and Ph.D. supervisor at Anhui University of Science and Technology. Her research interests include information security, data mining, etc.

Received: November 16, 2023; Accepted: February 05, 2024.

Computer Science and Information Systems 21(3):851–897 https://doi.org/10.2298/CSIS230804016L

The Security and Privacy Challenges Toward Cybersecurity of 6G Networks: A Comprehensive Review

Yanlu Li[1,2], Yufeng Xiao[1,2], Wei Liang[1,2⋆], Jiahong Cai[1,2], Ronglin Zhang[1,2], Kuan-Ching Li[3,⋆] and Muhammad Khurram Khan[4]

[1] School of Computer Science and Engineering, Hunan University of Science and Technology, China
[2] Hunan Key Laboratory for Service Computing and Novel Software Technology, China
yanluli@mail.hnust.edu.cn, hnxiaoyf@hnust.edu.cn, wliang@hnust.edu.cn, jiahongcai@mail.hnust.edu.cn, ronglin@mail.hnust.edu.cn
[3] Dept. of Computer Science and Information Engineering (CSIE), Providence University, Taiwan
kuancli@pu.edu.tw
[4] Center of Excellence in Information Assurance, King Saud University, Saudi Arabia
mkhurram@ksu.edu.sa

Abstract. The integration of 6G networks with emerging key technologies such as blockchain, artificial intelligence, and digital twins continues to improve. However, it carries many issues with security threats and challenges of 6G networks. In this article, we analyzed the security issues of 6G networks and presented some possible solutions. First, we discussed the developments of mobile communication technology, the research motivation of 6G networks, the comparison of Key Performance Indicators (KPIs) between 5G/6G networks, and the key technologies of 6G networks. Next, security threats in the 6G network were analyzed concerning architecture, major visions and related applications. This was followed by solutions to security issues in applying key technologies for 6G networks. We also presented the application of AI in solving 6G network security problems. Firstly, we illustrated the impact of AI on 6G networks from two aspects: AI promotes the construction of 6G networks, and AI brings security threats to 6G networks. Then, we demonstrated that AI can assist 6G networks in solving security problems in many ways. Lastly, the is summarized, and the future directions in this area are proposed.

Keywords: 6G security, privacy protection, AI, VLC security, THz security, blockchain security.

1. Introduction

Mobile communication technologies have evolved from 1G to 5G, and the development of 6G is underway [1]. The first generation of communication technology (1G) utilizes analog technology and Frequency Division Multiple Access (FDMA), such as the Advanced Mobile Phone (AMPS) in the United States, which is a typical representative of 1G networks [2]. However, 1G is limited by distance and network capacity during transmission and also has disadvantages such as low spectrum utilization, limited service variety, and poor confidentiality performance. The second-generation communication technology

⋆ Corresponding authors

(2G) uses digital technologies such as Time Division Multiple Access (TDMA) and Code Division Multiple Access (CDMA) to provide better voice quality and data transmission rates. Later, the third-generation communication technology (3G) was based on broadband wireless connectivity technologies such as TD-SCDMA in China, CD-MA2000 in the United States, and WCDMA in Europe. 3G networks connect the international Internet and wireless communication, providing better voice quality and data transmission rates [3]. The fourth generation communication technology (4G) combines 3G communication technology and WLAN technology, abandoning CDMA, with Orthogonal Frequency Division Multiple Access (OFDM) and Multiple-In Multiple-Out antenna (MIMO) technology as core technologies [4]. 4G has higher data transmission rates and much lower network latency than 3G. The fifth generation communication technology (5G) uses the millimeter wave band and MIMO technology [5]. Compared with 4G networks, it has lower energy consumption, lower latency, and faster speed, with the characteristics of ubiquitous networks and interconnection of all things [6].

In March 2019, the first 6G summit was held in Finland [7], where communication research experts worldwide worked together to draft the first 6G white paper [8]. Subsequently, governments and regional organizations began to research next-generation mobile communication technologies. China established a national 6G technology R&D promotion working group and expert group in November 2019, Huawei also launched a 6G R&D lab in Canada. China Mobile and Tsinghua University established a strategic partnership to collaborate in scientific research on 6G networks. The Korea Institute of Communication and Information Science held a 6G forum in April 2019 to officially announce the start of research on 6G networks. At the Mobile World Congress America (MWCA 2018) in September 2018 [9], US Federal Communications Commission (FCC) officials looked ahead to 6G technology for the first time in a public forum. The New York University Wireless Center (NYU Wireless) is working on wireless technology using terahertz frequency channels with transmission rates of up to 100 Gbps, and the University of California's ComsenTer Research Center has been awarded a $27.5 million grant to research the fusion of terahertz communication and sensor.

6G will outperform 5G in many aspects, such as latency, peak rate, connection density, traffic density, mobility, and spectral efficiency [10]. Regarding data transmission rate, 6G is expected to reach 50 times that of 5G, with peak transmission speeds of 100Gbps-1Tbps compared to 10Gbps for 5G. Regarding network latency, 6G networks will be reduced to one-tenth of 5G networks. Regarding connection number density, indoor positioning accuracy reaches 10 cm and outdoor is 1 m, which is 10 times higher compared to 5G and the density of connected devices reaches more than 100 per cubic meter. 6G networks use Terahertz (THz) band communication, and network capacity will be significantly increased [11]. Along with the rapid development of advanced technologies such as digital twins, blockchain technology, artificial intelligence, and space communication, the 6G network will realize the vision of a "smart connection of everything".

However, as network coverage expands and network heterogeneity increases, the security and privacy of 6G networks will be worse than in previous generations of mobile communication networks. Notably, 6G networks will face more severe security challenges, since (1) 6G networks use AI algorithms to train data, and service providers will collect a large amount of user data for model training to improve the model. These data will cover a large amount of user privacy. Once the data are leaked or attacked, the conse-

quences will be severe [12], (2) 6G network combined with the Internet of Things (IoT), the access and management of large-scale, low-power IoT devices can easily cause signaling storms. Moreover, the direct interaction of IoT devices is more frequent, as many IoT devices have higher requirements and security for group communication [13]. Thereafter, if the nodes are the wrong ones, it may bring out large-scale security problems [14], and (3) 6G is a crucial way to achieve air-space-ground-sea integration, this will expose it to more complex security issues. For example, at the UAV network level, UAVs are more susceptible to malicious attacks than ground-based stations because they cannot support complex encryption algorithms [15].

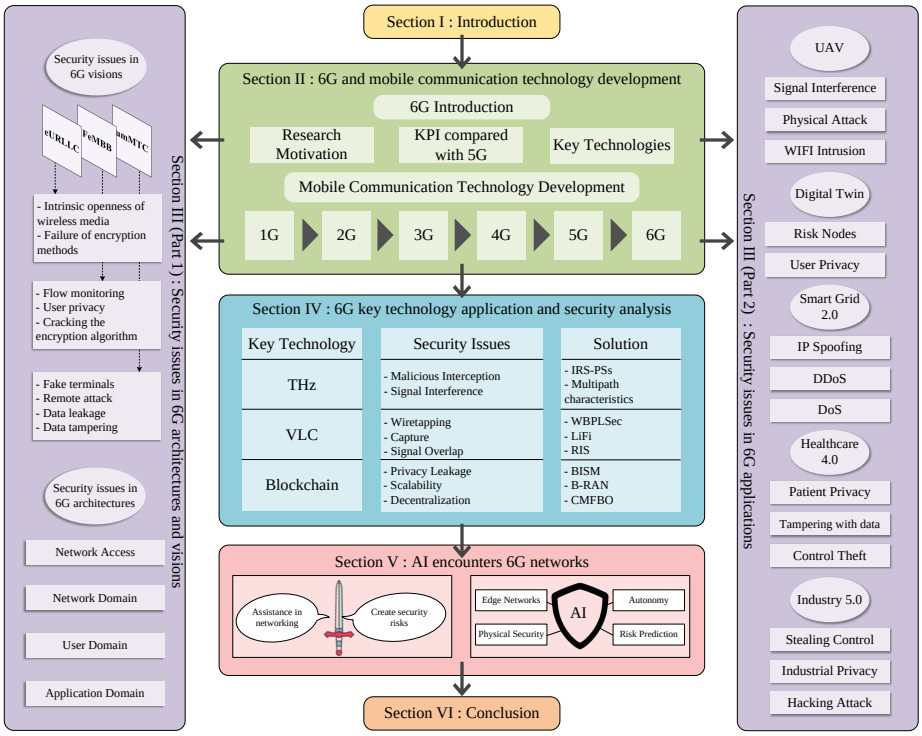


Fig. 1. The overall idea of this article

This work addressed the security issues in 6G networks, the structure and ideas of this work are shown in Figure 1. The remainder of this article is distributed as follows: In section 2, the development process of mobile communication technology from 1G to 6G is briefly introduced, and the research motivation of 6G networks is discussed. In addition, the KPIs of 6G networks are compared with 5G networks, and some key technologies of 6G networks are listed. Section 3 starts with the security issues of 6G in the network architecture, then compares the three main visions of 5G and 6G networks and lists the security issues. Also, in this section, we analyze the security issues of 6G networks in UAV mobile communication, digital twin, smart grid 2.0, healthcare 4.0, and

industrial 5.0. In section 4, we list the key technologies of 6G networks, such as terahertz communication, visible light communication, and blockchain technology, which are listed to analyze the security problems of using these technologies in 6G networks. We also offer some possible solutions to these problems. In section 5 of this work, we first mention that the application of AI in 6G networks is a double-edged sword and illustrate this point in terms of AI driving the construction of 6G networks while AI creates cyber risks for 6G networks. Next, four perspectives are presented to demonstrate that AI can assist 6G networks in addressing security threats. Finally, the work is summarized, and the future research directions are prospected.

2. Development of the Mobile Communication Technology and 6G

The second part introduces the development of mobile communication technology, summarizing the characteristics of each generation of communication networks from 1G to 5G and the enhancements over previous generations of each communication technology. Later, we introduced 6G networks in detail to raise the issue of security threats in 6G networks.

2.1. Development and Security Analysis of Mobile Communication Technology

Since the 1980s, mobile communication has produced a new generation of revolutionary technologies in a ten-year cycle. It continues to accelerate the iterative upgrade of the information industry and continuously promotes the prosperity of the economy and society. Today, it has become an indispensable primary information network to connect human society. However, the exploration of communication mobile networks is endless, communication technologies have evolved from 1G communication networks to 5G communication networks, and 5G networks are now being deployed on a large scale worldwide. At the same time, research on 6G communication networks is underway [16]. This part is a general overview of the development of the first five generations of communication networks. We briefly introduce the evolution of mobile communication technology in Figure 2.

The First Generation (1G) In 1974, Bell Labs developed the world's first mobile cellular telephone system, the Advanced Mobile Phone System (AMPS). 1G is developed to simulate information through electrical signals or electromagnetic waves directly. In the early stages, it was developed mainly to enable cell phone communication through distributed transceivers and networks.

As mentioned in [17], 1G networks can provide data rates of 1kbps to 2.8kbps and use circuit switches. It uses a bandwidth of 40 MHz, and its frequency range is between 800 and 900 MHz. The underlying technologies used in 1G networks are Advanced Mobile Phone Services (AMPS), Nordic Mobile Talk (NMT), and Cellular Digital Packet Data (CDPD).

Since the 1G network is an analog technology, call information must be transmitted through open-air electromagnetic waves. Therefore, lawbreakers can use technical means to intercept and decrypt this information, thus eavesdropping on the content of the call. It shows that the security of the communication network at that time was a matter of concern.

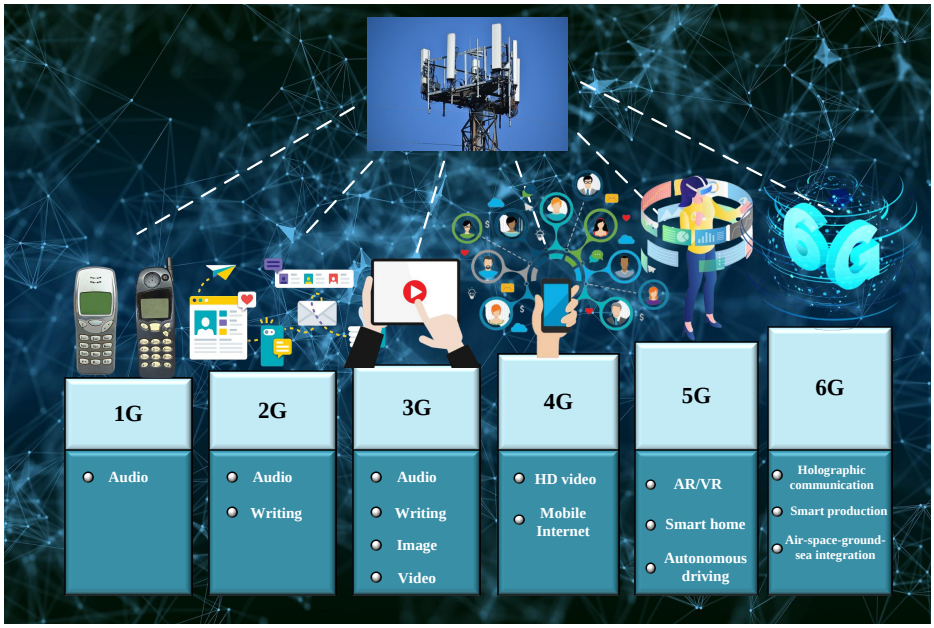


Fig. 2. History of mobile communication technology development

The Second Generation (2G) The second-generation mobile communication technology (2G) is based on digital language transmission technology as the core. In addition to the call function, some systems also introduce SMS functions and support data transmission and simulation. But the user experience rate is 10 kbps, and the peak rate is only 100 kbps, so it is only suitable for transmitting low volume of information such as e-mail, software, and others. 2G technology is divided into two types: one is based on TDMA, represented by GSM, and the other is CDMA, a form of multiplexing [18].

The 2G networks have the disadvantages of low transmission rate, unstable network, and high maintenance cost. The confidentiality of communication in 2G networks is also poor, which can easily cause security problems such as privacy leakage.

The 2G networks use weak encryption between the base station and the device, an encryption that can be easily cracked by an attacker in real time to intercept a call or text message. There is also no base station authentication technology in 2G networks, so anyone can seamlessly impersonate a 2G base station, and cell phones using the 2G protocol cannot recognize the authenticity. Fake BTSs sometimes capitalize on 2G's security vulnerabilities in order to intercept users' communications messages.

The Third Generation (3G) 3G mobile communication network is a cellular mobile communication technology supporting high-speed data transmission. Its rate is generally above several hundred kbps, it can transfer sound and data information. The main difference between 3G and 2G is the transmission rate. 3G can provide 144kbps transmission speed, which is faster compared to 2G. 3G has higher transmission rates, supports more advanced multimedia, and supports access and global roaming. It also allows you

to connect your phone to the Internet or other IP networks for voice calls, video calls, downloads, and data uploads [19].

Although 3G has a higher transmission rate, larger system capacity, and better communication quality than the previous two generations of communication networks, it has disadvantages such as limited capacity, incompatibility, inability to provide automatic roaming, and difficulty in achieving call confidentiality.

The 3G networks do not introduce public key cryptosystems in their security design, which makes it challenging to implement user digital signatures, thus reducing the security of the network. Although advanced cryptographic algorithms such as AES were used in the 3G network, the latest cryptographic achievements of the time, such as the ECC elliptic curve cryptographic algorithm, were not introduced. These pose security threats to 3G networks.

The Forth Generation (4G) The fourth-generation mobile network technology (4G) combines 3G and WLAN in one, it can download at speeds of over 100 Mbps, which is 25 times faster than home broadband ADSL. It meets many requirements of users for wireless services. In addition, 4G can be deployed in areas not covered by DSL and cable modems and then expanded throughout the area [20].

The core technologies of 4G are access methods, multiple access schemes, modulation technologies, coding technologies, high-performance receivers, smart antenna technologies, MIMO technologies, software radio technologies, etc. 4G networks provide mobile network access, IP telephony, gaming services, HD mobile TV, and video conferences.

The communication speed of 4G can be limited by the capacity of the communication system. As a result, if the amount of users is large, the communication speed will become slow. Moreover, due to the limited research conditions at that time, many difficulties were encountered during its construction and development.

4G networks may face the threat of Distributed Denial-of-Service (DDoS) attacks. This refers to an attacker sending a large number of invalid or high-traffic network requests to a target server by controlling a large number of botnet hosts, causing the server to run out of resources and not be able to respond appropriately to legitimate user requests.

The Fifth Generation (5G) The fifth generation mobile communication technology (5G) is a new communication technology with high speed, low latency, and ample connectivity, it is a network that enables the interconnection of people and machines [21]. 5G has a peak data download speed of up to 10 Gbps, nearly 100 times higher than 4G networks. It also has a significant increase in capacity compared to 4G networks due to the construction of many base stations. At the same time, 5G networks also have lower latency, higher bandwidth, and more excellent connectivity [22]. 5G networks are combined with technologies such as digital twin, blockchain, and IoT. And 5G is applied in various fields such as healthcare, education, logistics, and travel. It brings great convenience to our life [20,23]. Nevertheless, as the 5G networks are used in a broader range of scenarios, this makes it trickier to deal with a wide variety of security threats.

2.2. 6G Communication Network Introduction

In this section, we will introduce 6G networks from three aspects: research motivation, key performance indicators, and key technologies.

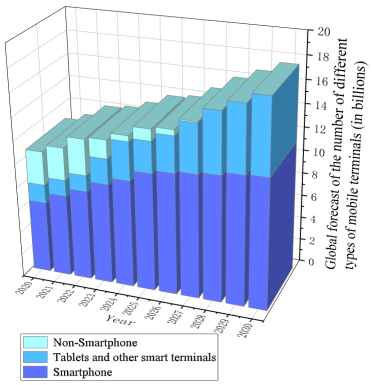


Fig. 3. Global forecast of the numberof different types of mobile terminals

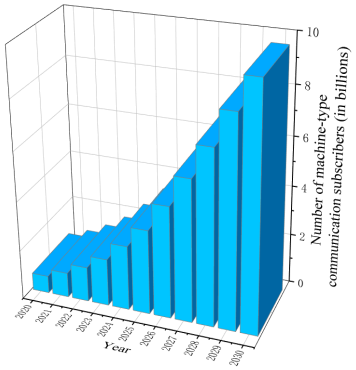


Fig. 4. Global forecast of the number of machine-based communication subscribers

6G Research Motivation With the rapid development of communication technology, 5G communication technology was born. It has the characteristics of low latency, high speed, ubiquitous network, and low power consumption. The International Telecommunication Union (ITU) defines three major types of 5G application scenarios [24], namely ultra-reliable low-latency communications (uRLLC), enhanced mobile broadband (eMBB), and massive machine-type-communications (mMTC).

5G networks have a massive impact on society. As a significant technology in digital transformation, 5G networks are changing how people live and work. Technologies such as smart cities, automated driving, and remote surgery are becoming a reality, promoting innovation in healthcare, logistics, agriculture, education, and other fields. Also, 5G is rapidly boosting the development of the world economy, with literature [25] indicating that 5G is expected to achieve approximately $12.3 trillion in global economic output in 2035, exceeding the total consumer spending of China, Germany, France, UK, and Japan in 2016.

Although 5G technology brings excellent convenience to society, as 5G is commercialized, many demands are raised that cannot be realized by 5G. Figure 3 and Figure 4

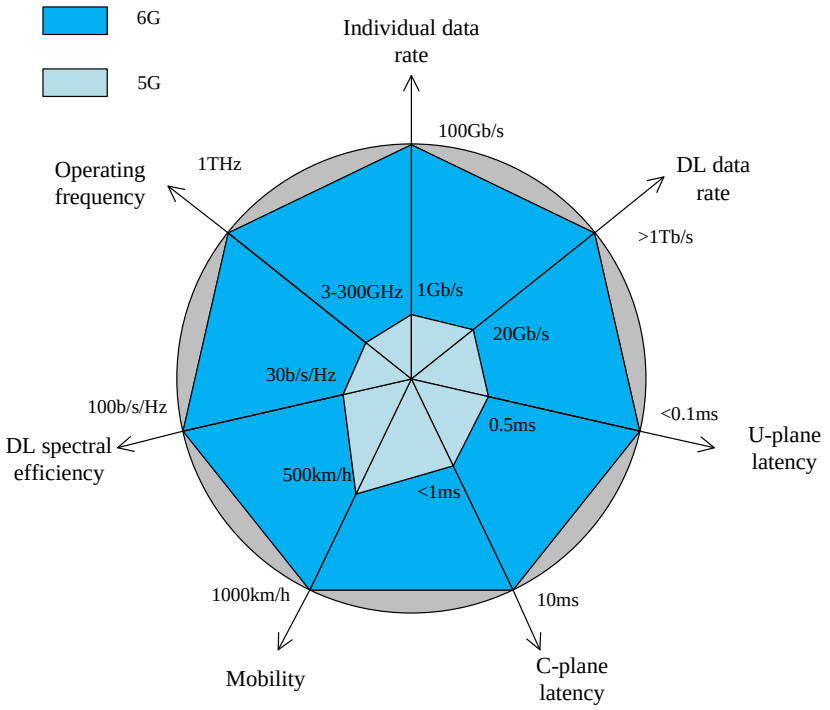


Fig. 5. Comparison of KPIs between 5G and 6G networks

show the bar graphs of future global mobile services and the number of terminal devices forecast in the Global Mobile Services Forecast 2020-2030 report [26] published by the International Telecommunication Union (ITU). Research on 6G wireless communication networks has been initiated in this context.

Comparisons of KPIs between 5G and 6G networks Compared with 5G networks, 6G networks will usher in more significant improvements in various aspects, such as transmission rate, traffic density, positioning accuracy, connection number density, mobility, etc. Due to the improved performance, 6G networks will also see an inevitable increase in energy consumption.

In March 2019, Bell Labs, USA, proposed some key performance indicators for 6G networks. In December 2020, the Chinese Communications Standardization Association also presented key indicators for 6G networks and compared 5G network indicators with 6G. Jiang *et al.* estimated the requirements and characteristics of 6G in [27]. The comparison of the key performance indicators between 5G and 6G networks is shown in Figure 5.

Key Technologies in 6G The widespread use of advanced technologies such as digital twins and blockchain for digital transactions, privacy protection, and others, accelerating the research of 6G communication networks and promoting the development of next-generation communication technologies [28,29,30]. Some key technologies for 6G

networks are listed here: terahertz communication technology, air-space-ground-sea integrated communication technology, and visible light communication technology. The concepts of these four key technologies will be described, and how the technology can be applied in 6G networks will be shown in this part.

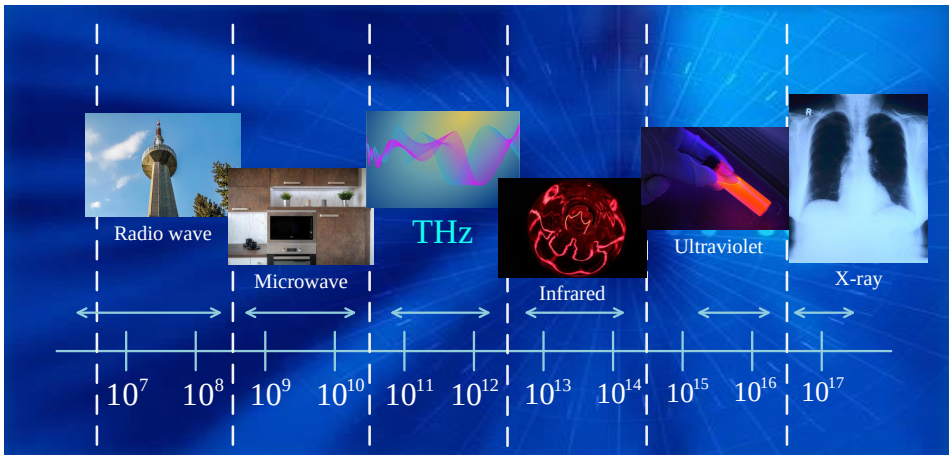


Fig. 6. The applications of electromagnetic waves in different frequency bands

A. Terahertz Communication Technology

As shown in Figure 6, terahertz waves have a frequency between 0.1 and 10 THz. In terms of frequency, terahertz waves are located between millimeter waves and infrared waves. In terms of energy, it is located between electrons and photons. The high frequency of terahertz waves can solve the problem of large-scale data information limited by bandwidth, and it can meet the future requirements of mass data transmission as well as high transmission rate communication. Terahertz waves are characterized by abundant spectrum resources, strong anti-interference capability, high transmission rate, and high penetration. Therefore, it can effectively complement today's null-port transmissions [31,32].

B. Air-space-ground-sea Integrated Communication Technology

Air-space-ground-sea integration means that the 6G network will cover the sky, space, ground as well as the sea, connecting all parts of the network to form an integrated communication network, which is shown in Figure 7. Air-space-ground-sea integrated communication technology takes ground-based networks as the basis, while space-based, air-based, and sea-based networks are considered as expansion and extension. The ground-based networks are composed of local area networks and cellular mobile networks, and the network service area in this area is relatively dense. The space-based networks are composed of various satellite networks, which can be classified into low-orbit satellites, medium-orbit satellites and high-orbit satellites. The air-based networks are composed of



Fig. 7. Air-space-ground-sea integrated communication

planes, airships, drones, and other flying communication equipment networks. The sea-based networks are composed of ocean underwater communication [33], ship communication and island communication networks, etc. [34]. We can connect the communication networks in each area of the integrated air-space-sea network and fully use the various resources in them to meet a wide range of requirements in the future [35].

C. Visible Light Communication (VLC) Technology

VLC is a communication method that uses light in the visible wavelength band as a message carrier to transmit optical signals directly in the air. The number and density of communication devices will multiply in the future. If traditional radio communication continues to be used, this will lead to a shortage of communication bands. And because of its unique information carrier, visible light communication technology provides a new choice for short-range wireless communication [36]. The applications of VLC are shown in Figure 8:

The RF signal of traditional radio communication will not only cause harm to the human body, but also cause interference to other devices. At the same time, visible light transmission will not produce radiation to the human body and will not be affected by external electromagnetic interference. It has a more enormous scope of application than radio communication [36]. From the perspective of environmental protection, visible light communication is green and low-carbon. Then, from the aspect of privacy protection, radio electromagnetic signals will have the problem of electromagnetic leakage, which will not be conducive to privacy protection. But visible light communication rarely appears to be eavesdropped and illegally appropriated.

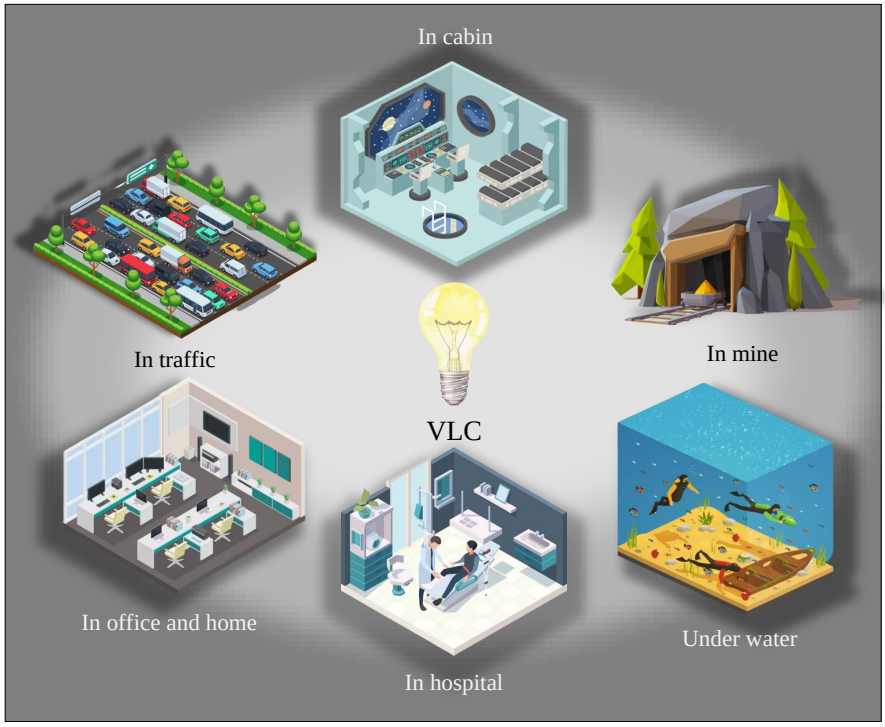


Fig. 8. Applications of the visible light communication technology in different scenarios

From the exposition in this section, we can see that people have great confidence and motivation in developing 6G communication technology, and all countries strongly support the research of 6G technology and invest a lot of time and money to promote the development of 6G. 6G has a vast improvement compared with previous generations of mobile communication technology. With the application of 6G technology in many fields, it will bring great convenience to people and change the face of society. However, some security issues in 6G networks will also arise. In the following parts of this article, we will analyze some security issues in 6G networks.

3. Security Challenges of 6G Networks in Architecture, Visions and Applications

After the introduction of the development of mobile communication networks and 6G networks in the previous section, this section will discuss some security issues existing in 6G networks. Here, we mainly introduce the security threats of 6G in network architecture, visions, and related applications, and we also list some related countermeasure strategies.

3.1. Security Issues in Architecture

The network architecture of 6G can be divided into network access, network domain, user domain, and application domain. This section will summarize the security issues in the 6G architecture and possible solutions, as shown in Table 1.

Table 1. Security issues and solutions in 6G architecture

Security Issues	New Requirements	Possible Solution	References
Network Access Security	• New authentication model and password system	• 6G-AKA • Quantum security codes • Physical layer security • A design of SD-WAN-oriented wide area	• Wang *et al.* [37] • Liang *et al.* [38] • Sun *et al.* [39] • Tanveer *et al.* [40] • Wang *et al.* [41] • Mine *et al.* [42]
Network Domain Security	• New open authentication method • Authentication methods during network domain switching	• Lightweight encryption methods • New authentication technology • Blockchain technology	• Khashan *et al.* [43] • Alresheedi *et al.* [44] • Arfaoui *et al.* [45] • Yazdinejad *et al.* [46]
User Domain Security	• New authentication method • New user experience	• Biometric-based authentication technology, smart cards, etc. • Multi-party information sharing and cooperation • Blockchain technology	• Liu *et al.* [47] • Liang *et al.* [48] • Sengupta *et al.* [49] • Cao *et al.* [50] • Lu *et al.* [51] • Long *et al.* [52]
Application Domain Security	• Solutions for different security needs in different areas	• Customized security solutions • Network security regulation and management • Data encryption and privacy protection	• Abdel *et al.* [53] • Akbar *et al.* [54] • Ramezanpour *et al.* [55] • Khan *et al.* [56] • Zhang *et al.* [57]

Network Access Security The 6G network will access a large number of devices, including IoT devices, sensors, smart terminals, and so on. The security of these devices is one of the core issues when 6G network devices access the network. Devices may have security vulnerabilities, such as hardware vulnerabilities, software vulnerabilities, communication protocol vulnerabilities, etc., which can be exploited by attackers to gain control of the device, steal data, or perform malicious operations. For example, the era of 6G networks will usher in immersive cloud XR, and multisensory XR applications will apply molecular communication technologies, THz technologies, and quantum communication technologies [41]. However, combining XR with uRLLC and eMBB has many outstanding security issues, such as malicious behavior, access control, internal communication,

and so on. Therefore, 6G networks require new authentication models and cryptographic systems, such as 6G-AKA and quantum security cryptography.

6G network design will gradually shift to a cloud-based and open programmable network platform, leading to changes in the authentication architecture. 6G networks may inherit the security model of 5G networks, such as providing a unified authentication platform for open and access network irrelevant individuals and flexible scheduling and dispatching of resources through network slicing mechanisms. However, the inherited security model is far from solving the network access security problems that 6G networks will face. Consequently, some solutions need to be adopted, such as the 6G-AKA protocol proposed by Nguyen *et al.* in [58]; this method needs to prove which component will determine authentication in cross-slice communication. 6G-AKA needs to be able to verify endpoint identity even in deeply sliced and open programmable network platforms.

Network Domain Security The 6G air-space-ground-sea integrated network architecture will be based on terrestrial cellular mobile networks. It can combine the broad coverage, flexible deployment, and efficient broadcasting features of broadband satellite communications and achieve full coverage of sea, land, space, and air through the deep integration of multiple heterogeneous networks, which will bring new opportunities for the markets of marine, airborne, cross-border and the fusion of heaven and earth. As the 6G networks will cover various network domains, security issues must be considered when switching between different network domains. To solve this problem, authentication methods can be designed during the switching process to ensure security. It is also important to note that the 6G coverage network domain includes non-terrestrial networks such as satellite and maritime communications, which will require new open authentication methods.

User Domain Security Nowadays, authentication is valued, and various researches have been carried out in this direction [59,60]. Nevertheless, most applications have been using password-based authentication, which not only makes it easy for hackers to crack passwords but also it is not excluded that users may forget their passwords, which does not guarantee user account security, and users will have a poor experience. 6G networks can use biometric-based authentication encryption methods [61] or provide password-free services, such as retina-based or brainwave-based authentication [47]-[52]. It may provide users with a more secure and better user experience in the future.

Application Domain Security 6G is expected to be commercially available in 2030, when it will become the key communication infrastructure for us to meet many requirements. People will apply 6G to Unmanned Aerial Vehicle(UAV), autonomous driving, Smart Grid 2.0, Industrial 5.0, super smart healthcare, digital twin, extended reality, collaborative robotics and many other areas. The application domain of 6G communication networks will continue to expand.

However, as 6G communication technology is applied to more and more fields, some security issues will continue to emerge. For example, different application domains have different usage scenarios and different levels of security requirements. Moreover, many domains are new, and people have little research on them. As a result, safety will be difficult to guarantee. People need to take corresponding solutions for different security

requirements in different domains. For instance, in identity verification, even though symmetric key mutual authentication is still used in 5G, it may benefit more from blockchain technology and distributed ledger technology in a 6G network environment. This means that there is still a great deal of work to be done in application domain security for 6G networks.

3.2. Security Issues in Visions

The three significant visions of 5G networks are uRLLC, eMBB, and mMTC. In contrast, the visions in 6G networks have further enhancements based on the visions of 5G networks, which are enhanced Ultra-Reliable Low-Latency Communication (eURLLC), Further enhanced Mobile Broadband (FeMBB), and ultra massive Machine-Type Communications (umMTC) [62]. In this section, some security issues of 6G networks in these three visions are analyzed. We summarize the contents, as shown in Table 2.

eURLLC The goal of uRLLC is to enable network characteristics and functions to manifest under extreme conditions [62]. To keep round-trip latency between critical infrastructure and computers within 1ms, techniques such as network slicing, beam forming, and packet re-transmission protocols are used to improve reliability and reduce latency. And the network reliability will be required to exceed 99.999%. The eURLLC proposed by 3GPP introduces enhanced ultra-reliable, low-latency communications, and the network's reliability will increase to 99.99999% [72]. Service applications for eURLLC include latency-sensitive services that require high reliability, such as autonomous driving, tactile Internet, telemedicine, and industrial automation. The end-to-end packet transmission delay for mission-critical applications should be between 10-100ms.

The security threats faced in eURLLC are as follows:

1. Intrinsic Openness of the Wireless Medium

   eURLLC should consider the impact of security workflow delays to ensure the quality of service [63]. At the same time, the ultra-high reliability of 6G networks will imply the need for more efficient security protection schemes to meet the high requirements they impose, so as to protect the availability of services and resources. As also mentioned in the literature [65], the security issues of eURLLC are raised due to the inherent openness of the wireless medium. Traditional cryptography-based approaches are able to solve the security problem of high-level information transmission under the assumption that the eavesdroppers (Eves) have limited computational ability. However, once the above assumptions fail, traditional cryptography will not be able to solve the security problem. Therefore, more sophisticated and innovative encryption methods are needed to solve the security problems of information transmission in eURLLC.
2. Multi-level Architecture Security

   Many agents and virtual spaces are generated with the continuous innovation of information technology [66]. Object communication (OCC) exhibits the multi-flow concurrency characteristics of multi-dimensional, heterogeneous, and multi Quality of Service (QoS) requirements. Data computing transitions from cloud computing

Table 2. Comparisons of 5G, 6G visions and security issues in 6G application scenarios

5G	6G	Differences	Security Issues	Possible Solutions	Applications
uRLLC	eURLLC	• Lower latency • Higher reliability • More device connections are supported • More technologies are adopted	• Intrinsic openness of wireless media [63,64] • Failure of encryption methods [65] • Multi-level architecture security [66] • Malicious software	• Secure network architecture design • Secure communication protocols • Security authentication and authorization • Safety monitoring and management	• UAV-based Mobility • IoT Healthcare • Drone-based systems
eMBB	FeMBB	• Higher bandwidth • Lower latency • Greater throughput • Faster transmission speed • Wider network coverage • Greater network capacity	• Flow monitoring [67] • User privacy [68] • DoS and resource attacks • Crack of the encryption algorithm	• Data encryption • Authentication and access control • Installation of firewall and intrusion detection system • Establishing an effective security monitoring and management mechanism • Lightweight authentication • Techniques for keys management	• UAV-based Mobility • Holographic Telepresence • Super-Fast Hotspot • IoT Healthcare • Industrial Internet of Smart Things • Ultra-dense cellular IoT networks
mMTC	umMTC	• More device connections are supported • Application to larger scale IoT communications	• Fake terminals [69] • Remote attack [70] • Device forgery and illegal access • Hacking and malware intrusion • Data leakage • Data tampering	• MTC endpoint embedded device security [71] • Implement authentication and data encryption for devices and users in umMTC • Risk assessment and management • Enhance physical equipment security management	• Holographic Telepresence • Super-Fast Hotspot • IoT supply chain • IoT Healthcare • Drone-based systems

to edge computing, then gradually convert from edge computing to distributed computing and heterogeneous computing. Therefore, the security of centralized and distributed multi-level architectures will also become a factor affecting the security of eURLLC requirements for 6G networks.

FeMBB eMBB is one of the three main 3G New Radio (NR) use cases defined by 5GPP. eMBB is defined by Dakhaz *et al.* in the literature [73] as follows: eMBB communication is intended to meet high throughput requirements and is used for several high-priority services. eMBB is proposed to meet the latest applications due to the rapidly increasing use of broadband data. In the era of 6G networks, a range of immersive applications, including 3D extended reality capabilities, 3D multimedia, and IoE are predicted to emerge [74], which will need to be achieved with high-quality services at tens of times the peak Gbps rate. Therefore, the 6G network mobile broadband speed must be further improved to provide a Tbps level mobile broadband data rate peak to meet the higher requirements, thus proposing FeMBB. Business applications for FeMBB include high-resolution video streaming and expanded reality with large data packets. This creates high bandwidth requirements for 6G networks.

The security threats of 6G networks in FeMBB are shown in the following points:

1. Flow Monitoring

   In the future, FeMBB will be applied to 4K/8K HD video and mobile roaming immersive services based on Virtual Reality (VR) and Augmented reality (AR), generating a large amount of data consumption. Existing network firewalls and security devices such as intrusion detection systems will not be able to ensure adequate security protection when large amounts of data monitoring and storage are involved [67].
2. User Privacy

   Extreme data rates in FeMBB will pose challenges for secure traffic processing, such as attack detection, AI/ML pipelines, traffic analysis, and universal encryption [63]. In FeMBB, a large amount of user data needs to be processed, including identity information, location information, communication information, and so on. Therefore, if the work related to secure traffic processing is not done correctly, then it can cause serious privacy leakage problems [68].

umMTC One of the key communication scenarios envisaged to be supported by 6G is the umMTC [75]. Its main feature is the large number of low-power devices to be connected, with up to 1 million sensors per square kilometer. IoT devices and their connectivity in the IoE world will dramatically change the scale of machine-type communication in the IoE revolution [76]. In IoE architectures, there may be a trillion brakes and sensors automatically transferring data back and forth. With such massive machine communication, current machine-type communication architectures will not be able to meet effective and efficient connectivity, thus requiring umMTC architectures to support reliable connectivity for large-scale networks.

The security issues in umMTC are as follows:

1. Fake Terminals

Since the IoT terminal is weak in resource processing and computation, this leads to a situation where authentication may not be achievable. Even if authentication is possible, the restricted computing power can only be used in simple authentication methods that are not effective. This situation leads to forged, counterfeit terminals that cannot be identified, which in turn creates security problems [69].

2. Remote Attack

   Intruders may take the shortcomings of the simplicity and weak protection of IoT terminals and then remotely attack the IoT terminals through software and hardware interfaces and then use the compromised terminals to launch cyber attacks [70].

3.3. Security Issues in Applications

6G networks will be used in the future in many fields, such as healthcare, transportation [77,78], logistics [79], and production, but applications in every field will face some security issues [63]. For the application of technology, we have selected UAV technology, which is the key technology to realize the concept of the air-space-ground-sea integration of the 6G network. As well as digital twin technology, a popular technology for applications in the future society. We have also selected people's visions of Smart Grid 2.0 and Healthcare 4.0 arising from the future 6G era as application examples. As well as Industry 5.0, an upcoming application domain, is also presented in this section. Figure 9 summarizes the security issues in applying 6G networks in different fields.

UAV based Mobility UAVs are mainly used in 6G networks for communication at the air-based network level. UAVs can be deployed in various scenarios, and the essential communication of ground-based networks can be expanded to air-based networks and connected with the satellite equipment of space-based networks and the communication equipment of sea-based networks, thus realizing the whole area coverage of 6G networks. In [80], Shrestha *et al.* proposed an advanced and fully autonomous 6G-based UTM system by applying UAVs.

Some security issues will also arise in the autonomous UAV system. For example, an attacker could gain control of a UAV by jamming the control signals or capturing it physically and then hacking inside it to obtain critical data. An attacker could also combine UAVs with other smart devices and thus convert from a cyber attack to a physical attack, which could lead to even more damage. Hooper *et al.* [81] predicted that an attacker could also gain remote control of a UAV by hacking into it through WiFi.

Smart Grid 2.0 The smart grid is the intellectualized grid, also known as "Smart Grid 2.0", based on integrated, high-speed two-way communication networks. Advanced sensing and measurement technologies, equipment technologies, control methods and decision support system technologies are applied to make the grid reliable, secure, economical, efficient, environmental-friendly and safe to use [82]. Its main features include (1) self-healing, (2) stimulating and protecting customers, (3) resisting attacks, (4) providing power quality that meets customer needs, (5) allowing access to different forms of generation, (6) initiating power markets, (7) optimizing the efficient operation of assets [83].

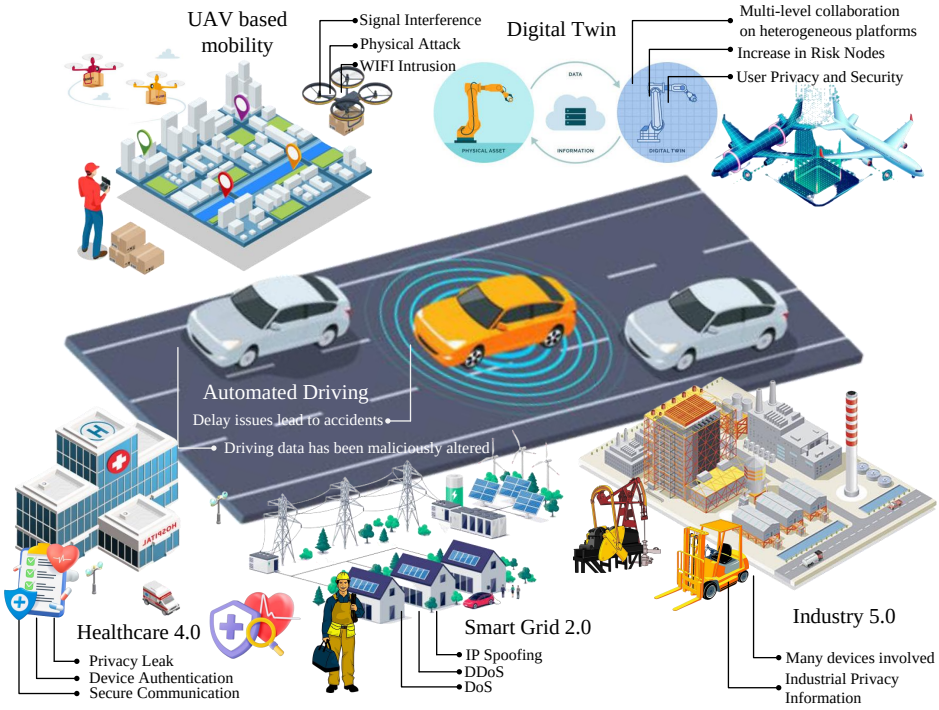


Fig. 9. Summary of security issues of 6G network applications in different fields

Smart grids improve network reliability by combining advanced sensing, communication, and control infrastructure. In the context of Smart Grid 2.0, the performance of the grid is affected by cyber-attacks and physical disturbances (e.g., symmetric and asymmetric faults), which can lead to power system instability. An attacker may analyze traffic data to gather helpful information or eavesdrop on smart grid entities(i.e., smart meters and control centers) [84]. Intruders or eavesdroppers can impact the security of smart grid network entities through component, DoS, DDoS, IP spoofing, and MitM attacks. From a cybersecurity perspective, a reliable and secure information system is required for the effective operation of a smart grid [85]. Especially when the smart grid is subjected to cyber-attacks, its physical layer performance is also affected, which may lead to decreased system reliability. In addition, improvements in trust management of transaction mechanisms are also critical to maintaining the security of the smart grid.

Healthcare 4.0 In the era of 6G networks, many smart medical devices applying communication technologies, sensors, vehicle technologies, and artificial intelligence will be deployed into the medical industry in large numbers [86]. 6G networks will facilitate the improvement of reliable remote monitoring systems in medical systems so that remote surgery, diagnosis, monitoring, and emergency care will become a reality in the 6G era. The 6G networks with high-speed rates, low latency, and ultra-reliable features will fa-

cilitate the fast and efficient transmission of data between many medical devices, thus improving the efficiency and quality of medical services.

Smart medical devices will collect large amounts of data, such as Body Area Networks (BANs) equipped with intelligent embedded systems to facilitate personalized management and health detection [87], and these BANs can collect human health data from various parts of sensors, dynamically share and interact with network services. 6G is likely to become the central communication platform for smart medical platforms, and there will be a large amount of private information involving patients transmitted in the network, which will have unimaginable consequences in case of security issues leading to information leakage. Healthcare monitoring frameworks are also increasing in their requirements for network performance as well as privacy-preserving means [88]. The vulnerable healthcare supply chain in networks is also risky [89]. This is because it may reduce the quality of some healthcare services, thereby threatening the privacy and security of patients as well as test results. In summary, device authentication, secure communication and access control for billions of micro medical devices will be the issues that need to be addressed for 6G network security.

Industry 5.0 Since mid-2010, the world has been in the fourth industrial revolution, also known as ”Industry 4.0”. The World Economic Forum defines this revolution as a combination of AI, advanced robotics, 3D printing, and IoT [90]. Recently, the automation level in the manufacturing industry has increased significantly [91], various smart devices are gradually appearing, and the way machines are connected by the Internet is fundamentally transformed. Some people describe this shift as ”Industry 5.0”, and the collaboration between humans, robots, and smart technologies will become the next industrial revolution breakthrough in the ”Industry 5.0” era [92].

The ”Industry 5.0” era involves many industrial privacy issues, such as the intellectual property rights of industries, production information of factories and the identity of employees. In an industrial Internet system, a company can collect raw data from its devices and sensors. These raw data allow the company to deliver scalable and reliable applications faster to meet the changing needs of the customers. Nevertheless, this raw information may contain the kind of private information mentioned above, and if this information is hacked illegally, it can lead to a serious privacy leakage problem. Therefore, in the ”Industry 5.0” era, applications must meet basic security requirements such as integrity, availability, authentication, and auditing.

Digital Twins The coming of the 6G era heralds the coming of the digital twin era. With the development of communication and artificial intelligence technologies, objects and processes will be able to produce a replica in digital form, and intelligent mapping of human-human and object-object interactions in digital environments will occur [93]. Through various complex algorithmic models, people can predict, plan and simulate the real world through the digital world. In this way, the optimal solution to solve some practical problems is thus obtained. In the field of industrial manufacturing, the digital twin can virtually construct a digital model of a product to simulate it for testing and verification. In the healthcare assurance industry, the same type of sensor data system can be used to track various health metrics and generate significant insight results [94].

Currently, technologies such as artificial intelligence, big data, and industrial Internet are developing rapidly. The connection between the virtual space of intelligent manufacturing based on the digital twin and the physical world is built on the basis of network information flow transmission. With the accelerated integration of digital twins and smart manufacturing, there will be a great possibility of transforming from a closed system to an open system in the future, and systemic security issues will be concentrated. With the integration of smart manufacturing with big data and cloud computing, as well as the deep involvement of third-party collaboration services and multi-level collaboration of a large number of heterogeneous platforms, the number of network security risk nodes will rapidly increase. It also brings more ways of intrusion, and many security issues are required to be solved [95].

4. 6G Key Technology and Security Analysis

There are many key technologies in 6G communication networks, the classic technologies are selected, such as THz communication, VLC, and blockchain. In this section, we analyze some security issues of the application of these key technologies in 6G and list some countermeasure strategies as references.

4.1. Terahertz Communication Technology

In recent years, the communications industry has developed rapidly. Due to the large number of people using wireless TV, mobile communication, amplitude modulation (AM), FM radio (FM), and other communication devices, the RF band has been almost wholly occupied. It is challenging to meet the massive demand for 6G networks, people need to explore higher-frequency electromagnetic waves to meet the new demand. Therefore, people focus on the THz electromagnetic waves. Terahertz, which ranges from 0.1 THz to 10 THz, will provide more bandwidth, greater capacity, and higher data transmission rates for 6G networks [96,97].

Security Threats Although the use of terahertz electromagnetic waves will facilitate the rapid development of 6G networks, there are many security issues in their application. For example, the high directionality of transmissions in narrow beams makes them robust to interception attacks, but there is still the possibility that malicious nodes can successfully intercept the signals. As a result, they can compromise access control, steal user information and essential data, and gain access to unauthorized resources. Singh *et al.* [98] confirmed that an unauthorized attacker could maliciously place some obstacles in the transmission path and thus make the attacker capture the communication through radiation scattering. The principle of this method is schematically shown in Figure 10. The attacker can also set the target of the jamming beam as the receiver of the signal and then jam the signal by electronic methods.

Possible solutions As 6G networks will use a higher frequency range, fine-grained frequency management, and coordination will be required to ensure that signals from different users do not interfere with each other. This can be achieved through techniques such as

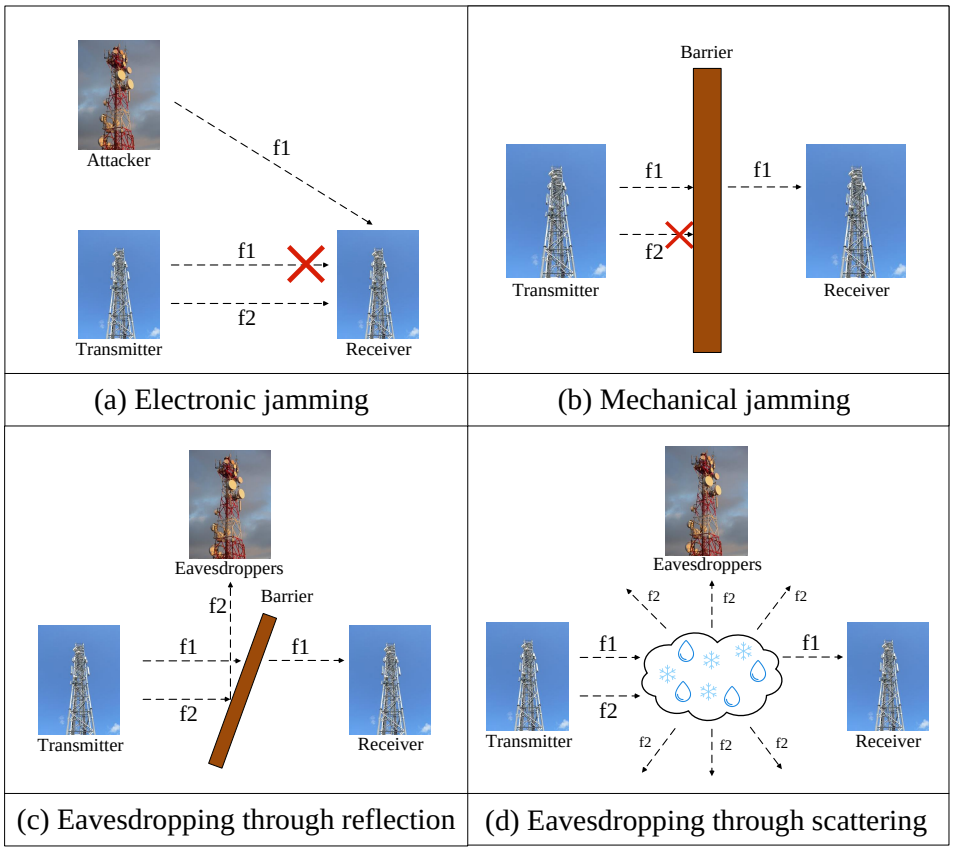


Fig. 10. Schematic of the principle of signal capture by radiation scattering

dynamic frequency allocation, coordinated spectrum sharing, and interference cancellation. Interference between different users can also be reduced by spatially isolating them through the use of massive antenna techniques. At the same time, beamforming techniques can concentrate signal energy in specific directions, improving signal quality and reducing interference to other users. Furthermore, Ning *et al.* [99] propose to maximize the secrecy rate by designing active beamformers at the base station (BS) and deploying passive reflective phase shifters (PSs) at the Intelligent Reflective Surface (IRS). To cope with the mixed integer nonconvex optimization problem involved, the authors propose a joint design with high-security performance and a low computational complexity design in the article. Apart from this, Li *et al.* applied IRS to wireless secure communication to study how to make the secure rate reach the optimal secure capacity from the perspective of optimization technique [100]. Petrov *et al.* in the literature [101] propose to exploit the inherent multi-path property of terahertz communication to reduce the probability of messages being eavesdropped. The authors suggest the ability to share data between communicating entities over the multiple terahertz propagation paths that are available. In this

way, even if eavesdroppers cooperate, this could have a significant effect in addressing the risk of eavesdropping. But at the same time, it will lead to a reduction in link capacity.

4.2. Visible Light Communication

VLC is a communication method that uses light in the visible wavelength band as an information carrier to transmit optical signals directly in the air [102]. VLC is considered a technology that supports 6G networks because of its huge bandwidth and terahertz frequency. The concepts related to VLC have been mentioned earlier in Section 2.2.3 and will not be elaborated on here. Possible security threats and their countermeasures are shown below:

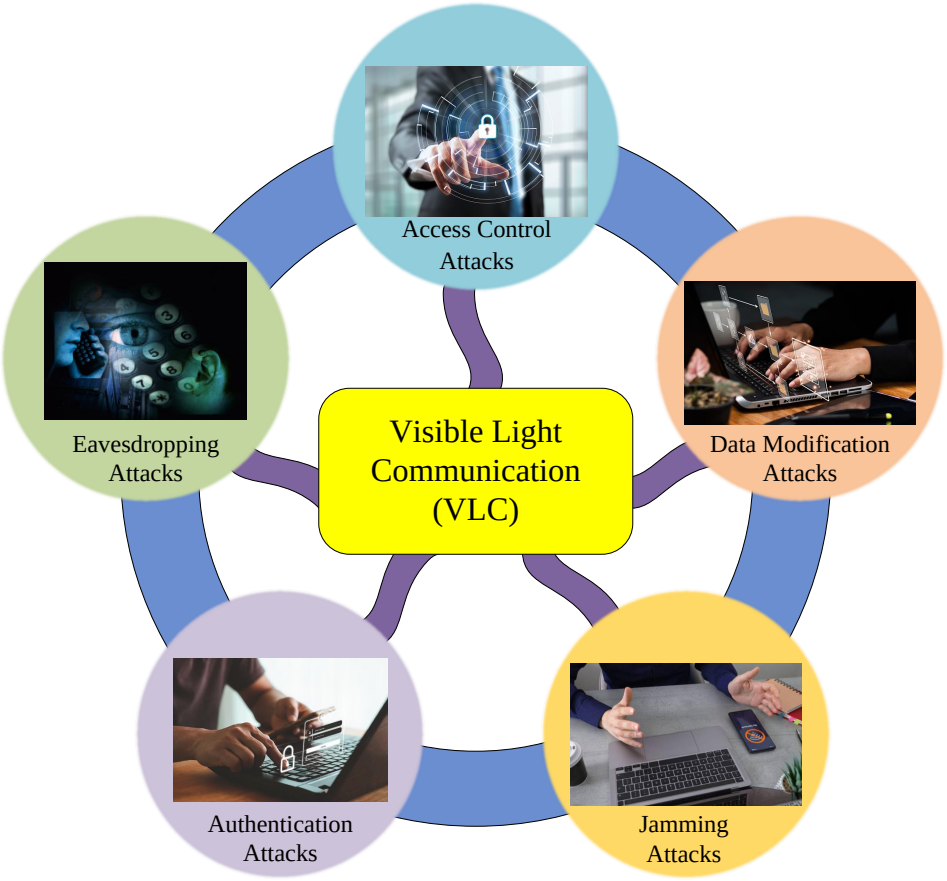


Fig. 11. Security issues in VLC

Security Threats Due to the open and broadcast characteristics of VLC, when the nodes of a VLC system are deployed in areas with large coverage, it may result in the overlap-

ping of signals from different transmitters, which makes VLC systems potentially exposed to some security issues [103]. At the physical level, attackers may attack the physical layer by eavesdropping, jamming, and capturing signals. Other access control attacks are carried out by illegally gaining access to the wireless medium. The security threats to the application of VLC technology in 6G networks are shown in Figure 11.

Possible solutions To address the security problems of VLC technology in 6G, it is necessary to strengthen technological innovation and establish a mature security system. For example, coding and modulation techniques with higher security can be developed to improve the reliability of data transmission. In [104], Soderi *et al.* propose a Watermark-Blind Physical Layer Security algorithm (WBPLSec) that combines watermarking and interference primitives to protect visible optical communications at the physical layer. They also use Intelligent Supersurface (RIS) techniques to improve the security performance of communications. In the literature [103], the authors enhance the confidentiality performance of MIMO VLC systems by linear pre-coding. Chen *et al.* proposed a visible light communication system of LiFi (light fidelity) network in the literature [105], which can provide high-speed broadband services to users with multiple addresses and mobile support. The authors also propose a transmitter and receiver design with omnidirectional characteristics in order to improve the channel quality and robustness of the uplink in LiFi systems.

4.3. Blockchain

Blockchain is a naturally distributed trust mechanism that is maintained by multiple parties collaboratively. It has the characteristics of openness and transparency, immutability, full trace, intelligent execution, etc. It can provide effective ideas for operation management, security trustworthiness enhancement and multi-party trust model of 6G networks [106]. The nature of blockchain distributed bookkeeping and its immutability provides a fair and open operating environment for every operator involved in the shared construction of the 6G network [107]. The smart contract in blockchain has the characteristics of value transfer, resource share, and enabling intelligent settlement [108,109]. Therefore, blockchain is applicable to resource sharing and settlement in the integrated air-space-ground-sea network [110], which makes the sharing of the infrastructure resources and the process of value transfer in the integrated network transparent, reliable, and real [111]. Blockchain can also be applied in the Internet of Vehicles (IoV) to prevent malicious participants from modifying the transmitted information and ensure the security of in-vehicle information transmission [112].

Security Threats The application of blockchain is exposed to a number of risks, and if it is deployed in a 6G network, these security threats will be manifested in the 6G network, as shown in Figure 12 [113]:

- 51% attack: Blockchain can build trust without the involvement of third-party institutions, when malicious people occupy 51% or more of the blockchain nodes, attackers may succeed in taking control of the network and thus tamper with information or affect transactions [114].

- Scalability: Blockchain-enabled systems present a number of challenges in terms of scalability in terms of throughput and storage [115,116]. Throughput depends on various factors such as hardware capacity, network latency, size of the P2P blockchain network, and others. The 6G network will have a very high data rate and connection density, coupled with the fact that many heterogeneous devices in the network will exist as nodes in the blockchain, which has a very high requirement for the throughput and storage capacity of the blockchain network.
- Privacy leakage: Since blockchain relies on transparent transactions to a certain extent, the process may cause problems such as the leakage of smart contract logic, transaction information [109], and user privacy during the execution of smart contracts.
- Decentralization: Blockchain is a decentralized network, and for untrusted nodes, blockchain needs to provide them with decentralized means of trust to prevent them from gaining more significant influence [117,118]. Whereas in a 6G network, there may be a large number of nodes with fake identities, which pose a security risk to the network as untrusted nodes. This poses a great challenge to the blockchain, and if these nodes are not handled well, the decentralization of the blockchain network will be compromised and will pose certain security threats.
- Consensus Algorithms: Consensus algorithms play a vital role in the operation of blockchain systems [119], and different types of consensus algorithms are developed and used for different applications [120]. With the development of quantum computing, there is a high possibility that 6G networks will also be integrated with quantum computing in the future. The current consensus algorithms existing in encryption mechanisms [121,122] are easily cracked by quantum computing, so we must propose new consensus algorithms encryption mechanisms for emerging technologies.

Possible solutions In order to cope with the above security issues of blockchain applications in 6G networks. We can adopt the techniques of sharding as well as a Directed Acyclic Graph (DAG) to cope with the high throughput requirements. Increase flexibility in network management and operation by decoupling the data plane and control plane in Software-Defined Networking (SDN) [123]. Quantify the degree of decentralization of blockchain by entropy method based on information theory. It is also vital to design cryptographic mechanisms with consensus algorithms against various attacks. New smart contract mechanisms can also be designed to protect private information on the blockchain.

A number of articles have proposed solutions to the security problems associated with the application of blockchain in 6G networks. In literature [124], Manogaran *et al.* introduce a blockchain-based integrated security measure (BISM) to provide secure access control and privacy protection to users. In this method, the access control process depends on the state of virtualized resources at different time instances, and the privacy protection depends on the corresponding duration of the service. Wang *et al.* propose a unified framework of Blockchain Radio Access Network (B-RAN) to improve the efficiency and security of 6G networks using blockchain technology [125]. Shahzad *et al.* [126] makes two contributions to address the security issues of blockchain applications in 6G networks. Firstly, at the process level, the authors propose a new smart contract mechanism. The specific process is done by embedding a hashed image into a message and then dividing the image into small blocks of pixels, all of which can be shuffled by

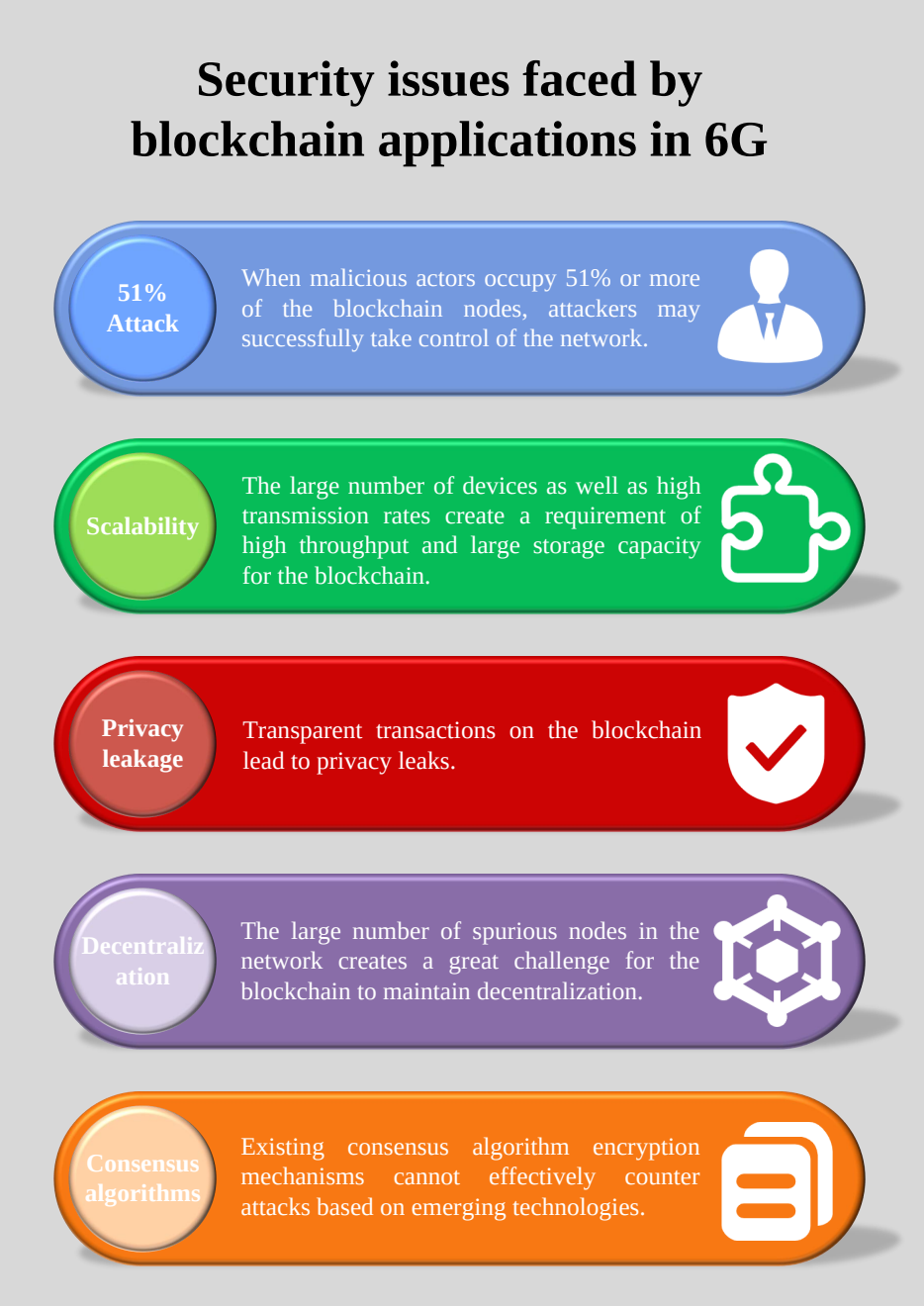


Fig. 12. Security issues in the application of blockchain technology in 6G networks

any matrix or algorithm to create a hash of that image. The receiver then restores the image to its initial state by means of the same matrix or algorithm. In this way, the record is guaranteed not to be tampered with. At the data level, the authors use a digital signature method that allows anonymization to verify and protect the blockchain without encrypting it. Abdulqadder *et al.* [127] has proposed a context-based authentication and security switching scheme using Markov Decision Method (MDM) and weighted product models to address the security issues that exist in network slicing in 6G environments. In terms of identity verification, the authors propose a context-based multi-factor biometric authentication (CMFBO) algorithm. A large number of IoT devices in 6G networks are working on insecure networks. However, the existing centralized signature schemes cannot address the challenges of security and efficiency in IoT identification. Liu *et al.* proposed a decentralized multi-signature protocol, IdenMultiSig [128], that combines Identity-Based Signature (IBS) with a Schnorr scheme under discrete logarithms on elliptic curves.

5. AI encounters 6G networks

6G networks will have greater flexibility and interoperability than 5G networks, intelligent network orchestration will be part of 6G networks, and 6G networks will move beyond the new Open Radio Access Network (O-RAN) [129]. O-RAN networks enable network flexibility by supporting techniques such as deep learning, machine learning, and reinforcement learning in AI. Applying AI algorithms to communication technologies is a way to improve spectrum utilization efficiency, reduce communication latency, and enhance security. Thus, the 6G paradigm is viewed as AI-based trust management, which is a promising one capable of delivering trusted and reliable services.

5.1. AI deployment in 6G networks is a double-edged sword

By applying AI to 6G networks, AI can help 6G networks build a flexible, intelligent, and dynamic network architecture. It assists 6G networks in building a secure and stable network environment regarding threat detection, resource allocation, and attack defense. However, while AI benefits 6G networks, some security vulnerabilities in AI will also be reflected in 6G networks with the deployment of AI in 6G networks. Therefore, AI will also bring some security threats to 6G networks, so it is said that the deployment of AI in 6G networks is a double-edged sword [130]. In this section, we will illustrate the application of AI in 6G networks from the aspects of AI facilitating network construction and the security risks brought by AI.

AI facilitates 6G network construction AI can provide a lot of convenience for the construction of 6G networks, which will help us quickly deploy 6G networks and put them into commercial use. AI can assist 6G networks in dynamically allocating network resources, so that limited resources can be reasonably allocated to different network devices. At the same time, AI can also provide personalized services according to user needs, providing users with a better experience. At the same time, AI can also realize the concept of distributed independence, autonomous management, and green communication in the 6G networks and provide a more secure and reliable authentication method for the 6G

networks. We list the contributions made by AI in driving 6G network construction and summarize some related papers in Table 3.

1. Dynamic management

   AI is expected to provide deep-learning-based security function virtualization (SFV) for 6G networks, so as to support the dynamic defense mechanism of Software Defined Security (SDS) architecture using SFV [131]. The mechanism enables intelligent monitoring of network traffic conditions at different network endpoints and segments, timely detection of anomalies and dynamic countermeasures to build an intelligent network management system.
2. Personalized service

   AI can also leverage heterogeneous network resources to provide intelligent personal customization to users by collecting their service needs and preferences. Yang *et al.* [132] proposed the concept of a Service Requirement Zone (SRZ), which is deployed at the users' end to describe and visualize the integrated service requirements and preferences for a single user's specific task. On the system side, the concept of User Satisfaction Rating (USR) is introduced to evaluate the overall service capability of the system to accomplish tasks in various SRZs [133]. We have also proposed a network AI architecture with integrated network resources and leveraging AI capabilities to guarantee the Quality of Experience (QoE) for each user based on the above. It also effectively meets diverse user needs and provides them with customized services.
3. Realization of a distributed independent autonomous system

   Deploying AI's intelligent agents to Base Stations (BSs), makes it possible for AI technology to not only optimize the network but also provide service and application intelligence to users. This means that AI will change the standards for future 6G networks. For example, future 6G networks will need to have self-aggregating capabilities for functional, situational, and location networks. Enhanced context-awareness capabilities, as well as contextual self-reconfiguration of networks and nodes, are also needed [134]. These make AI the driving force of 6G networks, enabling the proliferation of large-scale fog computing clusters driven by distributed independent autonomous systems and related common goals.
4. Authentication

   AI can automatically combine multiple non-cryptographic attributes related to users, resources and environments for determining the identity of a given entity and authorizing it after confirming successful authentication. In this way, continuous authentication and dynamic enforcement of fine-grained access policies are provided. Peng *et al.* in [135] proposed a framework for suspicious person re-recognition based on end-to-end edge devices in real public scenes. Firstly, the image information intercepted from the video is fed into the detector to obtain the pedestrian location information, and then the pedestrian attributes are recognized and re-recognized using Omni-Scale Network (OSNet). Generalized learning systems (BLSs) as well as Cycle-consistent Adversarial Networks (CycleGAN) are also utilized to remove the noisy data and unify the data style. This improves the performance of the identification model.
5. Green communications

   In the future 6G era, the number of basic communication equipment will increase dramatically, and the communication between equipment and the maintenance

Table 3. Summary of papers related to 6G network construction driven by AI

Authors&Years	Ref	Assistance in the management	Enhance network security	Green communications	Improve network architecture	Optimize user experience	Remarks
Rahman *et al.* 2022	[131]	H	M	×	L	×	By studying the security challenges posed by the convergence of operational and information technologies for 6G networks, they proposed a distributed DL-assisted SDS for 6G vertical networks, which will autonomously detect, localize and isolate security threats through SFV.
Yang *et al.* 2022	[132]	M	L	×	H	H	The authors propose SRZ to collect user requirements and preferences and introduce USR to assess the system service capability strength. Finally, a network AI architecture with the ability to integrate network resources and the capability to popularize AI is proposed.
Stoica *et al.* 2019	[134]	H	L	×	M	×	They propose to deploy intelligent agents of AI to base stations, which makes it possible for AI technology to not only optimize the network but also provide service and application intelligence to users.
Peng *et al.* 2023	[135]	×	H	×	×	L	They propose an end-to-end edge device suspicious person re-identification framework based on real public scenarios.
Mao *et al.* 2021	[136]	H	L	H	L	×	They propose to combine DL algorithms or ML algorithms for AI with traditional AI algorithms and mathematical models to reduce the complexity of the algorithms and increase the efficiency of network resource management.
Wu *et al.* 2022	[137]	H	L	×	H	×	This paper presents an AI-native network slicing architecture for 6G networks to enable synergy between AI and network slicing to facilitate intelligent network management and support emerging AI services.
Shafin *et al.* 2020	[138]	H	×	L	M	×	In this work, the authors present a vision of AI cellular networks for 6G networks, which reduces the operational costs as well as the overhead in management and planning for mobile network operators.
Yang *et al.* 2020	[139]	M	L	×	H	L	In this paper, the authors propose an AI-based intelligent architecture for 6G networks, which can realize intelligent resource management, automatic network adjustment and intelligent service issuance.
Letaief *et al.* 2021	[140]	H	M	×	M	L	In this paper, the authors present their vision for scalable and trustworthy edge AI systems through the integrated design of wireless communication strategies and decentralized machine learning models.
Xiao *et al.* 2020	[141]	M	×	×	H	L	The authors present a self-learning architecture based on self-supervised Generative Adversarial Networks (GANs) to demonstrate the potential performance improvements that can be realized through automated data learning at the edge of the network.

H :High Coverage M :Medium Coverage L :Low Coverage × :Not Mentioned

of enormous communication networks will lead to a large amount of energy consumption. It will become an urgent task to promote the research of green communication networks. However, the requirements of 6G on QoS, security, flexibility, and intelligence will become increasingly strict and diversified, which will bring significant challenges to the development of green communication. Today's research in green communication is based on two main ideas: using energy harvesting techniques and combining artificial intelligence algorithms and traditional network management models. Energy harvesting technology refers to the rapid and wide-scale deployment of energy harvesting units such as solar panels and wind turbines to collect and convert various sustainable energy sources into electrical energy for Information and Communication Technology (ICT) devices [142]. Another idea is to combine AI's DL algorithm or ML algorithm with traditional AI algorithms and mathematical models, applying them in the management of 6G networks to reduce the complexity of the algorithms and improve the efficiency of network resource management [136].

Security Issues in the Combination of AI and 6G Artificial intelligence refers to systems and machines that can mimic humans to perform tasks and iteratively improve themselves based on the information collected. Today, AI has become a ubiquitous term, many applications that can automatically perform complex tasks can be called AI. The application of AI in 6G networks will contribute to the optimization and design of architectures, protocols, and operations. Therefore, AI is of great interest in the development of 6G networks [143]. AI can adaptively adjust some models and architectures through a large amount of data training. Since 6G networks will face a large amount of data transmission and various architectural models, it is unrealistic to configure, monitor, and repair them manually. Therefore, 6G is in need of AI, which can help it achieve "intelligent endogenous" [144]. The application of AI in 6G networks will face many security issues [145]. We list some cybersecurity issues AI brings to 6G networks in Table 4 and provide some possible solutions.

1. Malicious samples

   Just because AI needs to collect a large amount of data for training, many attackers will attack 6G networks from this point. For example, an attacker may lure the AI to obtain malicious samples that are deliberately prepared(e.g., manipulating labeled data and weak labeling) for training. Then, the training data can then be tampered with to affect the learning results, leading to wrong classification and regression results. The attacker focuses on interfering with the learning algorithm and the learning logic. They can attack the agent model in the distributed learning model, manipulate the partial model parameters through malicious agents, and thus corrupt the global model. The principle of this attack method is shown in Figure 13:

   Barreno *et al.* [146] proposed a Rejection Of Negative Impact (RONI) method, which is used to detect whether the training samples are malicious. They first trained a classifier on a base training set, then added the training samples to be detected to the training set, and subsequently trained a second classifier on the training set used for comparison. They apply both classifiers to the same set of instances with known labels and compare the accuracy difference between them. The samples can be considered harmful if the first classifier produces more classification errors. Li *et*

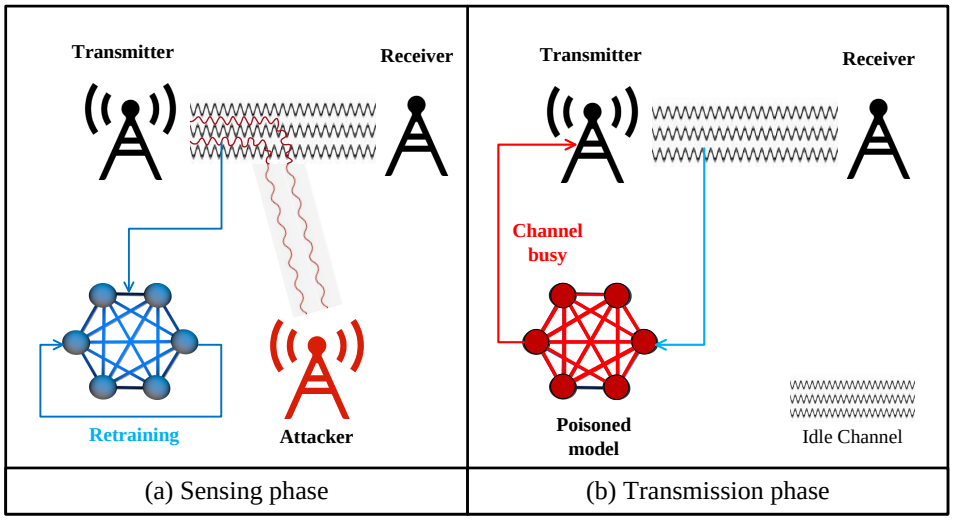


Fig. 13. The principle of malicious samples attack

al. [147] proposed a homomorphic encryption scheme with ciphertext comparison to secure machine learning training and classification.

2. Attacks based on AI

   AI can have the ability to make network-wide intelligent decisions with a distributed edge-based architecture. Attackers can discover vulnerabilities in the network by revealing patterns in large amounts of data at different levels (smart radio, edge, cloud) through an artificial intelligence-based approach. For example, AI can learn the vulnerable IoT device, compromise it, gain control of it, and subsequently launch a DDoS attack on a critical node [148].

   The solution to this AI-based attack is to use AI to train a more complete and more secure defense system by distributed intelligence. The moving target defense technique proposed by Cho *et al.* [149] is a proactive measure. It is an approach that invokes dynamism for 6G networks, thus weakening the learning speed and the learning level of AI attackers. Biamonte *et al.* [150] proposes that quantum machine learning can also design advanced defense techniques to defend against AI attacks.

3. Backdoor attack

   This type of attack generally does not affect the accuracy of the main task. A backdoor is first inserted into the AI model, or a malicious task is injected. The backdoor attack is triggered once a set event or a specific condition is triggered during the normal training process. These attacks are difficult to detect and can spoof ML models with high confidence [151].

   Aramoon *et al.* [152] introduced the concept of meta-federated learning, a new federated learning framework that facilitates the defense against backdoor attacks and the protection of user privacy. They also showed that shifting the defense execution point from a single update level to an aggregation level can effectively mitigate the impact of backdoor attacks without violating user privacy. A previous study by Liu *et al.* [153] showed that pruning and fine-tuning are two effective ways to defend

against backdoor attacks, while in later research, the combination of pruning and fine-tuning was evaluated and shown experimentally to be effective in removing or even eliminating backdoor attacks.

Table 4. Security threats raised by AI in 6G networks and possible solutions

Security threats	Possible solutions		
	Authors&Year	**Ref**	**Remarks**
Malicious samples	Barreno *et al.* 2010	[146]	The authors propose a Rejection of Negative Influence (RONI) method, which is used to detect whether a training sample is malicious or not.
	Li *et al.* 2020	[147]	The authors propose a homomorphic encryption scheme with ciphertext comparison to secure machine learning training and classification.
AI based attacks	Cho *et al.* 2020	[149]	In this paper, the authors propose a moving target defense technique to cripple the learning speed as well as the learning level of an AI attacker.
	Biamonte *et al.* 2017	[150]	The authors propose to design advanced defense techniques through quantum machine learning to defend against AI attacks.
Backdoor attacks	Aramoon *et al.* 2021	[152]	The authors propose a meta-federation learning framework that facilitates defense against backdoor attacks.
	Liu *et al.* 2018	[153]	The combination of pruning and fine-tuning is evaluated and experimentally shown to remove or eliminate backdoor attacks effectively.
Attacks against AI algorithms	Karimireddy *et al.* 2020	[154]	A SCAFFOLD algorithm is proposed, which utilizes the similarity of client data to enable the model to converge faster.
	Babaeer *et al.* 2020	[155]	The authors propose a lightweight security approach based on threshold-sensitive energy-efficient sensor network protocols and watermarking techniques to ensure the integrity and authenticity of sensed data.
	Fang *et al.* 2020	[156]	A lightweight encryption protocol is designed, and a new efficient federated learning scheme with strong privacy protection in cloud computing is proposed.

4. Attacks against AI algorithms

Aggregation algorithms orchestrate AI training models in a centralized or peer-to-peer setup. Attacks against these algorithms are typically carried out either through the aggregator or by altering the algorithmic parameters. And the system's integrity will be compromised in the event of a successful attack [151]. First, in terms of aggregators, when suffering from attacks such as data poisoning or model poisoning, non-

robust aggregators will produce compromised models. As a result, the training results will be affected. In terms of parameters, the attacker may manipulate parameters such as learning rate, batch size, among others, through some compromised clients, thus forcing the global model convergence to fail or the model's training process to be terminated.

For the problem of malicious parameter modification leading to non-convergence or slow convergence of the global model, Karimireddy *et al.* [154] proposed a SCAFFOLD algorithm that utilizes the similarity of the client's data to enable the model to converge faster. As for how to solve the problem of tampering with algorithm parameters, Babaeer *et al.* [155] proposed a lightweight security method based on threshold-sensitive energy-efficient sensor network protocols and watermarking techniques, which employs Homomorphic Encryption (HE) to ensure the integrity and authenticity of the sensed data sent from the sensor nodes to the base station. Moreover, this method also detects if the data has been tampered with. Fang *et al.* also designed a lightweight encryption protocol and proposed a new efficient federated learning scheme with solid privacy preservation in cloud computing [156], which is used to prevent the risk of model parameters shared directly between cloud servers and clients from being used for model inversion attacks. The scheme also employs an efficient optimization strategy to improve training efficiency.

5.2. AI assists 6G networks in addressing security threats

In the third part of this work, we analyze the security issues of 6G networks in terms of network architecture, network requirements, and different application areas. Meanwhile, in the fourth part, we analyze the security issues and corresponding solutions of other technologies applied in 6G networks. These two sections give us a preliminary understanding of the security threats in 6G networks. In this section, we analyze the perspectives from which AI assists 6G networks in addressing security threats. As depicted in Table 5, We list examples of AI assisting 6G networks to address security threats.

Security issues on edge networks The development of IoT, 6G networks, and autonomous driving is driving the emergence of more and more connected end devices. However, due to these devices' limited communication and computing resources, they are vulnerable to malicious attacks by intruders. Thereby gaining control of these end devices, leading to DDoS attacks on the devices. This, in turn, generates a large number of computation and processing tasks, exhausting the resources of edge servers [157]. With location-based services becoming more common, the risk of location privacy breaches is increasing [57]. Emerging network services such as telemedicine, teleconferencing, and remote construction realize real-time, remote, and convenient control and management. However, these remote control operations are susceptible to malware attacks. Once the manipulation right is intercepted by malware, it will not only lead to the leakage of a large amount of private information, but also the deprivation of the control of the computation part of the terminal devices and edge servers [158]. This leads to unauthorized computing tasks, resulting in the loss of many resources. Some hackers also provide spurious data to edge servers. As the AI model collects data for training, it enters inside the model and disrupts the AI training process [159]. This leads to a reduction in the accuracy of the training or makes

the training proceed in the wrong direction. It even leads to the failure of the entire training process, ultimately leading to the degradation of the network performance of the edge server [160].

For the above scenario of 6G networks subject to malicious attacks on edge servers, Lovén *et al.* [161] proposed that AI can be used to build machine learning models at the edge based on large amounts of real-time data on system control and application execution. Service deployment for IoT applications in Mobile Edge Computing (MEC) environments is a key problem to be solved. Considering the knowledge of mobile users' service requests and the processing power of edge servers, the microservice deployment problem in the MEC environment can be modeled as a nonlinear optimization problem. Tang *et al.* proposed an adaptive dynamic deployment optimization method, Adapt-SD [162], based on Adam and weighted round-robin scheduling algorithms, to solve the microservice deployment problem. The edge-native AI approach provides ideas for solving security problems on 6G edge networks and optimizes the execution of edge applications. It also improves the network's connectivity, reactivity, adaptability, and proactivity.

Physical security protection Physical security in 6G networks mainly reflects security issues on IoT, since the commissioning of the future 6G network requires many IoT devices as support. The security problem of communication between these devices becomes an essential threat to constructing a secure and stable network environment. There are many kinds of common physical attacks, such as the capture of sensor devices mentioned in [172,173]. The attacker obtains access to the sensor device and then uses the captured device to attack the whole network and cause significant damage. Another example is the physical attacks on network devices mentioned in [93,174], where the attacker damages the network devices by artificial physical means, thus affecting the normal operation of the communication network. There are also attacks that obtain control of devices by forging identities, which threaten network security [175].

For the abovementioned attacks against IoT physical devices, one can first consider deploying symmetric encryption systems [163,164] or asymmetric encryption systems [165] in IoT environments to prevent devices from being intruded or data from being tampered with. Communication between two IoT devices or nodes via key and encryption/decryption mechanisms can also be considered, thus avoiding eavesdropping on the communication or interruption of the communication process. Based on the above ideas, AI can provide better learning and adaptive capabilities for these safeguards and integrate new technologies to enhance the security-based of these technologies.

Autonomous mechanisms for addressing security risks For 6G networks, which are substantial communication networks, it is unrealistic to manually monitor each device's status in real-time, detect anomalies, and make timely responses. However, we can use AI to monitor device status and detect security threats and then adaptively take countermeasures, such as dynamically adjusting the network configuration or security authentication model to defend against security threats.

Shen *et al.* proposed an AI-based 6G network architecture whose core network consists of SDN [166]. Operational data is collected by emulating the virtualized network infrastructure. The AI learns from the collected data and determines appropriate network

Table 5. AI's contribution to solving 6G cybersecurity problems

Contributions	Examples		
	Authors&Year	**Ref**	**Remarks**
Assist 6G in solving security problems on edge networks	Lovén *et al.* 2019	[161]	The authors propose that AI can build machine learning models on edge networks based on large amounts of real-time system control and application execution data.
Assistance in responding to physical attacks	Jawad *et al.* 2019	[163]	Deploying symmetric encryption systems in IoT environments to prevent device intrusion or data tampering.
	Alotaibi *et al.* 2018	[164]	
	Alaya *et al.* 2021	[165]	The authors propose to deploy asymmetric encryption systems on physical devices to prevent devices from being compromised.
Provide autonomous mechanism to address security risks	Shen *et al.* 2021	[166]	An AI-based 6G network architecture is proposed, where the AI learns from the simulation data to realize intelligent resource scheduling and security configuration management and then realizes dynamic adjustment to external influences.
	Fang *et al.* 2022	[167]	This article introduces autonomous collaborative authentication and proposes a collaborative authentication scheme with privacy preservation based on federated learning.
	Xiao *et al.* 2019	[168]	The authors propose RL-based intelligent beamforming techniques to provide optimal beamforming strategies against eavesdropper attacks in VLC systems.
Future security risk forecasting	Husák *et al.* 2018	[169]	The authors propose that deep learning-based attack prediction methods are promising for malicious behavior detection.
	Tariq *et al.* 2020	[170]	The authors propose that the combination of future 6G networks with AI has the potential to realize the concept of cognitive radio. And by storing behavioral data about the environment, AI can be used to predict suspicious activities of malicious nodes.
	Zhang *et al.* 2021	[171]	This paper proposes the need to investigate intelligent and flexible security mechanisms to deal with different forms of security threats.

management policies for intelligent resource scheduling and security configuration management. In turn, it realizes dynamic adjustment to external influences. Due to the complex heterogeneous networks in 6G and the distributed nature of the devices and information involved, entities in 6G networks are susceptible to privacy leakage. Fang *et al.* [167] introduced autonomous collaborative authentication and proposed a federated learning-

based collaborative authentication scheme with privacy-preserving features. The scheme analyzes the heterogeneous information of the authenticated devices and then updates the respective authentication models locally. Then, the corresponding authentication model is used to realize privacy protection. In VLC systems for 6G networks, intelligent beamforming techniques based on Reinforcement Learning (RL) provide optimal beamforming strategies against eavesdropper attacks [168].

Future security risk projections In the future, we may face various cyber attacks in 6G networks, and they may use different methods to launch attacks from different angles. We need to detect these attacks in advance to address these security threats better and minimize the loss of benefits. We must also anticipate future security risks as attackers may change their attack methods. By detecting security threats and predicting security risks in advance, we can take countermeasures to increase the success rate of defending against cyberattacks.

In complex 6G network environments, many network parameters and states are in constant flux, which increases the difficulty of predicting risks. However, deep learning-based attack prediction methods in this changing environment are promising for malicious behavior detection [169]. Tariq *et al.* [170] also suggested that the combination of future 6G networks with AI has the potential to realize the concept of cognitive radio proposed by Mitola *et al.* in 1999. It is also argued that 6G Mitola radios will also store behavioral data of the environment so that suspicious activities of malicious nodes can be predicted through AI. We must study intelligent and flexible security mechanisms to predict, detect, mitigate, and prevent security attacks to cope with different security threats. Thus, the propagation of such vulnerabilities in 6G networks can be limited [171].

6. Concluding Remarks and Outlook

As a next-generation mobile communication technology, 6G network, based on 5G communication technology, will span the connection between people and people and the connection between people and things and eventually move towards the smart relationship of everything, allowing intelligence to be reflected in all areas of society and bringing a new face to human society. However, the large-scale use of 6G networks and their application in new fields will also make their safety challenging.

This work analyzed and discussed the security issues surrounding 6G networks and their applications. By reviewing relevant research articles, this paper introduced the development process of mobile communication technology and discussed the research motivation of 6G networks. This paper also compared the key performance index differences between 5G networks and 6G networks and analyzed the key technologies of 6G networks. This work examined the security of 6G architecture regarding network access, network domain, user domain, and application domain. Thus, this work also compared the three significant visions of 5G and 6G and summarized the security threats in the 6G visions. Security issues of 6G applications in areas such as UAV mobile communication, Smart Grid 2.0, and Industrial 5.0 were also analyzed, listing the applications of terahertz communication, visible light communication, and blockchain technology in 6G. For the above applications, we analyzed some security issues and selected the corresponding solutions proposed in recent research articles. The RONI method for detecting malicious samples,

the WBPLSec method for protecting visible light communication, and the context-based multi-factor biometric authentication algorithm mentioned in this paper all provided effective ideas for solving the security problems of 6G networks. We also introduced the application of AI to the issue of 6G network security. Firstly, we illustrated the impact of AI on 6G networks from two aspects: AI promotes the construction of 6G networks, and AI brings cyber risks to 6G networks, and we demonstrated that AI can assist 6G networks in solving security threats from four perspectives.

Because 6G networks cover a wide range of areas and have a complex architecture, much work remains to be done on 6G network security. Network attacks in the future will become increasingly sophisticated, such as the one mentioned in this work, which uses AI to carry out attacks. This requires defenders to research more comprehensive and complete solutions to address such security threats. In terms of artificial intelligence, we hope to propose more effective detection methods for malicious sample detection. In terms of terahertz technology, we hope to take measures to deal with the problem of signal interference. In blockchain technology, we hope to use the distributed ledger mechanism and smart contracts to make corresponding improvements in privacy and security protection measures. In future directions, we will conduct more in-depth research on the security problems of 6G networks in blockchain and AI, and we expect to contribute to the research of 6G network security by collecting information and conducting simulation experiments by profoundly analyzing the security risks and proposing the corresponding countermeasures.

Acknowledgment

This work was partially supported by the National Key Research and Development Program of China under Grant 2021YFA1000600, the National Natural Science Foundation of China under Grant 62072170, the Science and Technology Project of the Department of Communications of Hunan Provincial under Grant 202101, the Key Research and Development Program of Hunan Province under Grant 2022GK2015, and the Hunan Provincial Natural Science Foundation of China under Grant 2021JJ30141. The co-author Muhammad Khurram Khan is supported by King Saud University, Riyadh, Saudi Arabia, under project number (RSP2024R12).

References

1. Qi Bi. Ten trends in the cellular industry and an outlook on 6g. *IEEE Communications Magazine*, 57(12):31–36, 2019.
2. Reeya Agrawal. Comparison of different mobile wireless technology (from 0g to 6g). *ECS Transactions*, 107(1):4799, 2022.
3. AFM Shahen Shah, Ahmed Nidham Qasim, Muhammet Ali Karabulut, Haci Ilhan, and Md Baharul Islam. Survey and performance evaluation of multiple access schemes for next-generation wireless communication systems. *IEEE Access*, 9:113428–113442, 2021.
4. Jessica Moysen and Lorenza Giupponi. From 4g to 5g: Self-organized network management meets machine learning. *Computer Communications*, 129:248–268, 2018.
5. Joyce Ayoola Adebusola, Adebiyi Ayodele Ariyo, Okeyinka Aderemi Elisha, Adebiyi Marion Olubunmi, and Okesola Olatunji Julius. An overview of 5g technology. In *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)*, pages 1–4. IEEE, 2020.

6. Klaus David and Hendrik Berndt. 6g vision and requirements: Is there any need for beyond 5g? *IEEE vehicular technology magazine*, 13(3):72–80, 2018.
7. Karthik Kumar Vaigandla, Nilofar Azmi, Ramya Podila, and Radha Krishna Karne. A survey on wireless communications: 6g and 7g. *International Journal Of Science, Technology & Management*, 2(6):2018–2025, 2021.
8. Matti Latva-aho, Kari Leppänen, Federico Clazzer, and Andrea Munari. Key drivers and research challenges for 6g ubiquitous wireless intelligence. 2020.
9. Zihang Song, Yue Gao, and Rahim Tafazolli. A survey on spectrum sensing and learning technologies for 6g. *IEICE Transactions on Communications*, 104(10):1207–1216, 2021.
10. Choongil Yeh, Gweon Do Jo, Young-Jo Ko, and Hyun Kyu Chung. Perspectives on 6g wireless communications. *ICT Express*, 9(1):82–91, 2023.
11. Muhammad Waseem Akhtar, Syed Ali Hassan, Rizwan Ghaffar, Haejoon Jung, Sahil Garg, and M Shamim Hossain. The shift to 6g communications: vision and requirements. *Human-centric Computing and Information Sciences*, 10:1–27, 2020.
12. Yushan Siriwardhana, Pawani Porambage, Madhusanka Liyanage, and Mika Ylianttila. Ai and 6g security: Opportunities and challenges. In *2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, pages 616–621. IEEE, 2021.
13. Zisang Xu, Wei Liang, Kuan-Ching Li, Jianbo Xu, Albert Y Zomaya, and Jixin Zhang. A time-sensitive token-based anonymous authentication and dynamic group key agreement scheme for industry 5.0. *IEEE Transactions on Industrial Informatics*, 18(10):7118–7127, 2021.
14. Dinh C Nguyen, Ming Ding, Pubudu N Pathirana, Aruna Seneviratne, Jun Li, Dusit Niyato, Octavia Dobre, and H Vincent Poor. 6g internet of things: A comprehensive survey. *IEEE Internet of Things Journal*, 9(1):359–383, 2021.
15. Chengxiao Liu, Wei Feng, Yunfei Chen, Cheng-Xiang Wang, and Ning Ge. Cell-free satellite-uav networks for 6g wide-area internet of things. *IEEE Journal on Selected Areas in Communications*, 39(4):1116–1131, 2020.
16. Y Li, Wei Liang, K Xie, D Zhang, S Xie, and KC Li. Lightnestle: quick and accurate neural sequential tensor completion via meta learning. In *IEEE Infocom*, 2023.
17. José A del Peral-Rosado, Ronald Raulefs, José A López-Salcedo, and Gonzalo Seco-Granados. Survey of cellular mobile radio localization methods: From 1g to 5g. *IEEE Communications Surveys & Tutorials*, 20(2):1124–1148, 2017.
18. Rani P Tidke, Pritee S Uttarwar, Deepak S Dandwate, and Umesh J Tupe. A literature review on: Wireless technologies from 0g to 7g. *IRE Journals*, 4(6):59–64, 2020.
19. AFM Shahen Shah. A survey from 1g to 5g including the advent of 6g: Architectures, multiple access techniques, and emerging technologies. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 1117–1123. IEEE, 2022.
20. Mamta Agiwal, Hyeyeon Kwon, Seungkeun Park, and Hu Jin. A survey on 4g-5g dual connectivity: road to 5g implementation. *Ieee Access*, 9:16193–16210, 2021.
21. Wan Siti Halimatul Munirah Wan Ahmad, Nurul Asyikin Mohamed Radzi, FS Samidi, Aiman Ismail, Fairuz Abdullah, Md Zaini Jamaludin, and MohdNasim Zakaria. 5g technology: Towards dynamic spectrum sharing using cognitive radio networks. *IEEE access*, 8:14460–14488, 2020.
22. Peter Backeman, Ashalatha Kunnappilly, and Cristina Seceleanu. Supporting 5g service orchestration with formal verification. *Computer Science and Information Systems*, 20(1):329–357, 2023.
23. Anutusha Dogra, Rakesh Kumar Jha, and Shubha Jain. A survey on beyond 5g network with the advent of 6g: Architecture and emerging technologies. *IEEE Access*, 9:67512–67547, 2020.
24. Samer Henry, Ahmed Alsohaily, and Elvino S Sousa. 5g is real: Evaluating the compliance of the 3gpp 5g new radio system with the itu imt-2020 requirements. *IEEE Access*, 8:42828–42840, 2020.

25. Karen Campbell, Jim Diffley, Bob Flanagan, Bill Morelli, Brendan O'Neil, Francis Sideco, et al. The 5g economy: How 5g technology will contribute to the global economy. *IHS economics and IHS technology*, 4(16):1, 2017.
26. IIMT Union. Imt traffic estimates for the years 2020 to 2030. *Report ITU*, 2370, 2015.
27. Wei Jiang, Bin Han, Mohammad Asif Habibi, and Hans Dieter Schotten. The road towards 6g: A comprehensive survey. *IEEE Open Journal of the Communications Society*, 2:334–366, 2021.
28. Wei Liang, Yang Yang, Ce Yang, Yonghua Hu, Songyou Xie, Kuan-Ching Li, and Jiannong Cao. Pdpchain: A consortium blockchain-based privacy protection scheme for personal data. *IEEE Transactions on Reliability*, 2022.
29. Wei Liang, Dafang Zhang, Xia Lei, Mingdong Tang, Kuan-Ching Li, and Albert Y Zomaya. Circuit copyright blockchain: blockchain-based homomorphic encryption for ip circuit protection. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1410–1420, 2020.
30. Ying Long, Yinyan Gong, Weihong Huang, Jiahong Cai, Nengxiang Xu, and Kuan-ching Li. Cryptography of blockchain. In *Smart Computing and Communication: 7th International Conference, SmartCom 2022, New York City, NY, USA, November 18–20, 2022, Proceedings*, pages 340–349. Springer, 2023.
31. Hui Chen, Hadi Sarieddeen, Tarig Ballal, Henk Wymeersch, Mohamed-Slim Alouini, and Tareq Y Al-Naffouri. A tutorial on terahertz-band localization for 6g communication systems. *IEEE Communications Surveys & Tutorials*, 2022.
32. Wei Liang, Yuhui Li, Jianlong Xu, Zheng Qin, Dafang Zhang, and Kuan-Ching Li. Qos prediction and adversarial attack protection for distributed services under dlaas. *IEEE Transactions on Computers*, 2023.
33. Qiannan Zhang, Huafeng Wu, Xiaojun Mei, Dezhi Han, Mario Donato Marino, Kuan-Ching Li, and Song Guo. A sparse sensor placement strategy based on information entropy and data reconstruction for ocean monitoring. *IEEE Internet of Things Journal*, 2023.
34. Gang Dong, Lin Pan, Yi Zhang, and Yanan Liu. Air-space-ground integrated information network: Technology, development and prospect. In *Signal and Information Processing, Networking and Computers: Proceedings of the 8th International Conference on Signal and Information Processing, Networking and Computers (ICSINC)*, pages 883–891. Springer, 2022.
35. Wei Liang, Zuoting Ning, Songyou Xie, Yupeng Hu, Shaofei Lu, and Dafang Zhang. Secure fusion approach for the internet of things in smart autonomous multi-robot systems. *Information Sciences*, 579:468–482, 2021.
36. Nan Chi, Yingjun Zhou, Yiran Wei, and Fangchen Hu. Visible light communication in 6g: Advances, challenges, and prospects. *IEEE Vehicular Technology Magazine*, 15(4):93–102, 2020.
37. Yiying Wang, Xin Kang, Tieyan Li, Haiguang Wang, Cheng-Kang Chu, and Zhongding Lei. Six-trust for 6g: Towards a secure and trustworthy 6g network. *arXiv preprint arXiv:2210.17291*, 2022.
38. Wei Liang, Songyou Xie, Dafang Zhang, Xiong Li, and Kuan-ching Li. A mutual security authentication method for rfid-puf circuit based on deep learning. *ACM Transactions on Internet Technology (TOIT)*, 22(2):1–20, 2021.
39. Zhen Sun, Liyuan Song, Qin Huang, Liuguo Yin, Guilu Long, Jianhua Lu, and Lajos Hanzo. Toward practical quantum secure direct communication: A quantum-memory-free protocol and code design. *IEEE Transactions on Communications*, 68(9):5778–5792, 2020.
40. Jawad Tanveer, Amir Haider, Rashid Ali, and Ajung Kim. Machine learning for physical layer in 5g and beyond wireless networks: A survey. *Electronics*, 11(1):121, 2022.
41. Minghao Wang, Tianqing Zhu, Tao Zhang, Jun Zhang, Shui Yu, and Wanlei Zhou. Security and privacy in 6g networks: New areas and new challenges. *Digital Communications and Networks*, 6(3):281–291, 2020.

42. Gao Mine, Jiao Hai, Luo Jin, and Zhou Huiying. A design of sd-wan-oriented wide area network access. In *2020 International Conference on Computer Communication and Network Security (CCNS)*, pages 174–177. IEEE, 2020.
43. Osama A Khashan, Rami Ahmad, and Nour M Khafajah. An automated lightweight encryption scheme for secure and energy-efficient communication in wireless sensor networks. *Ad Hoc Networks*, 115:102448, 2021.
44. Yahya M Al-Moliki, Mohammed T Alresheedi, Yahya Al-Harthi, and Ali H Alqahtani. Robust lightweight-channel-independent ofdm-based encryption method for vlc-iot networks. *IEEE Internet of Things Journal*, 9(6):4661–4676, 2021.
45. Ghada Arfaoui, Pascal Bisson, Rolf Blom, Ravishankar Borgaonkar, Håkan Englund, Edith Félix, Felix Klaedtke, Prajwol Kumar Nakarmi, Mats Näslund, Piers O'Hanlon, et al. A security architecture for 5g networks. *IEEE access*, 6:22466–22479, 2018.
46. Abbas Yazdinejad, Reza M Parizi, Ali Dehghantanha, Qi Zhang, and Kim-Kwang Raymond Choo. An energy-efficient sdn controller architecture for iot networks with blockchain-based security. *IEEE Transactions on Services Computing*, 13(4):625–638, 2020.
47. Chun-Yu Liu, Shanq-Jang Ruan, Yu-Ren Lai, and Chih-Yuan Yao. Finger-vein as a biometric-based authentication. *IEEE Consumer Electronics Magazine*, 8(6):29–34, 2019.
48. Wei Liang, Yuhui Li, Kun Xie, Dafang Zhang, Kuan-Ching Li, Alireza Souri, and Keqin Li. Spatial-temporal aware inductive graph neural network for c-its data recovery. *IEEE Transactions on Intelligent Transportation Systems*, 2022.
49. Sushanta Sengupta. A secured biometric-based authentication scheme in iot-based patient monitoring system. In *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*, pages 501–518. Springer, 2020.
50. Jin Cao, Maode Ma, Hui Li, Ruhui Ma, Yunqing Sun, Pu Yu, and Lihui Xiong. A survey on security aspects for 3gpp 5g networks. *IEEE communications surveys & tutorials*, 22(1):170–195, 2019.
51. Yanrong Lu and Dawei Zhao. Providing impersonation resistance for biometric-based authentication scheme in mobile cloud computing service. *Computer Communications*, 182:22–30, 2022.
52. Jing Long, Wei Liang, Kuan-Ching Li, Yehua Wei, and Mario Donato Marino. A regularized cross-layer ladder network for intrusion detection in industrial internet of things. *IEEE Transactions on Industrial Informatics*, 19(2):1747–1755, 2022.
53. Shimaa A Abdel Hakeem, Hanan H Hussein, and HyungWon Kim. Security requirements and challenges of 6g technologies and applications. *Sensors*, 22(5):1969, 2022.
54. Muhammad Sajjad Akbar, Zawar Hussain, Quan Z Sheng, and Subhas Mukhopadhyay. 6g survey on challenges, requirements, applications, key enabling technologies, use cases, ai integration issues and security aspects. *arXiv preprint arXiv:2206.00868*, 2022.
55. Keyvan Ramezanpour and Jithin Jagannath. Intelligent zero trust architecture for 5g/6g networks: Principles, challenges, and the role of machine learning in the context of o-ran. *Computer Networks*, page 109358, 2022.
56. Sulaiman Khan, Anwar Hussain, Shah Nazir, Fazlullah Khan, Ammar Oad, and Mohammad Dahman Alshehri. Efficient and reliable hybrid deep learning-enabled model for congestion control in 5g/6g networks. *Computer Communications*, 182:31–40, 2022.
57. Shiwen Zhang, Biao Hu, Wei Liang, Kuan-Ching Li, and Brij B Gupta. A caching-based dual k-anonymous location privacy-preserving scheme for edge computing. *IEEE Internet of Things Journal*, 2023.
58. Van-Linh Nguyen, Po-Ching Lin, Bo-Chao Cheng, Ren-Hung Hwang, and Ying-Dar Lin. Security and privacy for 6g: A survey on prospective technologies and challenges. *IEEE Communications Surveys & Tutorials*, 23(4):2384–2428, 2021.
59. Zisang Xu, Wei Liang, Kuan-Ching Li, Jianbo Xu, and Hai Jin. A blockchain-based roadside unit-assisted authentication and key agreement protocol for internet of vehicles. *Journal of Parallel and Distributed Computing*, 149:29–39, 2021.

60. Wei Liang, Songyou Xie, Jing Long, Kuan-Ching Li, Dafang Zhang, and Keqin Li. A double puf-based rfid identity authentication protocol in service-centric internet of things environments. *Information Sciences*, 503:129–147, 2019.
61. Shiwen Zhang, Ziwei Yan, Wei Liang, Kuan-Ching Li, and Ciprian Dobre. Baka: Biometric authentication and key agreement scheme based on fuzzy extractor for wireless body area networks. *IEEE Internet of Things Journal*, 2023.
62. Mehdi Rasti, Shiva Kazemi Taskou, Hina Tabassum, and Ekram Hossain. Evolution toward 6g multi-band wireless networks: A resource management perspective. *IEEE Wireless Communications*, 29(4):118–125, 2022.
63. Pawani Porambage, Gürkan Gür, Diana Pamela Moya Osorio, Madhusanka Livanage, and Mika Ylianttila. 6g security challenges and potential solutions. In *2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, pages 622–627. IEEE, 2021.
64. Naijian Zhang, Qi Jiang, Long Li, Xindi Ma, and Jianfeng Ma. An efficient three-factor remote user authentication protocol based on bpv-fourq for internet of drones. *Peer-to-Peer Networking and Applications*, 14:3319–3332, 2021.
65. Kan Yu, Jiguo Yu, and Anming Dong. Cooperative communication and mobility for securing urllc of future wireless networks. *IEEE Transactions on Vehicular Technology*, 71(5):5331–5342, 2022.
66. Zhengyu Zhu, Gengwang Hou, Zheng Chu, Xingwang Li, Gangcan Sun, Wanming Hao, and Paramjit S Sehdev. Research and analysis of urllc technology based on artificial intelligence. *IEEE Communications Standards Magazine*, 5(2):37–43, 2021.
67. IMT CAICT. Promotion group, 5g security report. *The China Academy of Information and Communications Technology (CAICT) and IMT*, 2020.
68. Mohamed Amine Ferrag, Leandros Maglaras, Antonios Argyriou, Dimitrios Kosmanos, and Helge Janicke. Security for 4g and 5g cellular networks: A survey of existing authentication and privacy-preserving schemes. *Journal of Network and Computer Applications*, 101:55–82, 2018.
69. Ding Wang and Ping Wang. Two birds with one stone: Two-factor authentication with security beyond conventional bound. *IEEE transactions on dependable and secure computing*, 15(4):708–722, 2016.
70. Marco Lourenco, Louis Marinos, and Lampros Patseas. Enisa threat landscape for 5g networks. In *Proc. Eur. Union Agency Cybersecurity (ENISA)*, page 87, 2019.
71. Najib Ahmed Mohammed, Ali Mohammed Mansoor, and Rodina Binti Ahmad. Mission-critical machine-type communication: An overview and perspectives towards 5g. *IEEE Access*, 7:127198–127216, 2019.
72. Cosmas Ifeanyi Nwakanma, Alifia Putri Anantha, Fabliha Bushra Islam, Jae-Min Lee, and Dong-Seong Kim. 3gpp release-16 for industrial internet of things and mission critical communications. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 403–406. IEEE, 2020.
73. Dakhaz Mustafa Abdullah and Siddeeq Y Ameen. Enhanced mobile broadband (embb): A review. *Journal of Information Technology and Informatics*, 1(1):13–19, 2021.
74. Zhenyu Na, Yue Liu, Jingcheng Shi, Chungang Liu, and Zihe Gao. Uav-supported clustered noma for 6g-enabled internet of things: Trajectory planning and resource allocation. *IEEE Internet of Things Journal*, 8(20):15041–15048, 2020.
75. Xiaohu You, Cheng-Xiang Wang, Jie Huang, Xiqi Gao, Zaichen Zhang, Mao Wang, Yongming Huang, Chuan Zhang, Yanxiang Jiang, Jiaheng Wang, et al. Towards 6g wireless communication networks: Vision, enabling technologies, and new paradigm shifts. *Science China Information Sciences*, 64:1–74, 2021.
76. Helga E Melcherts. The internet of everything and beyond: the interplay between things and humans. *Human bond communication: the holy grail of holistic communication and immersive experience*, pages 173–185, 2017.

77. Lijun Xiao, Wei Liang, Jiahong Cai, Ming Wang, Jiahong Xiao, Yinyan Gong, and Weigang Zhang. An efficiency control strategy of dual-motor multi-gear drive algorithm. *Connection Science*, 35(1):2249264, 2023.
78. Chunyan Diao, Dafang Zhang, Wei Liang, Kuan-Ching Li, Yujie Hong, and Jean-Luc Gaudiot. A novel spatial-temporal multi-scale alignment graph neural network security model for vehicles prediction. *IEEE Transactions on Intelligent Transportation Systems*, 24(1):904–914, 2022.
79. Yinyan Gong, Kuanching Li, Lijun Xiao, Jiahong Cai, Jiahong Xiao, Wei Liang, and Muhammad Khurram Khan. Vaserp: an adaptive, lightweight, secure, and efficient rfid-based authentication scheme for iov. *Sensors*, 23(11):5198, 2023.
80. Rakesh Shrestha, Rojeena Bajracharya, and Shiho Kim. 6g enabled unmanned aerial vehicle traffic management: A perspective. *IEEE Access*, 9:91119–91136, 2021.
81. Michael Hooper, Yifan Tian, Runxuan Zhou, Bin Cao, Adrian P Lauf, Lanier Watkins, William H Robinson, and Wlajimir Alexis. Securing commercial wifi-based uavs from common security attacks. In *MILCOM 2016-2016 IEEE Military Communications Conference*, pages 1213–1218. IEEE, 2016.
82. Shurui Wang, Aifeng Song, and Yufeng Qian. Predicting smart cities' electricity demands using k-means clustering algorithm in smart grid. *Computer Science and Information Systems*, (00):13–13, 2023.
83. Foysal Ilahi, Shaswata Dutta, Md Mehedi Hasan, Sadia Afrin Rumpa, and AKM Baki. Development of a novel uwb antenna for 6g-iot based smart grid device monitoring system. In *2021 International Conference on Green Energy, Computing and Sustainable Technology (GECOST)*, pages 1–5. IEEE, 2021.
84. Muhammad Tariq, Mansoor Ali, Faisal Naeem, and H Vincent Poor. Vulnerability assessment of 6g-enabled smart grid cyber–physical systems. *IEEE internet of things journal*, 8(7):5468–5475, 2020.
85. Xin Xu, Guangli Zhu, Houyue Wu, Shunxiang Zhang, and Kuan-Ching Li. See-3d: Sentiment-driven emotion-cause pair extraction based on 3d-cnn. *Computer Science and Information Systems*, 20(1):77–93, 2023.
86. Sabuzima Nayak and Ripon Patgiri. A vision on intelligent medical service for emergency on 5g and 6g communication era. *EAI Endorsed Transactions on Internet of Things*, 6(22), 2020.
87. Wenkang Liu, Yuxuan He, Xiaoliang Wang, Ziming Duan, Wei Liang, and Yuzhen Liu. Bfg: privacy protection framework for internet of medical things based on blockchain and federated learning. *Connection Science*, 35(1):2199951, 2023.
88. Jianqiang Hu, Wei Liang, Osama Hosam, Meng-Yen Hsieh, and Xin Su. 5gss: a framework for 5g-secure-smart healthcare monitoring. *Connection Science*, 34(1):139–161, 2022.
89. Jiatao Li, Dezhi Han, Zhongdai Wu, Junxiang Wang, Kuan-Ching Li, and Arcangelo Castiglione. A novel system for medical equipment supply chain traceability based on alliance chain and attribute and role access control. *Future Generation Computer Systems*, 142:195–211, 2023.
90. Lingwei Xu, Xinpeng Zhou, Ye Tao, Xu Yu, Miao Yu, and Fazlullah Khan. Af relaying secrecy performance prediction for 6g mobile communication networks in industry 5.0. *IEEE Transactions on Industrial Informatics*, 18(8):5485–5493, 2021.
91. Lijun Xiao, Dezhi Han, Ce Yang, Jiahong Cai, Wei Liang, and Kuan-Ching Li. Ts-dp:an efficient data processing algorithm for distribution digital twin grid for industry 5.0. *IEEE Transactions on Consumer Electronics*, pages 1–1, 2023.
92. Jamie Snudden. Progression to the next industrial revolution: Industry 4.0 for composites. *Reinforced Plastics*, 63(3):136–142, 2019.
93. Latif U Khan, Walid Saad, Dusit Niyato, Zhu Han, and Choong Seon Hong. Digital-twin-enabled 6g: Vision, architectural trends, and future directions. *IEEE Communications Magazine*, 60(1):74–80, 2022.

94. Michael Batty. Digital twins, 2018.
95. David Holmes, Maria Papathanasaki, Leandros Maglaras, Mohamed Amine Ferrag, Surya Nepal, and Helge Janicke. Digital twins and cyber security–solution or challenge? In *2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, pages 1–8. IEEE, 2021.
96. Kari Rikkinen, Pekka Kyosti, Marko E Leinonen, Markus Berg, and Aarno Parssinen. Thz radio communication: Link budget analysis toward 6g. *IEEE Communications Magazine*, 58(11):22–27, 2020.
97. Chong Han, Yongzhi Wu, Zhi Chen, and Xudong Wang. Terahertz communications (teracom): Challenges and impact on 6g wireless systems. *arXiv preprint arXiv:1912.06040*, 2019.
98. Rohit Singh and Douglas Sicker. Thz communications-a boon and/or bane for security, privacy, and national security. In *TPRC48: The 48th Research Conference on Communication, Information and Internet Policy*, 2020.
99. Boyu Ning, Zhi Chen, Wenjie Chen, and Lingxiang Li. Improving security of thz communication with intelligent reflecting surface. In *2019 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6. IEEE, 2019.
100. Yiming Li, Xitao Liang, Wenwu Xie, and Juan Zhu. Security performance analysis of active intelligent reflective surface assisted wireless communication. *Computer Science and Information Systems*, (00):11–11, 2023.
101. Vitaly Petrov, Dmitri Moltchanov, Josep Miquel Jornet, and Yevgeni Koucheryavy. Exploiting multipath terahertz communications for physical layer security in beyond 5g networks. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 865–872. IEEE, 2019.
102. Marcos Katz and Iqrar Ahmed. Opportunities and challenges for visible light communications in 6g. *2020 2nd 6G wireless summit (6G SUMMIT)*, pages 1–5, 2020.
103. Mohamed Amine Arfaoui, Mohammad Dehghani Soltani, Iman Tavakkolnia, Ali Ghrayeb, Majid Safari, Chadi M Assi, and Harald Haas. Physical layer security for visible light communication systems: A survey. *IEEE Communications Surveys & Tutorials*, 22(3):1887–1908, 2020.
104. Simone Soderi, Alessandro Brighente, Federico Turrin, and Mauro Conti. Vlc physical layer security through ris-aided jamming receiver for 6g wireless networks. In *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 370–378. IEEE, 2022.
105. Cheng Chen, Rui Bian, and Harald Haas. Omnidirectional transmitter and receiver design for wireless infrared uplink transmission in lifi. In *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6. IEEE, 2018.
106. Ali Hussain Khan, Naveed UL Hassan, Chau Yuen, Jun Zhao, Dusit Niyato, Yan Zhang, and H Vincent Poor. Blockchain and 6g: the future of secure and ubiquitous communication. *IEEE Wireless Communications*, 29(1):194–201, 2021.
107. Jelena Marjanović, Nikola Dalčeković, and Goran Sladić. Blockchain-based model for tracking compliance with security requirements. *Computer Science and Information Systems*, 20(1):359–380, 2023.
108. Wei Liang, Lijun Xiao, Ke Zhang, Mingdong Tang, Dacheng He, and Kuan-Ching Li. Data fusion approach for collaborative anomaly intrusion detection in blockchain-based systems. *IEEE Internet of Things Journal*, 9(16):14741–14751, 2021.
109. Wei Liang, Yongkai Fan, Kuan-Ching Li, Dafang Zhang, and Jean-Luc Gaudiot. Secure data storage and recovery in industrial blockchain network environments. *IEEE Transactions on Industrial Informatics*, 16(10):6543–6552, 2020.
110. Sisi Zhou, Kuanching Li, Lijun Xiao, Jiahong Cai, Wei Liang, and Arcangelo Castiglione. A systematic review of consensus mechanisms in blockchain. *Mathematics*, 11(10):2248, 2023.

111. Tharaka Hewa, Gürkan Gür, Anshuman Kalla, Mika Ylianttila, An Bracken, and Madhusanka Liyanage. The role of blockchain in 6g: Challenges, opportunities and research directions. *2020 2nd 6G Wireless Summit (6G SUMMIT)*, pages 1–5, 2020.
112. Hua-Yi Lin. Secure cloud internet of vehicles based on blockchain and data transmission scheme of map/reduce. *Computer Science and Information Systems*, 20(1):137–156, 2023.
113. Hao Xu, Paulo Valente Klaine, Oluwakayode Onireti, Bin Cao, Muhammad Imran, and Lei Zhang. Blockchain-enabled resource management and sharing for 6g communications. *Digital Communications and Networks*, 6(3):261–269, 2020.
114. Fredy Andres Aponte-Novoa, Ana Lucila Sandoval Orozco, Ricardo Villanueva-Polanco, and Pedro Wightman. The 51% attack on blockchains: A mining behavior study. *IEEE Access*, 9:140549–140564, 2021.
115. Qiheng Zhou, Huawei Huang, Zibin Zheng, and Jing Bian. Solutions to scalability of blockchain: A survey. *Ieee Access*, 8:16440–16455, 2020.
116. Songyou Xie, Lijun Xiao, Dezhi Han, Kun Xie, Xiong Li, and Wei Liang. Hcvc: A high-capacity off-chain virtual channel scheme based on bidirectional locking mechanism. *IEEE Transactions on Network Science and Engineering*, 2023.
117. Javad Zarrin, Hao Wen Phang, Lakshmi Babu Saheer, and Bahram Zarrin. Blockchain for decentralization of internet: prospects, trends, and challenges. *Cluster Computing*, 24(4):2841–2866, 2021.
118. Yaqin Liu, Wei Liang, Kun Xie, Songyou Xie, Kuanching Li, and Weizhi Meng. Lightpay: A lightweight and secure off-chain multi-path payment scheme based on adapter signatures. *IEEE Transactions on Services Computing*, 2023.
119. Weiyu Zhong, Ce Yang, Wei Liang, Jiahong Cai, Lin Chen, Jing Liao, and Naixue Xiong. Byzantine fault-tolerant consensus algorithms: A survey. *Electronics*, 12(18):3801, 2023.
120. Seyed Mojtaba Hosseini Bamakan, Amirhossein Motavali, and Alireza Babaei Bondarti. A survey of blockchain consensus algorithms performance evaluation criteria. *Expert Systems with Applications*, 154:113385, 2020.
121. Md Sadek Ferdous, Mohammad Jabed Morshed Chowdhury, Mohammad A Hoque, and Alan Colman. Blockchain consensus algorithms: A survey. *arXiv preprint arXiv:2001.07091*, 2020.
122. Xiang Fu, Huaimin Wang, and Peichang Shi. A survey of blockchain consensus algorithms: mechanism, design and applications. *Science China Information Sciences*, 64:1–15, 2021.
123. Xian Guo, Chen Wang, Laicheng Cao, Yongbo Jiang, and Yan Yan. A novel security mechanism for software defined network based on blockchain. *Computer Science and Information Systems*, 19(2):523–545, 2022.
124. Gunasekaran Manogaran, Bharat S Rawal, Vijayalakshmi Saravanan, Priyan Malarvizhi Kumar, Oscar Sanjuán Martínez, Rubén González Crespo, Carlos Enrique Montenegro-Marin, and Sujatha Krishnamoorthy. Blockchain based integrated security measure for reliable service delegation in 6g communication environment. *Computer Communications*, 161:248–256, 2020.
125. Jiaheng Wang, Xintong Ling, Yuwei Le, Yongming Huang, and Xiaohu You. Blockchain-enabled wireless communications: a new paradigm towards 6g. *National science review*, 8(9):nwab069, 2021.
126. Khurram Shahzad, Ahmad O Aseeri, and Munam Ali Shah. A blockchain-based authentication solution for 6g communication security in tactile networks. *Electronics*, 11(9):1374, 2022.
127. Ihsan H Abdulqadder and Shijie Zhou. Sliceblock: context-aware authentication handover and secure network slicing using dag-blockchain in edge-assisted sdn/nfv-6g environment. *IEEE Internet of Things Journal*, 9(18):18079–18097, 2022.
128. Han Liu, Dezhi Han, Mingming Cui, Kuan-Ching Li, Alireza Souri, and Mohammad Shojafar. Idenmultisig: identity-based decentralized multi-signature in internet of things. *IEEE Transactions on Computational Social Systems*, 2023.

129. Wei Liang, Songyou Xie, Jiahong Cai, Jianbo Xu, Yupeng Hu, Yang Xu, and Meikang Qiu. Deep neural network security collaborative filtering scheme for service recommendation in intelligent cyber–physical systems. *IEEE Internet of Things Journal*, 9(22):22123–22132, 2021.
130. Wei Liang, Songyou Xie, Jiahong Cai, Chong Wang, Yujie Hong, and Xiaoyan Kui. Novel private data access control scheme suitable for mobile edge computing. *China Communications*, 18(11):92–103, 2021.
131. Md Abdur Rahman and M Shamim Hossain. A deep learning assisted software defined security architecture for 6g wireless networks: Iiot perspective. *IEEE Wireless Communications*, 29(2):52–59, 2022.
132. Yang Yang, Mulei Ma, Hequan Wu, Quan Yu, Ping Zhang, Xiaohu You, Jianjun Wu, Chenghui Peng, Tak-Shing Peter Yum, Sherman Shen, et al. 6g network ai architecture for everyone-centric customized services. *arXiv preprint arXiv:2205.09944*, 2022.
133. Jiahong Cai, Wei Liang, Xiong Li, Kuanching Li, Zhenwen Gui, and Muhammad Khurram Khan. Gtxchain: A secure iot smart blockchain architecture based on graph neural network. *IEEE Internet of Things Journal*, 2023.
134. Razvan-Andrei Stoica and Giuseppe Thadeu Freitas de Abreu. 6g: the wireless communications network for collaborative and ai applications. *arXiv preprint arXiv:1904.03413*, 2019.
135. Xiting Peng, Yichao Wang, Xiaoyu Zhang, Haibo Yang, Xiongyan Tang, and Shi Bai. A 6g-enabled lightweight framework for person re-identification on distributed edges. *Electronics*, 12(10):2266, 2023.
136. Bomin Mao, Fengxiao Tang, Zubair Md Fadlullah, and Nei Kato. An intelligent route computation approach based on real-time deep learning strategy for software defined communication systems. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1554–1565, 2021.
137. Wen Wu, Conghao Zhou, Mushu Li, Huaqing Wu, Haibo Zhou, Ning Zhang, Xuemin Sherman Shen, and Weihua Zhuang. Ai-native network slicing for 6g networks. *IEEE Wireless Communications*, 29(1):96–103, 2022.
138. Rubayet Shafin, Lingjia Liu, Vikram Chandrasekhar, Hao Chen, Jeffrey Reed, and Jianzhong Charlie Zhang. Artificial intelligence-enabled cellular networks: A critical path to beyond-5g and 6g. *IEEE Wireless Communications*, 27(2):212–217, 2020.
139. Helin Yang, Arokiaswami Alphones, Zehui Xiong, Dusit Niyato, Jun Zhao, and Kaishun Wu. Artificial-intelligence-enabled intelligent 6g networks. *IEEE Network*, 34(6):272–280, 2020.
140. Khaled B Letaief, Yuanming Shi, Jianmin Lu, and Jianhua Lu. Edge artificial intelligence for 6g: Vision, enabling technologies, and applications. *IEEE Journal on Selected Areas in Communications*, 40(1):5–36, 2021.
141. Yong Xiao, Guangming Shi, Yingyu Li, Walid Saad, and H Vincent Poor. Toward self-learning edge intelligence in 6g. *IEEE Communications Magazine*, 58(12):34–40, 2020.
142. Man Chu, Hang Li, Xuewen Liao, and Shuguang Cui. Reinforcement learning-based multi-access control and battery prediction with energy harvesting in iot systems. *IEEE Internet of Things Journal*, 6(2):2009–2020, 2019.
143. Khaled B Letaief, Wei Chen, Yuanming Shi, Jun Zhang, and Ying-Jun Angela Zhang. The roadmap to 6g: Ai empowered wireless networks. *IEEE communications magazine*, 57(8):84–90, 2019.
144. Shunliang Zhang and Dali Zhu. Towards artificial intelligence enabled 6g: State of the art, challenges, and opportunities. *Computer Networks*, 183:107556, 2020.
145. Mulumba Banza Gracia, Vusumuzi Malele, Sphiwe Promise Ndlovu, Topside Ehleketani Mathonsi, Lebogang Maaka, and Tonderai Muchenje. 6g security challenges and opportunities. In *2022 IEEE 13th International Conference on Mechanical and Intelligent Manufacturing Technologies (ICMIMT)*, pages 339–343. IEEE, 2022.
146. Marco Barreno, Blaine Nelson, Anthony D Joseph, and J Doug Tygar. The security of machine learning. *Machine Learning*, 81:121–148, 2010.

147. Jing Li, Xiaohui Kuang, Shujie Lin, Xu Ma, and Yi Tang. Privacy preservation for machine learning training and classification based on homomorphic encryption schemes. *Information Sciences*, 526:166–179, 2020.
148. Chafika Benzaid and Tarik Taleb. Ai for beyond 5g networks: a cyber-security defense or offense enabler? *IEEE network*, 34(6):140–147, 2020.
149. Jin-Hee Cho, Dilli P Sharma, Hooman Alavizadeh, Seunghyun Yoon, Noam Ben-Asher, Terrence J Moore, Dong Seong Kim, Hyuk Lim, and Frederica F Nelson. Toward proactive, adaptive defense: A survey on moving target defense. *IEEE Communications Surveys & Tutorials*, 22(1):709–745, 2020.
150. Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.
151. Deepika Sirohi, Neeraj Kumar, Prashant Singh Rana, Sudeep Tanwar, Rahat Iqbal, and Mohammad Hijjii. Federated learning for 6g-enabled secure communication systems: a comprehensive survey. *Artificial Intelligence Review*, pages 1–93, 2023.
152. Omid Aramoon, Pin-Yu Chen, Gang Qu, and Yuan Tian. Meta federated learning. *arXiv preprint arXiv:2102.05561*, 2021.
153. Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pages 273–294. Springer, 2018.
154. Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.
155. Huda A Babaeer and Saad A Al-Ahmadi. Efficient and secure data transmission and sinkhole detection in a multi-clustering wireless sensor network based on homomorphic encryption and watermarking. *IEEE Access*, 8:92098–92109, 2020.
156. Chen Fang, Yuanbo Guo, Na Wang, and Ankang Ju. Highly efficient federated learning with strong privacy preservation in cloud computing. *Computers & Security*, 96:101889, 2020.
157. Zubair Md Fadlullah, Bomin Mao, and Nei Kato. Balancing qos and security in the edge: Existing practices, challenges, and 6g opportunities with machine learning. *IEEE Communications Surveys & Tutorials*, 2022.
158. Hamed Haddadpajouh, Alireza Mohtadi, Ali Dehghantanaha, Hadis Karimipour, Xiaodong Lin, and Kim-Kwang Raymond Choo. A multikernel and metaheuristic feature selection approach for iot malware threat hunting in the edge layer. *IEEE Internet of Things Journal*, 8(6):4540–4547, 2020.
159. Jiawen Kang, Zehui Xiong, Dusit Niyato, Yuze Zou, Yang Zhang, and Mohsen Guizani. Reliable federated learning for mobile networks. *IEEE Wireless Communications*, 27(2):72–80, 2020.
160. Pawani Porambage, Tanesh Kumar, Madhusanka Liyanage, Juha Partala, Lauri Lovén, Mika Ylianttila, and Tapio Seppänen. Sec-edgeai: Ai for edge security vs security for edge ai. *The 1st 6G Wireless Summit,(Levi, Finland)*, 2019.
161. Lauri Lovén, Teemu Leppänen, Ella Peltonen, Juha Partala, Erkki Harjula, Pawani Porambage, Mika Ylianttila, and Jukka Riekki. Edgeai: A vision for distributed, edge-native artificial intelligence in future 6g networks. *6G Wireless Summit, March 24-26, 2019 Levi, Finland*, 2019.
162. Bing Tang, Feiyan Guo, Buqing Cao, Mingdong Tang, and Kuanching Li. Cost-aware deployment of microservices for iot applications in mobile edge computing environment. *IEEE Transactions on Network and Service Management*, 2022.
163. Khwaja Jawad, Khwaja Mansoor, Ahmed Fraz Baig, Anwar Ghani, and Azmat Naseem. An improved three-factor anonymous authentication protocol for wsn s based iot system using symmetric cryptography. In *2019 International Conference on Communication Technologies (ComTech)*, pages 53–59. IEEE, 2019.

164. Majid Alotaibi. An enhanced symmetric cryptosystem and biometric-based anonymous user authentication and session key establishment scheme for wsn. *IEEE Access*, 6:70072–70087, 2018.
165. Bechir Alaya and Lamaa Sellami. Clustering method and symmetric/asymmetric cryptography scheme adapted to securing urban vanet networks. *Journal of Information Security and Applications*, 58:102779, 2021.
166. Shuaiqi Shen, Chong Yu, Kuan Zhang, Jianbing Ni, and Song Ci. Adaptive and dynamic security in ai-empowered 6g: From an energy efficiency perspective. *IEEE Communications Standards Magazine*, 5(3):80–88, 2021.
167. He Fang, Xianbin Wang, Zhenlong Xiao, and Lajos Hanzo. Autonomous collaborative authentication with privacy preservation in 6g: From homogeneity to heterogeneity. *IEEE Network*, 36(6):28–36, 2022.
168. Liang Xiao, Geyi Sheng, Sicong Liu, Huaiyu Dai, Mugen Peng, and Jian Song. Deep reinforcement learning-enabled secure visible light communication against eavesdropping. *IEEE transactions on communications*, 67(10):6994–7005, 2019.
169. Martin Husák, Jana Komárková, Elias Bou-Harb, and Pavel Čeleda. Survey of attack projection, prediction, and forecasting in cyber security. *IEEE Communications Surveys & Tutorials*, 21(1):640–660, 2018.
170. Faisal Tariq, Muhammad RA Khandaker, Kai-Kit Wong, Muhammad A Imran, Mehdi Bennis, and Merouane Debbah. A speculative study on 6g. *IEEE Wireless Communications*, 27(4):118–125, 2020.
171. CL Zhang, YL Fu, H Li, et al. Research on security scenarios and security models for 6g networking. *Chinese Journal of Network and Information Security*, 7(1):28–45, 2021.
172. Shehzad Ashraf Chaudhry, Azeem Irshad, Muhammad Asghar Khan, Sajjad Ahmad Khan, Summera Nosheen, Ahmad Ali AlZubi, and Yousaf Bin Zikria. A lightweight authentication scheme for 6g-iot enabled maritime transport system. *IEEE Transactions on Intelligent Transportation Systems*, 2021.
173. Mohammad Wazid, Ashok Kumar Das, Neeraj Kumar, and Mamoun Alazab. Designing authenticated key management scheme in 6g-enabled network in a box deployed for industrial applications. *IEEE Transactions on Industrial Informatics*, 17(10):7174–7184, 2020.
174. Sabuzima Nayak and Ripon Patgiri. 6g communication: Envisioning the key issues and challenges. *arXiv preprint arXiv:2004.04024*, 2020.
175. Poonam Yadav, Angelo Feraudo, Budi Arief, Siamak F Shahandashti, and Vassilios G Vassilakis. Position paper: A systematic framework for categorising iot device fingerprinting mechanisms. In *Proceedings of the 2nd international workshop on challenges in artificial intelligence and machine learning for internet of things*, pages 62–68, 2020.

Yanlu Li is an undergraduate student at Hunan University of Science and Technology. His research interests include network measurements, network security, deep learning and blockchain.

Yufeng Xiao received the Ph.D. degree from Hunan University in 2020. He is currently an Assistant Professor at the School of Computer Science and Engineering, Hunan University of Science and Technology. His research interests include machine learning, network measurements, network security and speech signal processing.

Wei Liang received a Ph.D. degree from Hunan University in 2013. He is the Dean and Professor at the School of Computer Science and Engineering, Hunan University of Science and Technology. He has authored or co-authored more than 110 papers in top-ranked

journals/conferences. Dr. Liang's research interests include blockchain security technology, network security protection, embedded system and hardware IP protection, fog computing, and security management in wireless sensor networks (WSN).

Jiahong Cai is a Ph.D. candidate at Hunan University of Science and Technology. He has published several high-quality peer-reviewed papers in journals and conferences. His research interests include network measurements, deep reinforcement learning, network security, and blockchain.

Ronglin Zhang is a graduate student at Hunan University of Science and Technology. His research interests include network measurements, service computing, network security, deep learning and blockchain.

Kuan-Ching Li is a Life Distinguished Professor at Providence University, Taiwan. Besides publications in high-quality conferences and journals, he is a co-author or co-editor of more than 40 books published by Taylor & Francis, Springer, Elsevier, and McGraw-Hill. His research interests include Big Data, parallel and distributed computing, and emerging technologies.

Muhammad Khurram Khan is currently working as a Professor of Cybersecurity at the Center of Excellence in Information Assurance, King Saud University, Kingdom of Saudi Arabia. His research areas of interest are Cybersecurity, digital authentication, IoT security, biometrics, multimedia security, cloud computing security, cyber policy, and technological innovation management.

Received: August 04, 2023; Accepted: February 07, 2024.

Computer Science and Information Systems 21(3):899–921 https://doi.org/10.2298/CSIS230912017M

Reaching Quality and Efficiency with a Parameter-Efficient Controllable Sentence Simplification Approach

Antonio Menta and Ana Garcia-Serrano

E.T.S.I. Informática (UNED)
C. de Juan del Rosal, 14, 28040 Madrid, Spain
amenta1@alumno.uned.es
agarcia@lsi.uned.es

Abstract. The task of Automatic Text Simplification (ATS) aims to transform texts to improve their readability and comprehensibility. Current solutions are based on Large Language Models (LLM). These models have high performance but require powerful computing resources and large amounts of data to be fine-tuned when working in specific and technical domains. This prevents most researchers from adapting the models to their area of study. The main contributions of this research are as follows: (1) proposing an accurate solution when powerful resources are not available, using the transfer learning capabilities across different domains with a set of linguistic features using a reduced size pre-trained language model (T5-small) and making it accessible to a broader range of researchers and individuals; (2) the evaluation of our model on two well-known datasets, Turkcorpus and ASSET, and the analysis of the influence of control tokens on the SimpleText corpus, focusing on the domains of Computer Science and Medicine. Finally, a detailed discussion comparing our approach with state-of-the-art models for sentence simplification is included.

Keywords: Text Simplification, Transfer Learning, Language Models.

1. Introduction

The textual information that we consume daily is often presented in a style and vocabulary that varies significantly depending on the source and the particular domain area. For example, the language used on social media platforms is very different from the language used in newspapers, academic journals, or health insurance policies. According to numerous studies conducted by the Organisation for Economic Co-operation and Development (OECD), it has been found that between 15-30% of the adult population can only comprehend basic vocabulary [58]. As a result, a considerable number of adults struggle to understand the content presented in various news articles and books.

This presents a major challenge, as every individual, regardless of their literacy level, has the right to access and comprehend information that is pertinent to their lives. This is particularly important when it comes to information that has a direct impact on their daily activities, decision-making processes, and overall well-being. Furthermore, research conducted across various countries has consistently shown that adults who possess lower levels of literacy are more prone to experiencing health-related issues, less likely to engage in political processes, and generally exhibit lower levels of trust toward others.

Areas such as computer science or medicine, present additional difficulties due to a large number of neologisms and technical words [10]. Several workshops [18,19] and shared tasks [20] have been set up to deepen access to this type of technical documentation. Similarly, tools such as MedSimples [73] can evaluate text and simplify suggestions for a global audience.

Over the last twenty years, several initiatives have been launched to produce easy-to-read versions of texts, with a particular focus on the inclusion of socially disadvantaged groups in society. Unfortunately, the manual production of basic versions is an expensive and time-consuming process. In addition, the content is often outdated by the time it is published simplified version. The aim is to disseminate universally accessible versions of documents on social issues [59] to improve texts for groups with some disabilities such as aphasia [11], autism [21], or dyslexia [51].

In early work, ATS was used as a pre-processing stage for other NLP tasks such as summarization [35], parsing [12], question generation [64] or clarification [74], and more recently, as an improvement of machine translation systems for languages with limited resources [62]. Due to its relationship with machine translation, a natural step was to adapt the same techniques to generate simplified versions as a monolingual translation task [67]. Over time, the task has been divided into different categories: lexical simplification, syntactic simplification, sentence simplification and a hybrid category [3]. Lexical simplification (LS) is based on identifying the most complicated words and replacing them with simpler words [45,48], avoiding unnecessary words or information and cross references and other. Syntactic simplification attempts to reduce grammatical complexity by identifying the most complex syntactic structures and eliminating or replacing them, as for example avoiding the use of relative clauses, which introduce subordinate structures that may be difficult to understand for certain groups of people [53,54]. Sentence Simplification involves operations like sentence splitting, removal of non-essential sentence components, lexical simplification, and more.

The first automatic models were rule-based, focusing on syntactic simplification in English [12]. Subsequently, data-driven approaches were used with supervised systems trained on parallel corpora [7]. Since 2014, neural networks have been widely used for this task. Initially, using word embeddings with systems based on lexical simplification [25] and later, using sequence-sequence architectures for neural machine translation (NMT) [44,47]. In recent years, efforts have shifted towards the use of deep neural networks with large language models for sentence-level simplification systems [58,63].

Within this approach, controllable sentence simplification is a versatile method that lets you fine-tune the output of a simplification model according to certain characteristics. In this work, we continue the research line started by [37] and [57] searching for the optimal feature values for sentence simplification to reach a suitable performance when powerful resources are not available. We select well-known attributes/features and define a new control feature to help the model to produce better lexical simplifications using a controllable sentence simplification approach.

The paper is structured as follows. It begins with a section dedicated to previous work on the task, followed by a description of the controllable sentence simplification approach using neural networks in combination with linguistic attributes. This is followed by a description of the framework developed, the experiments carried out with it, and the datasets used for training and evaluation. Then, a comparison of the results with the state of the art

in the task and an analysis of the relevance of each of the features are presented to detail the contribution of the paper. Finally, some conclusions and future work are included.

2. Related Works

Early work on lexical simplification was developed for the English language using word frequencies in dictionaries and word semantic resources such as Wordnet to find simple synonyms for particular words. This approach was replaced by corpus-based methods. One example is Wikipedia. [26] used a dataset of aligned sentence pairs between English Wikipedia and Simple English Wikipedia to identify word substitutions. More recently, approaches based on BERT [15] propose substitution candidates for complex words in a sentence with high semantic similarity. For other languages, such as Spanish, several initiatives have been launched in recent years. For example, the EASIER system [2] evaluates lexical simplification for people with cognitive impairments.

Similar to the work done on lexical simplification using Wikipedia, sentence-level approaches have exploited the creation of parallel corpora between sentences extracted from Wikipedia and Simple English Wikipedia. The success of neural machine translation and the emergence of several large parallel datasets have enabled the use of deep neural network models. Early approaches were based on reinforcement learning [69] or the use of an augmented memory function known as a neural semantic encoder [65].

One of the disadvantages of the neural machine translation encoder-decoder models is that the original text and the simplified version often share many terms. Therefore, a few changes need to be made for simplification. This, combined with the fact that these operations must be learned implicitly, means that a large amount of data is required to learn the rules to be applied. In many cases, this leads to repetition of the original text because the learned patterns are not good enough [70].

Edit-based ATS systems overcome the previous limitation by learning the necessary transformations to convert complex sentences into their simpler counterparts. To do so, they use a predefined set of explicit editing operations (ADD, DELETE, MOVE, REWRITE) to transform only the complex terms during the generation of the simple sentence. Thus, the task can be approached as a sequence tagging problem using a BiLSTM-based architecture [4] or by proposing a programmer-interpreter neural network, inspired by the way humans iteratively simplify a text [16].

Subsequently, [14] added a convolutional graph module with the syntactic information of sentences to aid the detection of complex structures in the simplification process. These models are more interpretable than usual because the words to be replaced and the operation performed on them are known. Also, the search vector space of the models is small, so they are not as dependent on a large dataset as those based on linguistic models.

Figure 1 shows an example of automatic tagging based on word alignment between an original sentence (top) and a simplified sentence (bottom) in the Wikilarge dataset. Words not aligned on the original side receive the label ”D” (DELETE), while words aligned with a different shape receive ”R” (REPLACE), and added words receive an “A” (ADD). Words aligned without an explicit label are given the label “C” (COPY).

Currently, deep learning-based language models have proven to be state-of-the-art in many NLP tasks. Their effectiveness comes from prior training on a self-supervised task, such as a language model or a masked token prediction process. Once the model is



Fig. 1. Example of automatic labelling based on word alignment between an original sentence (top) and a simplified sentence (bottom). Extracted from Wikilarge dataset [69]

trained, its use in other tasks requires a much smaller data set than that needed to train a model from scratch. In particular, pre-trained Bidirectional Encoder Representations from Transformers (BERT) [15] and transfer learning techniques since ULMFit [27], have revolutionized sentence simplification task approaches.

Within this type of model, those based on an encoder-decoder architecture have excelled in tasks such as text summarisation [71] or machine translation. These include BART [34] and T5 [49]. For example, in simpleT5 [42], the authors propose the use of T5 and the union of different datasets to generate sentence simplifications. Another possibility is to include external information in the parallel corpus itself to simplify the model. An example is the BART model and the keywords provided together with the dataset [52]. In short, several works have demonstrated the transfer capabilities of pre-trained language models in the resolution of the sentence simplification task [41,61].

There are also approaches focusing on multiple languages using mT5 [68], a version trained on a massive dataset with more than a hundred languages. One example is the use of the model to simplify Arabic sentences [28]. Other work has used an unsupervised approach for each language using web mining techniques to obtain a dataset based on paraphrasing between sentences [37]. They also propose the addition of external features to the language model to control the simplification generated. This is called "conditional training". This process refers to the process of training a machine learning model to simplify sentences based on specific conditions or constraints. These conditions can include factors such as the intended audience, the level of simplification required, and the type of content being simplified.

For example, a model trained to simplify medical texts may have different conditions and constraints than a model trained with news articles. By incorporating these conditions and constraints into the training process, the model can learn to generate simplified sentences that are more appropriate for the intended use case. This can help improve the accuracy and effectiveness of sentence simplification systems, particularly in specialized domains where specific terminology and language conventions are used.

Conditional training has been applied in various NLP tasks to control text generation using sequence-to-sequence models. These tasks include text summarization [30], sentence compression [23], and text generation with specific linguistic styles [24]. The control of text generation can be achieved in the decoding phase by adjusting the output based on a specific attribute [30], or by modifying the training process through the addition of new features to the input vector, such as vector length [36]. [55] add the value of certain control features to the text to select the appropriate level among grade-level learners. [39] use grammatical attributes as model features to study their impact on simplification. Subsequently, MUSS [37], extend their work by using a massive dataset based

on paraphrasing rather than direct simplifications. In [57], they propose to use T5 as a language model to take advantage of its transfer learning capabilities. This model was pre-trained on different NLP tasks (e.g., text summarisation, question answering, machine translation) using a massive dataset such as C4[1].

In the past year, studies have begun to use Large Language Models for text simplification tasks [22]. However, the computational demands of fine-tuning these models for technical domains such as medicine have prevented many researchers from using them. The use of smaller models, but adapted to these complex domains by using a specific dataset, is therefore recommended.

3. Controllable Sentence Simplification

A controllable sentence simplification is a flexible approach that allows the output of a simplification model to be fine-tuned based on specific attributes. These attributes can be defined and customized for different use cases, such as adjusting the level of simplification for different audiences or content types. For example, a simplification model can be configured to simplify text for readers with cognitive disabilities, with output conditioned on text length and lexical complexity. Alternatively, a model can be trained to simplify news articles for readers with limited English proficiency, where output is conditioned on the amount of paraphrasing in the sentence to ensure that meaning is preserved. By incorporating controllable features into the simplification process, models can be tailored to meet specific needs and improve the effectiveness of the simplified output.

The first step in investigating the potential of language models and controllable sentence simplification in this work is to evaluate their transfer learning capabilities across different domains. The evaluation process starts with the general domain, followed by more technical vocabulary such as computer science and medical articles. The language model selected was T5, which has been pre-trained on summarisation and machine translation tasks, making it a good starting point for adding simplification features to the model. This, together with the use of a large dataset from the web, such as C4, makes it suitable for use in different domains. The T5 [49] model was also chosen because the input and output texts are in plain text format, which facilitates the addition of lexical, grammatical, and structural features without the need to modify the model's architecture. In our work, we adopt a simplified approach that involves conditioning the generation process by concatenating plain text control attributes to the source text.

Regarding T5, there are currently 5 pre-trained model versions. Depending on the number of neurons, they vary from T5-small with 60 million to T5-11B with 11 billion. In general, the greater the number of neurons the more information can be stored in the model. One of the main goals of our work is to study text simplification with language models with limited resources for use by as many researchers as possible. For this reason, we have selected the smaller model, T5-small, which can be used on most available hardware.

[1] https://www.tensorflow.org/datasets/catalog/c4

3.1. Control Attributes

Different linguistic attributes related to a simplification process are defined by [39], grouping them according to their typology:

- Text compression: Text compression is a process that involves shortening the length of sentences in simplification-oriented corpora. Typically, the length of the simplified sentences is shorter than that of the original sentences, and this compression of information results in a reduction in sentence complexity [38]. Text compression can be achieved through various techniques such as lexical substitution, deletion of non-essential words, and syntactic restructuring.
- Paraphrasing: The level of paraphrasing used in the simplification process is another attribute to highlight. If the level of paraphrasing is low, the meaning of the simplified version will be very similar to the original, and there may be no real simplification of the text. On the other hand, if the level of paraphrasing is high, there is a risk that the information in the sentence may be entirely changed, resulting in a loss of accuracy and relevance. Therefore, it is essential to select a balance between the level of paraphrasing and the level of simplification to ensure that the simplified text is both accurate and understandable.
- Lexical complexity: Lexical complexity is an attribute that indicates the level of complexity of the words used in a sentence. This attribute is based on the task of lexical simplification, which involves replacing complex words with simpler synonyms or alternative words. In technical fields with many specialized terms, substituting these terms with simple synonyms can significantly reduce the complexity of the sentence.
- Syntactic complexity: Based on the syntactic simplification task, syntactic complexity is an attribute that measures the complexity of the sentence structure in the context of syntactic simplification. The difficulty of reading a sentence can vary depending on various factors, such as the verb tense used, the use of the passive voice, or the depth of the sentence dependency tree. For example, using simpler verb tenses such as the present tense can make the sentence structure less complex and more accessible.

In this work, the above conceptual attributes are reflected in several features, the so-called control tokens, to modify the features of each sentence. The control tokens used in this work are character count, word count, Levenshtein-based similarity, dependency tree depth, word rank, and language model Fill-Mask.

Character count (CLR): This feature reflects text compression. That is, it uses the number of characters that each sentence is made of. Previous research has shown that sentence length can be a good indicator of complexity, with longer sentences generally being more complex than shorter ones [38].

Word count (WLR): [39] utilizes a feature based on the hypothesis that sentence simplicity is positively correlated with word frequency. The authors also suggest that word ratio is another important factor to consider in this context. While word frequency is typically associated with familiarity, the use of longer words can present difficulties for readers. Additionally, previous studies have found that simplified sentences tend to incorporate shorter and more frequently used words in comparison to their original versions [17].

Levenshtein-based similarity (LR): This feature is a measure of the changes made between the source and the simplification to determine their degree of paraphrase. It is calculated using Levenshtein's normalized similarity [33]. This is a measure of the similarity between two strings or sequences of characters. It is based on the Levenshtein distance, which is the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into the other.

Dependency Tree Depth (DTDR): This feature is used to add syntactic information to the model. A dependency tree is a hierarchical representation of the grammatical structure of a sentence, where the nodes represent words or groups of words, and the edges represent the syntactic relationships between them. The depth of a dependency tree is an important metric in NLP as it can be used to measure the complexity of a sentence. The deeper the dependency tree, the more complex the sentence. It is calculated by dividing the maximum dependency tree depth value of the simplified sentence by the original sentence [39].

Word Rank (WRR): In the context of lexical simplification, the word rank refers to the frequency with which a particular word appears in a given language or text. Words that have a higher rank are considered to be more common and familiar to readers, while words with a lower rank may be less familiar and more difficult for readers to understand. By considering word rank, lexical simplification tools can help writers identify and replace complex or uncommon words with simpler, more familiar alternatives to improve the overall readability of their text. Word frequency is one of the best indicators of word complexity [46].

Language Model Fill-Mask (LMFMR): This feature is based on masked language modelling. This is a type of language modeling in which certain words or tokens are masked or hidden, and the task is to predict the masked words based on the context of the surrounding words, as can be found in Table 1. We hypothesize that a language model trained on a masking task will be able to predict the set of masked words in a simple sentence earlier than in a complex sentence. In a simple sentence, the words used will be common, so the model will have been trained with them more often than with some complex and less-used words.

Table 1. Illustrates the masking process using COVID-SciBERT as a model and the SimpleText dataset [20], highlighting the dependence of the prediction model on the dataset's domain. This dataset was created to focus on both technological and medical domains simultaneously

Model	Text
Before Masking	Based on the inception - v3 architecture, our system performs better in terms of processing complexity and accuracy than many existing models for imitation learning.
COVID-SciBERT	Based on the [MASK] - v3 architecture, our system performs better in terms of processing complexity and accuracy than many existing models for imitation learning.

For every original and simplified sentence pair, features are computed by dividing the value of the simplified version by the original. This results in a value for each pair, which is then added as a prefix to the original text and displayed in Table 2 with its corresponding feature identifier. During the prediction stage, the feature values can be adjusted to cater to specific audiences, allowing for versatile feature combinations without modifying the model's structure. This input format streamlines the process by enabling pre-calculation and storage of values, which can be added as training progresses.

Table 2. Modified example with calculated features

Model	Text
Original	In the modern era of automation and robotics, autonomous vehicles are currently the focus of academic and industrial research.
Modified	simplify: CLR_0.65 DTDR_0.75 LMFMR_0.65 LR_0.65 WLR_0.8 In the modern era of automation and robotics, autonomous vehicles are currently the focus of academic and industrial research.

In our experiments, the feature values are categorized into fixed-width bins of 0.05 and are limited to a maximum ratio of 1.5. Subsequently, these features are evaluated across two datasets using different models. Selected values for the TurkCorpus dataset [67] included WLR_1.25, CLR_0.85, LR_0.65, DTDR_0.9, WRR_0.6, and LMFR_1.15. For the ASSET dataset [5], we utilized all control tokens, WLR_1.25, CLR_0.8, LR_0.65, DTDR_0.8, WRR_0.65, and LMFR_1.05.

3.2. Developed Framework for Experimentation

Language models are essential tools in natural language processing and the development of these models requires a substantial amount of data, as well as careful selection and tuning of various hyperparameters. This is because the performance of the model depends on the quality and quantity of the training data, as well as the choice of features and the model's architecture. These models can be fine-tuned on specific tasks with a smaller amount of data, enabling researchers to develop state-of-the-art models with limited resources.

The training dataset is created by incorporating selected feature values that capture the essential aspects of the sentence, such as its syntactic structure and lexical complexity. The feature values are fine-tuned using the validation dataset, which is different from the training set. The feature values vary across datasets due to distinct creation criteria, which reflect the underlying differences in the data distribution and the task requirements. The test dataset is used to evaluate the model's performance and compare it with prior work. This is an essential step, as it provides a benchmark for the model's performance and enables researchers to identify areas for improvement. The test set is typically held out from the training and validation sets and should be representative of the data distribution to ensure reliable evaluation.

We have created a Python framework for testing different models and datasets. The code is available on our GitHub repository[2] . The advantage of using this framework is that it facilitates the reproducibility of the experiments performed and allows the comparison between the results obtained with different hyperparameters. The model was trained using Pytorch-lightning and HuggingFace [62] on an Nvidia GeForce GTX 1070 Ti GPU with 8 GB of memory.

4. Benchmarking

Selecting the correct pre-trained language model is crucial for achieving accurate and effective sentence simplifications. From all available T5 pre-trained models in the HuggingFace community, we have trained with the smallest, T5-small (60 million parameters) to find a suitable solution when no powerful resources are available. Other pre-trained described in the original paper [49] are: T5-Base (220 million parameters), T5-Large (770 million parameters), T5-3B (3 billion parameters), and T5-11B (11 billion parameters).

All tests have been performed with the same hyperparameters such as 5 epochs (total number of iterations of all the training data), a batch size (the number of training examples utilized in one iteration) of 6 for both training and validation, a maximum number of 256 tokens, a learning rate (the rate at which an algorithm updates or learns the values of the parameters) of 3e-4 and a weight_decay (regularization technique applied to the weights of the neural network) of 0.1. A beam search size (decision-making layer to choose the best output given target variables) of 8 and a seed value of 22 have also been used.

These hyperparameter values are the result of an initial experimental exploration with the selected model and the training dataset. After the model was trained on the training dataset, we took advantage of the advanced capabilities of Optuna [1], an automatic hyperparameter optimization software framework, specifically designed for machine learning.

With Optuna, we conducted a meticulous hyperparameter search, performing 500 experiments to ensure the best possible results for our work. To ensure accuracy, we limited the maximum and minimum values of each control feature to 1.5 and 0.5, respectively, while using an incremental search value of 0.05. Finally, the ultimate results were obtained using the test dataset, ensuring that our model was both well-trained and well-optimized for maximum performance.

4.1. Training Dataset

Most available datasets are small, limiting their use with deep neural networks. Therefore, we chose the Wikilarge dataset [69]. This is the largest and most widely used dataset for text simplification. It consists of 296,402 automatically aligned sentence pairs extracted from the English version of Wikipedia and Wikipedia Simple. The dataset is not a parallel corpus, but a comparable corpus, because the articles in Wikipedia Simple could have been written independently of those with the same title in Wikipedia [59].

Several papers have questioned the quality of the Wikilarge dataset [8,60,66]. We use Sentence-BERT [50] to evaluate its quality. This model is a modification of the BERT architecture that uses siamese and triplet networks [56] to obtain a sentence-level embedding with a semantic representation of the sentence, that is, a transformation of the

[2] https://github.com/Hisarlik/Simplification_experiments

sentence into a vector of numbers, where sentences with similar meanings have similar numerical values [29]. This representation allows comparison with other sentences using cosine similarity. The model can be used as part of the Sentence Transformers toolkit[3] which enables the creation of high-quality, semantically meaningful sentence embeddings. The toolkit employs transformer-based neural networks to encode sentences as dense vectors, which can then be used for a variety of NLP tasks.

Other semantic similarity representations can be used to measure the similarity between words based on their meaning. Latent semantic analysis or graph-based models are some of the popular representations used for this purpose. These representations employ different techniques to capture the semantic relationships between concepts. Ontology-based similarity uses a structured knowledge base to measure the similarity formally and explicitly, which makes it useful for reasoning and inference. This measure has been explored in domains such as medicine. For example, the HESML library [32] uses an ontology-based semantic similarity measure to match related words.

In our work, we aim to improve the semantic quality of the training data used for automatic text simplification. To achieve this, we decided to select several threshold values between 0 and 1, which would allow us to generate different training datasets. For each pair of original and simplified text, we calculated the semantic difference and removed pairs with a similarity value lower than the threshold.

Table 3 summarizes the number of training documents with a similarity value higher than each of the selected thresholds. By finding the ideal threshold value, we were able to ensure that the original and simplified texts were semantically related, resulting in a reduction of the training data, but with higher semantic quality. Table 4 provides some examples of pairs that were eliminated during this process. Overall, our approach allowed us to improve the effectiveness and accuracy of automatic text simplification by enhancing the quality of the training data used.

Table 3. Pair sentence similarity using Sentence-BERT

Similarity	Pair sentences total
TOTAL	296.402
≥ 0.3	257.720
≥ 0.5	228.612
≥ 0.7	187.890
≥ 0.85	131.816
≥ 0.9	103.321

Next, we compare results obtained by selecting different similarity values between pairs of sentences in the training set. Deep learning-based models, studied in the previous section, typically perform better on larger datasets. However, we also want to test the quality of Wikilarge's automatic simplifications. For this purpose, we trained different models using each of the selected thresholds: 0.3, 0.5, 0.7, 0.85, and 0.9 similarities. We evaluated the models on the TurkCorpus dataset, using the best hyperparameters found during the validation stage and obtaining results using the test data (Table 5).

[3] https://www.sbert.net/

Table 4. Similarity ≤ 0.3 Sample deleted from the original corpus

Model	Text
Original	He is credited with many recording innovations.
Simplified	He was born in Waukesha, Wisconsin, and died in White Plains, New York.
Similarity Score	0.242623
Original	Content-control software determines what content will be available.
Simplified	In the USA, there is a filter used with TV called the V-chip.
Similarity	0.133548

Table 5. Pair sentence similarity threshold results on the TurkCorpus dataset

Similarity threshold	SARI	FKGL
ALL	42.58	5.38
≥ 0.3	42.65	5.74
≥ 0.5	42.89	5.77
≥ 0.7	42.92	6.37
≥ 0.85	43.25	6.38
≥ 0.9	43.10	6.27

Turkcorpus is a dataset in which each sentence has 8 simplifications created by Amazon Mechanical Turk workers. Each labeller was asked to paraphrase the original sentence, preserving as much meaning as possible. It contains 2000 sentence pairs for validation and 359 for testing. We also used ASSET, which uses the same data to crowdsource simplifications with a series of rewrite transformations (10 simplifications per original sentence).

In addition, to analyze the output of the models we have used a newly created dataset for the SimpleText lab in CLEF 2022 [20]. The lab focused on the challenges encountered for scientific text simplification in two domains as Computer Science and Medicine. For the former, 13 topics were selected from the Guardian newspaper headlines on the topic. For the second, the texts came from medical articles retrieved from Google Scholar and Pubmed. The texts were simplified by a master's student in translation or by an expert, either a subject matter expert or a professional translator. Each example was rewritten several times until it was clear to a non-expert. The dataset consists of 648 original-simplified sentence pairs for the validation and 116,763 for the test without the simplified version.

The results reported in the following subsections have been obtained by training the model with Wikilarge with the documents that have a similarity greater than 0.85 as explained in the training data section. From this model, we have selected the feature values in the validation dataset and the result using the TurkCorpus test set. The model used the following features: WLR_1.25, CLR_0.85, LR_0.65, DTDR_0.9, WRR_0.6, and LMFR_1.15.

For the ASSET dataset, all control tokens are used: WLR_1.25, CLR_0.8, LR_0.65, DTDR_0.8, WRR_0.65, and LMFR_1.05. Overall, by utilizing these datasets and selecting the optimal feature values, we were able to test and validate the effectiveness of our approach to automatic text simplification. These results could help improve the accessibility and readability of texts for a wider public.

4.2. Evaluation Metrics

In the field of automatic text simplification, it is essential to have reliable metrics for evaluating the effectiveness and accuracy of different models and approaches. Following [5,55,70], SARI is chosen as the main metric to evaluate the final results [67]. SARI stands for System Output Against Reference and Input, and it is a widely used automatic metric for evaluating sentence simplification.

SARI works by comparing the system output with the ground truth and the input sentence. It averages the F1 scores of additions, keeping, and deletions operations. The addition score measures the ability of the system to add new information to the output sentence while preserving the meaning of the input. The keeping score measures the ability of the system to retain the information from the input sentence in the output. Finally, the deletion score measures the ability of the system to remove redundant or unnecessary information from the input sentence.

In addition to using SARI as the primary metric for evaluating the effectiveness and accuracy of our models, we also incorporated the FKGL metric [31] in our study. FKGL, or Flesch-Kincaid Grade Level, is a widely used readability metric that measures the complexity of text based on the number of words per sentence and the number of syllables per word. Unlike SARI, which measures the semantic quality of the simplifications created, FKGL only reflects the simplicity of the text and does not consider other attributes such as the preservation of meaning or grammatical features. It is calculated using the following formula:

$$FKGL = 0.39\frac{TotalWords}{TotalSentences} + 11.8\frac{TotalSyllables}{TotalWords} - 15.59 \tag{1}$$

FKGL score is expressed as a grade level, which indicates the level of education required to understand the text. In our work, we utilized the Python EASSE library [6] to calculate the SARI and FKGL values for our models. The EASSE library is specifically designed for simplification metrics and provides a range of functions for evaluating the effectiveness and accuracy of different models and approaches to automatic text simplification.

In the succeeding section, we present the results of other approaches from their respective papers since reproducing them on our hardware was challenging. Nonetheless, we were able to replicate the results of the t5-small model [57].

4.3. Discussion

Table 6 and Table 7 present a comparative analysis of our best model against other state-of-the-art models for sentence simplification using the TurkCorpus and ASSET datasets, respectively. The models selected for comparison are:

- DMASS-DCSS [72]: This model is based on a sequence-sequence architecture with an attention mechanism. The model combines deep learning and rule-based methods. The authors propose a two-stage model that first employs a transformer architecture (DMASS) to generate multiple simplification candidates and then ranks them using a rule-based system (DCSS) to select the best paraphrase [74].

- ACCESS [39]: This sequence-sequence model incorporates an explicit attribute control mechanism for simplification. The attributes used in this model reflect text compression, paraphrasing, and lexical, and syntactic complexity.
- MUSS [37]: This BART-based model uses the attributes of the ACCESS model to perform sentence simplification. The model was trained using sentence-level paraphrase data rather than traditional simplification data. It also uses unsupervised pre-training and adjustable generation mechanisms, allowing flexible control over attributes such as length and lexical complexity during the inference process.
- Control prefixes [13]: This BART-based model employs a technique known as prefix-tuning to incorporate the representation of attributes in different layers of the model. Prefix-tuning is a powerful lightweight technique for adapting a large pre-trained language model to a downstream application. These attributes guide the generation of text toward a particular task, in this case, text simplification.
- T5-base [57]: In this paper, the authors explore the use of T5 fine-tuning for text simplification, combined with a controllable mechanism to regulate system output, which can help generate adapted text for different audiences. It includes a new feature, #words, to assist the model in replacing long words with shorter ones.
- GPT3.5 and chatGPT [22]: This approach uses OpenAI's GPT3.5 (text-davinci-003) and ChatGPT with different prompts to guide the model to a particular desired output. The model demonstrates the zero-shot transfer capabilities of LLMs. The author also uses a one-shot method with multiple combinations of examples on the model.

Table 6. Comparison of state-of-the-art models with the TurkCorpus dataset

Model	Type	Data	SARI	FKGL
DMASS-DCSS	Transformers	WikiLarge	40.45	8.04
GPT3.5 / ChatGPT	LLM	Single Shot Example	41.82	6.97
ACCESS	Transformers	WikiLarge	41.87	7.22
Control prefixes	BART	WikiLarge	42.32	7.73
MUSS – Supervised	BART	Wikilarge	42.62	6.98
T5-small (Sheang & Saggion, 2021)	T5-small	Wikilarge	40.83	6.78
T5-base (Sheang & Saggion, 2021)	T5-base	Wikilarge	**43.31**	6.17
T5-small (This work)	T5-small	Wikilarge	43.25	6.38

Upon comparing our methodology to other studies utilizing TurkCorpus, our model demonstrated nearly identical performance to the current state-of-the-art, achieving a SARI score of 43.25. This score was marginally lower than the T5-base model's 43.31 [57]. The unexpectedly low score of the LLM-based model in TurkCorpus, compared to the score it obtained in another evaluation dataset (ASSET), is noteworthy. The authors indicates that LLMs sentence simplification methods possess a penchant for excising segments of intricate sentences.

Our approach achieved promising results when evaluated on the ASSET dataset. Specifically, we obtained a SARI value of 44.56 by employing all control tokens. This result was

Table 7. Comparison of state-of-the-art models with the ASSET dataset

Model	Type	Data	SARI	FKGL
ACCESS	Transformers	WikiLarge	40.13	7.29
Control prefixes	BART	WikiLarge	43.58	5.97
MUSS – Supervised	BART	WikiLarge	44.15	6.05
T5-small (Sheang & Saggion, 2021)	T5-small	WikiLarge	39.12	6.99
T5-base (Sheang & Saggion, 2021)	T5-base	Wikilarge	45.04	5.88
GPT3.5 / ChatGPT	LLM	Single Shot Example	**47.94**	7.00
T5-small (This work)	T5-small	Wikilarge	44.56	5.76

the third-best among all models evaluated. The result of the chatGPT-based model is high compared to the other approaches, 47.94. Notably, our approach is the smallest of the models evaluated, containing almost four times fewer parameters than the T5 base. Despite this, our approach still achieves very competitive results, suggesting that it is very resource efficient. Our language model utilizes a smaller version of T5, with only 60 million parameters, compared to T5-base, which has 220 million parameters. This demonstrates the effectiveness of our approach in reducing the number of parameters required to achieve optimal results.

Our approach achieves this by eliminating the most semantically distant sentence pairs from WikiLarge and incorporating the LMFMR feature into the model. The same applies to the other work based on BART: the smallest BART model has 140 million parameters, and the largest model goes up to 400 million parameters. So, we get better results with less than half as many parameters as the BART base version.

5. Parameter-Efficient Controllable Sentence Simplification Results

Parameter-efficient controllable sentence simplification plays an important role in improving the readability and comprehension of complex texts for the wider public. By using a reduced number of parameters, these models aim to achieve meaningful and controllable text simplification without compromising overall performance. This results in a more resource-efficient approach. It requires less computing power and memory while preserving the integrity and meaning of the original content. By incorporating controllable features, these models can be fine-tuned to meet specific requirements, allowing for greater flexibility and adaptability. As a result, they can be applied to various tasks beyond text simplification, such as summarization, paraphrasing and translation.

5.1. Ablation Studies of Control Tokens

In this section, we study the influence of the control tokens we have used in the training. Each model has been trained with the reduced version of Wikilarge (a minimum similarity threshold of 0.85 between original and simplified sentences) to improve the quality of the results. A list of the most relevant training models according to the number of control tokens used can be found in Table 8 and Table 9.

To analyze the results, we began with the T5-small model without any additional tokens. Using this configuration, the model achieved a SARI value of 33.17 in TurkCorpus and 32.85 in the ASSET corpus. Next, we added each token individually and found that the best improvement with a single token was achieved with Work Rank feature (+6.96) in TurkCorpus and with Levenshtein-based similarity (+7.56) in ASSET. It's also worth noting that adding this feature resulted in an improvement of (+6.38) in TurkCorpus and adding Word Rank resulted in an improvement of (+7.03) in ASSET. These results are consistent with those published in [39]. Both features have the most significant impact on improving the Wikilarge TurkCorpus and ASSET based models.

After analyzing the TurkCorpus dataset, we discovered that Character Count feature enhanced the model's performance by (+4.25), Word Count by (+3.94), and Dependency Tree Depth by (+3.55). These results indicate that the control tokens had a similar impact on the model's performance. In comparison, the ASSET dataset showed that Word Count improved the model's performance by (+5.6), Dependency Tree Depth by (+2.96), and Character Count by (+2.94). This suggests that using only one control token can result in a significant improvement in the quality of simplifications. However, when analyzing the Language Model Fill-Mask control token on its own, the improvement achieved was only +1.3 in TurkCorpus and +0.21 in ASSET. This indicates that this feature, on its own, does not lead to a significant SARI improvement.

The incorporation of additional control tokens results in a progressive improvement in the quality of the generated output, and the best model for this work is achieved when using all control tokens. Both datasets exhibit this behaviour, with TurkCorpus achieving a SARI value of 43.25 with our model, representing a significant improvement of (+10.08) over the base model T5-small. Similarly, with ASSET, the best model obtains a SARI value of 44.56 using all the tokens.

When we analyzed the values of the control tokens for both the TurkCorpus and the ASSET datasets, we found that the values were either the same or very similar. This suggests that the control tokens have a stable effect on the performance of the model across different datasets. Furthermore, an analysis of the impact of the Language Model Fill-Mask control token used in conjunction with the other tokens reveals a slight improvement compared to models that do not use it. This denotes that the feature can be consistently incorporated into the simplification process.

5.2. Analysis of the Influence of the Control Tokens

The aim of including control tokens in the language model is to improve the quality of the output generated by the model. This section focuses on investigating whether the presence of control tokens affects the output of the model. The SimpleText corpus [20], which contains texts from the fields of computer science and medicine, was selected for this analysis.

The corpus is characterized by a large number of technical terms, making it challenging to simplify the texts for an audience without technical expertise in Computer Science or Medicine. The evaluation set was used to derive the control token values, and the sentences for the analysis were generated using the test set. The obtained control values were +WLR0.85+CLR0.75+LR0.65+DTDR0.95+WRR0.5+LMFR1.1, and the SARI value for the test set was 40.09.

Table 8. Ablation study on several trained models with different control tokens values on Wikilarge dataset (similarity≥ 0.85). SARI values were obtained on TurkCorpus

Model	SARI	FKGL
T5-small (No tokens)	33.17	10.01
+ LMFMR1.2	34.47	9.11
+DTDR0.8	36.72	9.01
+WLR0.75	37.11	8.10
+CLR0.75	37.42	7.97
+LR0.85	39.55	8.09
+WRR0.65	40.13	8.40
+WLR1.15+CLR0.75+DTDR0.8+WRR0.55+LMFR0.8	42.37	6.39
+WLR1.3+LR0.6+DTDR0.85+WRR0.75+LMFR1.2	42.62	6.78
+CLR0.95+LR0.7+DTDR0.9+WRR0.5+LMFR0.85	42.66	7.06
+WLR1.15+CLR0.9+DTDR1.45+LR0.65+LMFR1.25	42.80	6.99
+WLR1.2+CLR0.85+LR0.75+WRR0.75+LMFR1.1	43.04	6.57
+WLR1.25+CLR0.85+LR0.7+DTDR0.95+WRR0.6	43.04	6.85
+WLR1.25+CLR0.85+LR0.65+DTDR0.9+WRR0.6+LMFR1.15	43.25	6.38

Table 9. Ablation study on several trained models with different control tokens values on Wikilarge dataset (similarity≥ 0.85). SARI values were obtained on ASSET test datasets

Model	SARI	FKGL
T5-small (No tokens)	32.85	9.61
+ LMFMR1.3	33.06	9.51
+DTDR0.8	35.81	8.90
+WLR0.7	38.45	7.96
+CLR0.85	35.79	9.14
+LR0.8	40.41	7.56
+WRR0.65	39.88	8.20
+WLR1.1+CLR0.9+DTDR0.8+WRR0.8+LMFR0.95	42.25	6.50
+WLR1.05+LR0.6+DTDR0.8+WRR0.75+LMFR1.3	44.30	6.22
+CLR0.95+LR0.6+DTDR0.8+WRR0.75+LMFR1.25	44.19	6.57
+WLR1.3+CLR0.9+DTDR0.8+LR0.6+LMFR1.0	43.55	5.78
+WLR1.15+CLR0.85+LR0.65+WRR0.8+LMFR0.95	44.36	6.26
+WLR1.0+CLR0.9+LR0.65+DTDR0.8+WRR0.55	44.41	6.11
+WLR1.25+CLR0.8+LR0.65+DTDR0.8+WRR0.65+LMFR1.05	44.56	5.76

When the examples were analysed, it was found that the T5-small model often repeated the original sentence, making changes only by deleting words it considered less important. For example, it may delete information contained in parenthesis and acronyms, such as changing ”A mock PDA-based survey (PBS)” to ”A mock PDA-based survey”. In a few cases, it can change verbs, such as changing ”are emerging to become more prevalent” to ”are becoming more prevalent”. In contrast, the output generated by our model shows several features provided by the control tokens used. In all examples, the generated text is shorter than the original, indicating sentence compression. We have also included syntactic modification and paraphrasing.

One example is as follows: ”In particular, the ubiquitous access to remote resources is one of the most interesting characteristics achievable by using mobile devices such as Personal Digital Assistants, cellular phones, and tablets.”. This sentence commences with an introductory phrase, ”In particular” and has a noun phrase ”the ubiquitous access to remote resources” as its subject. In the simplified sentence the main clause is ”Using mobile devices such as Personal Digital Assistants, cellular phones, and tablets” which directly emphasizes the role of these devices in accessing remote resources. The simplification also changes the focus. The first sentence focuses on the ”interesting characteristics” achieved by using mobile devices, while the second sentence emphasizes that using these devices is ”one of the best ways” to access remote resources. This shifts the focus from the characteristics of the technology itself to the benefits it brings to the user.

One of the most frequently used features in the model is lexical substitution, where words are replaced with simpler alternatives. For instance, the word ”prevalent” is replaced with ”popular”. Similarly, ”browsed” is substituted with ”used”. Another interesting substitution is ”ambulatory setting” for ”hospital”.

Another example is the sentence: “In distributed computing paradigm, mobile surrogate systems migrates from one host in a network to another”. The term “paradigm” can be confusing for some people. Additionally, the use of the word ”migrates” implies a more technical process than is necessary to convey the idea. The simplified version removes the first term and uses the more commonly understood term “move” instead of “migrates”. This simplification clarifies the idea by removing unnecessary technical language.

6. Conclusions and Future Works

The approach we have presented utilizes a Deep Learning model known as T5 that has been pre-trained on a large corpus of text. To make the text simplification process more effective, we have incorporated several features into the model. By incorporating these features, our model can transform the complex text into a more straightforward and accessible form while still retaining the original meaning and context. Overall, our approach demonstrates the potential of leveraging pre-trained language models and transfer learning for text simplification tasks.

Our approach allows us to control the simplification result by selecting specific values for each of the previously trained features. This increases the flexibility of the system, as it is possible to generate different simplified versions, controlling, for example, the length of the text or its lexical richness. In addition, the architecture supports new features without having to modify any previous components, making it an ideal approach for experimentation. In order to train our text simplification model, we have used Wikilarge as our primary

source of data. However, we recognize that while Wikilarge is a valuable resource, it is a large and diverse dataset that may contain noise and irrelevant data. Therefore, we have created a smaller dataset with higher semantic quality. This dataset contains high-quality text that has been selected to ensure that it is relevant and suitable for text simplification. By using a smaller dataset with higher semantic quality, we aim to achieve more accurate and reliable results from our model.

To validate and test the effectiveness of our text simplification approach, we have used two different datasets: Turkcorpus and ASSET. These datasets were selected for their diversity in text complexity and their suitability for text simplification tasks. By testing our model on these datasets, we aimed to evaluate its performance and effectiveness in simplifying complex text.

Our model has demonstrated considerable success in simplifying text and achieved results that are on par with state-of-the-art approaches. It has obtained a SARI value of 43.25 on the TurkCorpus corpus and a value of 44.56 on the ASSET corpus, which are notable accomplishments in the field. Remarkably, these results were achieved using a smaller model size, with only 60 million parameters, as compared to other studies in the field. This suggests that our model is not only accurate but also more efficient and resource-friendly than some of the existing approaches, making it a promising solution for various applications requiring text simplification.

Using language models with reduced parameters is crucial for several reasons. Firstly, it ensures that these models are accessible to a wider range of users, including researchers and developers who may not have access to high-end hardware or computational resources. This democratization of AI technology allows more individuals and organizations to take advantage of language models in their work.

Moving forward, we plan to extend our approach to an additional area of interest, specifically Digital Humanities (DH). This interdisciplinary field involves the application of computational methods and tools to the study of human culture, history, and literature. The use of text simplification in historical texts is important because it can make these texts more accessible to a wider public, including students, researchers, and the general public. By simplifying complex linguistic structures and archaic vocabulary, more people can learn about and appreciate the rich cultural heritage contained in these works.

Over time, languages have experienced significant evolution, with some words and syntactic structures becoming outdated and supplanted by others. Additionally, the meanings of many words have shifted, taking on new nuances that differ from their original usage. As a result, modern readers may find it challenging to understand a literary piece from three centuries ago, due to elements such as extended sentences and the employment of particular words. Consequently, contemporary readers may struggle to comprehend a literary work from three centuries ago due to factors such as lengthy sentences and the utilization of specific words [40]. To make the simplifications, we pretend to use the same feature control tokens but using the mT5 architecture [68]. This model has been shown to give good results in text simplification approaches in Spanish [58].

Acknowledgments. This work has been partially funded by the Spanish Research Agency (AEI) through the CLARA-NLP project with sub-projects in three domains: Digital Humanities [40] (PID2020-116001RB-C32), Medicine [9] (PID2020-116001RA-C33) and Finance [43] (PID2020-116001RB-C31).

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework (2019)
2. Alarcon, R., Moreno, L., Martínez, P., Macías, J.A.: Easier system. evaluating a spanish lexical simplification proposal with people with cognitive impairments. International Journal of Human–Computer Interaction 0(0), 1–15 (2022)
3. Althunayyan, S., Azmi, A.: Automated text simplification: A survey. ACM Computing Surveys 54, Article no. 43 (03 2021)
4. Alva-Manchego, F., Bingel, J., Paetzold, G., Scarton, C., Specia, L.: Learning how to simplify from explicit labeling of complex-simplified text pairs. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 295–305. Asian Federation of Natural Language Processing, Taipei, Taiwan (Nov 2017)
5. Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., Specia, L.: ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4668–4679. Association for Computational Linguistics, Online (Jul 2020)
6. Alva-Manchego, F., Martin, L., Scarton, C., Specia, L.: EASSE: Easier automatic sentence simplification evaluation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. pp. 49–54. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
7. Alva-Manchego, F., Scarton, C., Specia, L.: Data-driven sentence simplification: Survey and benchmark. Computational Linguistics 46(1), 135–187 (2020)
8. Amancio, M., Specia, L.: An analysis of crowdsourced text simplifications. In: Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR). pp. 123–130. Association for Computational Linguistics, Gothenburg, Sweden (Apr 2014)
9. Campillos-Llanos, L., Terroba, A., Zakhir, S., Valverde-Mateos, A., Capllonch-Carrión, A.: Building a comparable corpus and a benchmark for spanish medical text simplification. Procesamiento del Lenguaje Natural 69, 189–196 (2022)
10. Campillos-Llanos, L., Valverde-Mateos, A., Capllonch-Carrión, A., Moreno-Sandoval, A.: A clinical trials corpus annotated with umls entities to enhance the access to evidence-based medicine (2021)
11. Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J.: Practical simplification of english newspaper text to assist aphasic readers. Proceedings of AAAI'98 Workshop on Integrating Artificial Intelligence and Assistive Technology pp. 7–10 (1998)
12. Chandrasekar, R., Doran, C., Srinivas, B.: Motivations and methods for text simplification. In: COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics (1996)
13. Clive, J., Cao, K., Rei, M.: Control prefixes for parameter-efficient text generation. In: Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM). pp. 363–382. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022)
14. Cumbicus-Pineda, O.M., Gonzalez-Dios, I., Soroa, A.: A syntax-aware edit-based system for text simplification. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). pp. 324–334. INCOMA Ltd., Held Online (Sep 2021)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)

16. Dong, Y., Li, Z., Rezagholizadeh, M., Cheung, J.C.K.: EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3393–3402. Association for Computational Linguistics, Florence, Italy (Jul 2019)
17. Drndarević, B., Saggion, H.: Towards automatic lexical simplification in Spanish: An empirical study. In: Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations. pp. 8–16. Association for Computational Linguistics, Montréal, Canada (Jun 2012)
18. Ermakova, L., Bellot, P., Braslavski, P., Kamps, J., Mothe, J., Nurbakova, D., Ovchinnikova, I., San-Juan, E.: Overview of simpletext clef 2021.workshop and pilot tasks. In Conference and Labs of the Evaluation Forum (2021)
19. Ermakova, L., Mothe, J., SanJuan, E.: Overview of the simpletext workshop at inforsid-2021: Scientific text simplification and popularization. In Conference and Labs of the Evaluation Forum (2021)
20. Ermakova, L., Ovchinnikov, I., Kamps, J., Nurbakova, D., Araújo, S., Hannachi, R.: Overview of the clef 2022 simpletext task 3: Query biased simplification of scientific texts. Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation, Bologna, Italy pp. 2792–2804 (2022)
21. Evans, R.J., Orasan, C., Dornescu, I.: An evaluation of syntactic simplification rules for people with autism. In: PITR@EACL (2014)
22. Feng, Y., Qiang, J., Li, Y., Yuan, Y., Zhu, Y.: Sentence simplification via large language models. ArXiv abs/2302.11957 (2023)
23. Févry, T., Phang, J.: Unsupervised sentence compression using denoising auto-encoders. In: Proceedings of the 22nd Conference on Computational Natural Language Learning. pp. 413–422. Association for Computational Linguistics, Brussels, Belgium (Oct 2018)
24. Ficler, J., Goldberg, Y.: Controlling linguistic style aspects in neural language generation. In: Proceedings of the Workshop on Stylistic Variation. pp. 94–104. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)
25. Glavaš, G., Štajner, S.: Simplifying lexical simplification: Do we need simplified corpora? In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 63–68. Association for Computational Linguistics, Beijing, China (Jul 2015)
26. Horn, C., Manduca, C., Kauchak, D.: Learning a lexical simplifier using Wikipedia. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 458–463. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014)
27. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
28. Khallaf, N., Sharoff, S., Soliman, R.: Towards Arabic sentence simplification via classification and generative approaches. In: Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP). pp. 43–52. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022)
29. Khan, A.W., Al-Obeidat, F., Khalid, A., Amin, A., Moreira, F.: Sentence embedding approach using lstm auto-encoder for discussion threads summarization. In: Computer Science and Information Systems, Vol. 20, No. 4. (2023)
30. Kikuchi, Y., Neubig, G., Sasano, R., Takamura, H., Okumura, M.: Controlling output length in neural encoder-decoders. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1328–1338. Association for Computational Linguistics, Austin, Texas (Nov 2016)

31. Kincaid, P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (1975)
32. Lastra-Díaz, J.J., Lara-Clares, A., Garcia-Serrano, A.: Hesml: a real-time semantic measures library for the biomedical domain with a reproducible survey. BMC bioinformatics 23(1), 23–23 (2022)
33. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics. Doklady 10, 707–710 (1965)
34. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880. Association for Computational Linguistics, Online (Jul 2020)
35. Li, B., Yang, P., Zhao, H., Zhang, P., Liu, Z.: Hierarchical sliding inference generator for question-driven abstractive answer summarization. ACM Trans. Inf. Syst. 41(1) (2023)
36. Mallinson, J., Sennrich, R., Lapata, M.: Sentence compression for arbitrary languages via multilingual pivoting. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2453–2464. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018)
37. Martin, L., Fan, A., de la Clergerie, É., Bordes, A., Sagot, B.: MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 1651–1664. European Language Resources Association, Marseille, France (Jun 2022)
38. Martin, L., Humeau, S., Mazaré, P.E., de La Clergerie, É., Bordes, A., Sagot, B.: Reference-less quality estimation of text simplification systems. In: Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA). pp. 29–38. Association for Computational Linguistics, Tilburg, the Netherlands (Nov 2018)
39. Martin, L., Sagot, B., de la Clergerie, É., Bordes, A.: Controllable sentence simplification. arXiv preprint arXiv:1910.02677 (2019)
40. Menta, A., Sánchez-Salido, E., García-Serrano, A.: Transcripción de periódicos históricos: aproximación clara-hd. SEPLN-PD 2022: Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations 69, 70–74 (2022)
41. Menta, A., García-Serrano, A.: Controllable sentence simplification using transfer learning. In: Simple Text Task at PAN - CLEF 2022 - Conference and Labs of the Evaluation Forum, Bolonia, Septb. 5-8, 2022. CEUR-WS, vol 3180, pp: 2818-2825. ISSN: 1613-0073 https://ceur-ws.org/Vol-3180/ (2022)
42. Monteiro, J.C., Aguiar, M., Araújo, S.: Using a pre-trained simplet5 model for text simplification in a limited corpus. In: Conference and Labs of the Evaluation Forum (2022)
43. Moreno Sandoval, A., Gisbert, A., Montoro Zamorano, H.: Fint-esp: A corpus of financial reports in spanish (01 2020)
44. Nisioi, S., Štajner, S., Ponzetto, S.P., Dinu, L.P.: Exploring neural text simplification models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 85–91. Association for Computational Linguistics, Vancouver, Canada (Jul 2017)
45. North, K., Ranasinghe, T., Shardlow, M., Zampieri, M.: Deep learning approaches to lexical simplification: A survey. arXiv (05 2023)
46. Paetzold, G., Specia, L.: SemEval 2016 task 11: Complex word identification. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 560–569. Association for Computational Linguistics, San Diego, California (Jun 2016)
47. Paetzold, G., Specia, L.: Lexical simplification with neural ranking. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume

2, Short Papers. pp. 34–40. Association for Computational Linguistics, Valencia, Spain (Apr 2017)
48. Paetzold, G.H., Specia, L.: A survey on lexical simplification. J. Artif. Int. Res. 60(1), 549–593 (sep 2017)
49. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21(1) (jan 2020)
50. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Conference on Empirical Methods in Natural Language Processing (2019)
51. Rello, L., Baeza-Yates, R.A., Dempere-Marco, L., Saggion, H.: Frequent words improve readability and short words improve understandability for people with dyslexia. In: IFIP TC13 International Conference on Human-Computer Interaction (2013)
52. Rubio, A., Martínez, P.: Hulat-uc3m at simpletext@clef-2022: Scientific text simplification using bart. In: Conference and Labs of the Evaluation Forum (2022)
53. Saggion, H., Hirst, G.: Automatic Text Simplification. Morgan & Claypool Publishers (2017)
54. Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., Drndarevic, B.: Making it simplext: Implementation and evaluation of a text simplification system for spanish. ACM Transactions on Accessible Computing 6(4) (may 2015)
55. Scarton, C., Specia, L.: Learning simplifications for specific target audiences. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 712–718. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
56. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2015)
57. Sheang, K.C., Saggion, H.: Controllable sentence simplification with a unified text-to-text transfer transformer. In: Proceedings of the 14th International Conference on Natural Language Generation. pp. 341–352. Association for Computational Linguistics, Aberdeen, Scotland, UK (Aug 2021)
58. Štajner, S., Sheang, K., Saggion, H.: Sentence simplification capabilities of transfer-based models. AAAI 36(11), 12172–12180 (2022)
59. Štajner, S.: Automatic text simplification for social good: Progress and challenges. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 2637–2652. Association for Computational Linguistics, Online (Aug 2021)
60. Štajner, S., Béchara, H., Saggion, H.: A deeper exploration of the standard PB-SMT approach to text simplification and its evaluation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 823–828. Association for Computational Linguistics, Beijing, China (Jul 2015)
61. Stajner, S., Ferres, D., Shardlow, M., North, K., Zampieri, M., Saggion, H.: Lexical simplification benchmarks for english, portuguese, and spanish (2022)
62. Štajner, S., Popović, M.: Automated text simplification as a preprocessing step for machine translation into an under-resourced language. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). pp. 1141–1150. INCOMA Ltd., Varna, Bulgaria (Sep 2019)
63. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. p. 3104–3112. NIPS'14, MIT Press, Cambridge, MA, USA (2014)
64. Tarouti, F.A., Kalita, J.K., McGrory, C.: Sentence simplification for question generation. In: Proceedings of the Conference of the North American Chapter of the ACL: Human Language Technologies. pp. 609–617 (2010)

65. Vu, T., Hu, B., Munkhdalai, T., Yu, H.: Sentence simplification with memory-augmented neural networks. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 79–85. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)
66. Xu, W., Callison-Burch, C., Napoles, C.: Problems in current text simplification research: New data can help. Transactions of the Association for Computational Linguistics 3, 283–297 (2015)
67. Xu, W., Napoles, C., Pavlick, E., Chen, Q., Callison-Burch, C.: Optimizing statistical machine translation for text simplification. Transactions of the Association for Computational Linguistics 4, 401–415 (2016)
68. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mT5: A massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 483–498. Association for Computational Linguistics, Online (Jun 2021)
69. Zhang, X., Lapata, M.: Sentence simplification with deep reinforcement learning. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 584–594. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)
70. Zhang, Y., Ye, Z., Feng, Y., Zhao, D., Yan, R.: A constrained sequence-to-sequence neural model for sentence simplification. ArXiv abs/1704.02312 (2017)
71. Zhao, H., Cao, J., Xu, M., Lu, J.: Variational neural decoder for abstractive text summarization. In: Computer Science and Information Systems, Vol. 17, No. 2, 537–552. (2020)
72. Zhao, S., Meng, R., He, D., Saptono, A., Parmanto, B.: Integrating transformer and paraphrase rules for sentence simplification. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3164–3173. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018)
73. Zilio, L., Braga Paraguassu, L., Leiva Hercules, L.A., Ponomarenko, G., Berwanger, L., Bocorny Finatto, M.J.: A lexical simplification tool for promoting health literacy. In: Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI). pp. 70–76. European Language Resources Association, Marseille, France (May 2020)
74. Zou, J., Aliannejadi, M., Kanoulas, E., Pera, M.S., Liu, Y.: Users meet clarifying questions: Toward a better understanding of user interactions for search clarification. ACM Transactions on Information Systems 41, 1 – 25 (2022)

Antonio Menta is a PhD student in Computer Science. He joined ETSI Informática - UNED in 2015 to study a Master's degree in Computer Systems and Languages. Subsequently, he started his PhD research on the application of Neural Networks in Digital Humanities. His current research uses language models to simplify Spanish newspapers from the 18th century. https://orcid.org/0000-0002-3172-2829 [1]

Ana Garcia-Serrano is a PhD in Computer Science. Associated Professor at the Artificial Intelligence Department - UPM, she joined the ETSI Informática - UNED in June 2007. Since then, she is participating in the NLP & IR consolidated research group of the university, being the technical coordinator in different research projects. For a list of her work visit http://orcid.org/0000-0003-0975-7205 [2].

Received: September 12, 2023; Accepted: February 05, 2024.

Computer Science and Information Systems 21(3):923–946 https://doi.org/10.2298/CSIS230610014P

Visit Planner: A Personalized Mobile Trip Design Application based on a Hybrid Recommendation Model

Harris Papadakis[1], Costas Panagiotakis[2], Paraskevi Fragopoulou[1], Georgios Chalkiadakis[3], Errikos Streviniotis[3], Ioannis-Panagiotis Ziogas[3], Michail Koutsmanis[3], and Panagiotis Bariamis[4]

[1] Department of Electrical and Computer Engineering,
Hellenic Mediterranean University, Estavromenos, Heraklion, Crete, Greece
adanar@hmu.gr
fragopou@ics.forth.gr
[2] Department of Management Science and Technology,
Hellenic Mediterranean University, Lakonia, Agios Nikolaos, Crete, Greece
cpanag@hmu.gr
[3] School of Electrical and Computer Engineering,
Technical University of Crete, Kounoupidiana, Chania, Crete, Greece
gchalkiadakis@tuc.gr
estreviniotis@tuc.gr
iziogas@tuc.gr
mkoutsmanis@tuc.gr
[4] Netmechanics LLC,
Solonos 23, Heraklion, Crete, Greece
panos@netmechanics.gr

Abstract. The paper presents Visit Planner (ViP), a mobile application prototype that provides a solution to the challenging tourist trip design problem. ViP follows a holistic approach offering personalized recommendations for Points of Interest (POIs) based on preferences either explicitly collected by the application, or inferred by the users' ongoing interaction with the system. ViP proposes to the final user, a trajectory of POIs calculated using an Expectation Maximization method that maximizes user satisfaction taking into consideration a variety of time and spatial constraints for both users and POIs. Additionally, POIs are divided into categories, so that a certain number of POIs from each category to be included in the final itinerary. The application is implemented as a user-interactive system that allows the flexibility for easy content adaptation and facilitates management of content and services by the user. The prototype has been implemented for Android-based smartphones, on an open application environment, using standard communication protocols and open database technology. Currently, it is applied to the city of Agios Nikolaos in Crete, and is available for download from Google play.

Keywords: Mobile application, Recommendation system, Personalized tour itinerary, Expectation maximization Content-based, Collaborative-filtering

1. Introduction

In recent years, mobile applications designed for tourism have gained significant popularity. Since many tourists rely heavily on their mobile devices to enhance their travel

experiences, the development of personalized mobile applications has become increasingly important. These applications offer targeted recommendations for POIs at a visiting site, which ultimately improve the cultural experience for the user [21, 6, 8, 9, 7]. This is particularly crucial for visitors who may have short stays in several locations, such as cruise ship tourists. Personalized mobile applications not only enhance the user's experience but also allow local communities and markets to promote their services and products in a more targeted manner, benefiting local economies. A recent comprehensive survey on the foundations and state of the art of mobile applications for tourism can be found in [23].

The designed mobile application prototype heavily relies on recommender algorithms to incorporate personalization into the final itinerary. Recommender systems aim to predict the preferences of users for items based on an analysis of preferences declared explicitly by the users or collected by the systems during previous use interactions [14]. Content-based recommender systems [15] rely solely on the user's declared preferences, while collaborative-based recommender systems [18] incorporate selections made by other users with similar preferences or similarity between items. At the same time, *Bayesian recommenders*, employ Bayesian updating of user models for efficient personalized recommendations [2, 1, 22].

The Visit Planner (ViP) application aims to provide users with the ultimate travel experience by offering personalized content and services. The application features a modern front-end that enables users to manage the content and services through an easy-to-use mobile application interface. The back-end of the designed prototype combines two innovative methods to provide users/visitors with the ultimate travel experience. The system uses a combination of four different approaches of personalized recommendation algorithms—collaborative filtering, content-based, and Bayesian—to extract from the POIs in the database those that best match the personalized interests of the traveler. After selecting the POIs that best match a user's interests using the recommender algorithms, an expectation maximization method is used to create a personalized itinerary recommendation that adheres to various real-time constraints such as opening and closing times of POIs, start/end time and other time constraints provided by the visitor, spatial locality of the POIs, budget constraints, preferred POI categories and locations, maximum trip duration, and other features such as preferred POI categories (e.g., travel, history).

This paper presents the functionality and the different components that comprise the current version of the Visit Planner App. The prototype has been currently applied to the Municipality of Agios Nikolaos in Crete, and is available though Google Play[5]. The ViP prototype can be adapted to any location by appropriately updating the content of its database module.

Among the main contributions of ViP is the fact that it incorporates an exhaustive variety of POI types, catering for every possible need and tase of the visotor (see Listing 1.3 in the Appendix). In addition, it is quite simple to use, with minimal participation from the user, while at the same time allowing for a large degree of control of the user over the created itinerary.

The remainder of the paper is organized as follows: Following the Introduction, Section 2 describes the related work. Section 3 presents the system requirements along with the user interface from the entering the user preferences and constraints to the derivation

[5] https://play.google.com/store/apps/details?id=com.netmechanics.vip

of the final personalized trip itinerary. The implementation and system architecture are presented in Section 4 including the front-end, back-end, database and middleware. Furthermore, in Section 4 the personalized recommendation algorithms used at the back-end of the application and the way they are invoked depending on each specific application scenario is described. Section 4 also presents the expectation maximization method that provides the final itinerary. We conclude in Section 5 with the main achievements and some directions for future research.

2. Related work

There is a significant number of publications on mobile applications for the tourism trip design problem, others focusing on making recommendations for POIs of interest, others employing data and user profile categorization and taxonomy data, while others related to the creation of personalized itineraries.

The authors of [16] elaborate on the challenges of designing apps for tourists through the evaluation of existing apps. They provides important implications for developers regarding application usability focusing on the parameters that guide the motivation of end users to use and reuse mobile apps that provide guided tours.

A systematic review of the personalized tourist trip design problem, also known as the orienteering problem in operations research is presented in [23]. The paper provides a review taxonomy and analyzes the main variants to the problem, the objectives, as well as the proposed solutions, an proposes threads for further research for solutions to new realistic problems. Another recent comprehensive survey on the foundations and state of the art of mobile applications for tourism can be found in [23].

In [21], a mobile application is developed that implements an algorithm for the Tourist Trip Design Orientering problem. The algorithm takes into consideration time dependencies, and analyzes in real time in combination the time constraints of the users and those of available POIs. The solution is based on a k-means algorithm, and is optimized using a genetic algorithm to improve the proposed tour itinerary. In order to facilitate recommendations, a parameterized fitness function is used to include any context element in the recommendation. The provided solution is scalable and adaptable to changes in the environment and in user preferences, thus offering a real time solution to the problem.

In [6], the authors present a hybrid planning service aiming to provide tourists with a sequence of attractions that interests them based on their previous interactions with the system. The service employs a model based recommendation system, named SCoR that operates on a synthetic coordinates principle. One of SCoR's main benefits is that it allows to incorporate additional training information on the fly without having to perform the training process from the beginning. The prototype has been implemented for Android-based smartphone and has been evaluated for St. Petersburg city. For the evaluation a database has been formed that includes attraction location information from the OpenStreeMaps platform, location description and media from Wikipedia, and ratings from Google Place.

In [8], the authors propose the ACUX recommendation system, in order to recommend POIs to visitors in a personalized manner. The proposed recommendation system replies on a collection of typologies in order to assign the visitors to one or more out of the eight available ACUX profiles. The classification is performed in order to capture the

nonverbal preferences of visitors, and to provide them with personalized suggestions of potential POIs that match their preferences.

The authors of [9] present LOOKER, a mobile application for Android devices, that implements a content-based filtering recommendation system. The system relies on tourist related content collected from the users' social media posts, to make personalized suggestions. The back-end of the application implements a multilayer user profile approach, spiting services of different kinds into layers (restaurants, hotels, POIs) to infer the interests of travelers for available items.

Other papers deal with the creation of tourist guides using innovative mobile technology. One of the first efforts is presented in [7]. This paper presents the implementation prototype of a city tourist guide generation system for the city of Mytilene, the myMytileneCity guide. The service allows tourists to declare their preferences, and based on those, a custom application is created, that is downloadable to a mobile device. Following its installation the application is fully functional without connection. A push model allows users to be signaled when new content becomes available by the administrator.

The ViP application presented in this paper follows a holistic approach including personalized recommendations, but also offering an itinerary creation component that takes a multitude of spatial and time constraints into consideration. Recommendations are based on preferences declared by the users but also on users' ongoing interaction with the service. The application allows easy content adaptation and facilitates management of content and services by the user.

3. System Requirements

The main goal of the system is to provide users with meaningful suggestions and a pleasurable itinerary regarding their short visit to the city. To this aim, we have made the following system design choices.

1. Preference specification must be as short as possible while at the same time providing necessary information for accurate and meaningful suggestions.
2. Final selection of POIs must be done by the users, in a quick and simple fashion, in order to increase their satisfaction.
3. The final output of the system will be an itinerary which will include arrival and departure times for each included POI, while satisfying the constraints set by the users.

3.1. Usage

In order to use the system, the user must perform an one time registration process in order for their profile to be created. As seen in Fig. 1 this entails the provision of an email and password as well as a few demographic information, essential to parts of the recommendation algorithms. In the "Trip Type" drop-down menu, the user specifies whether they are travelling alone, with family or friends, children or a group.

After registration, the user will also be asked to specify his/her preferences, by being asked a small number of questions. This process is described in detail in the Section 3.1. After the completion of this process and/or after each subsequent successful login, the

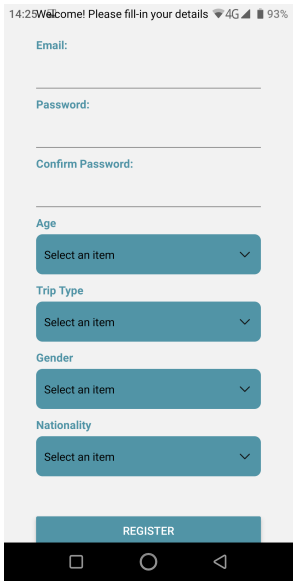


Fig. 1. New user registration form

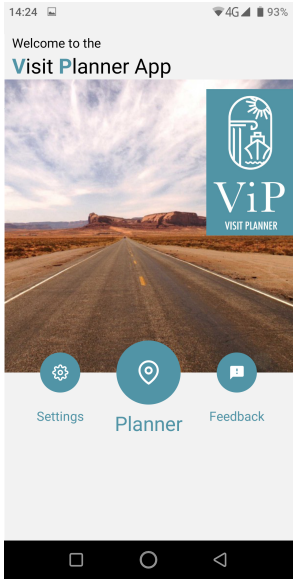


Fig. 2. Welcome screen

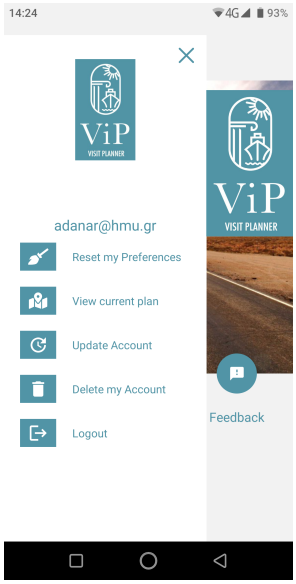


Fig. 3. Settings

user will be presented with the "Welcome" screen (Fig. 2). The available options on this screen include:

1. The "Settings" option which displays the "Settings" screen to the user, enabling them to Reset their preferences by repeating the "Specify preferences" procedure described in Section 3.1, View their most recent itinerary, Update their account by re-submitting the demographic information given during the Registration process, Delete their account from the system and finally Logout from the system.
2. The "Feedback" option which enables the user to write a textual review of the application
3. The "Planner" option where the user may begin the process of obtaining a new itinerary. This process is described in detail in Section 3.1.

Specifying preferences Depending in the recommendation algorithm that will be invoked for each user (see Section 4.4), we have implemented three different ways for the user to specify his/her preferences. Only one will be used for each user, depending on the assigned recommendation algorithm, during the registration process. However, the user can, at any time, request the repetition of the process in order to change his/her preferences (see the "Settings" menu in Section 3.1). In the first way of specifying preferences, the user is presented with a small number (10) of POI categories (not specific POIs, see Listing 1.3) and is asked to rate them (see Fig 4). In the second manner of preferences specification, the process is similar, with the only difference that the user is asked to simply specify whether they like or dislike each one of the POI categories they are presented (see Fig. 6). Finally, the third method, the user is presented with 5 different POI categories

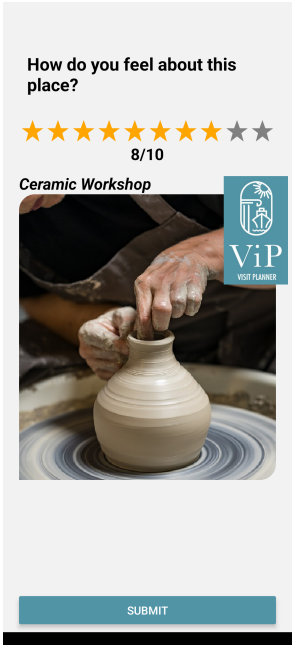


Fig. 4. POI category rating screen

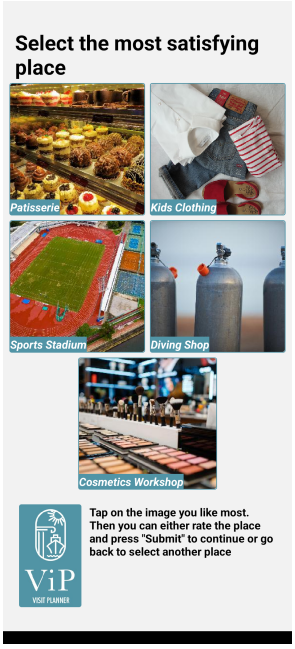


Fig. 5. POI category preference screen

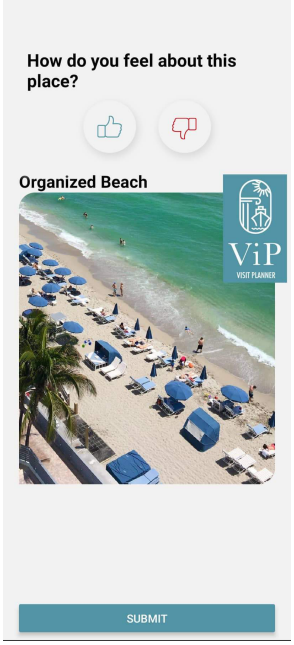


Fig. 6. POI category like/dislike screen

and is asked to choose the most preferred one (see Fig. 5). They are then asked to rate that particular POI category as per the first method (see Fig. 4).

Obtaining itinerary The process of obtaining an itinerary is the main functionality of the system. The user initiates the process by selecting the "Planner" option from the "Welcome" screen. They are then presented with a form where they are able to specify the constraints of their desired itinerary (Fig. 7). The constraints include the visit starting time, the visit duration (available time), the expenditure level (expressed in 5 different levels) as well as whether their itinerary should include venues for food.

After specifying the desired constraints and selecting the "Create Itinerary" option, the recommender module of the system will be used in order to provide 20 POIs which best fit the specified preferences of the user (see Fig. 8). The user is then asked to select which ones will be included in the itinerary. This design choice was made in order to maximize the user satisfaction by asking him/her to participate in the final selection, in a simple and quick manner. During this selection process, the user is presented at the top of the screen with the remaining time left, given their current selections. In addition, during the selection process, some of the recommended POIs may be unavailable (greyed out) depending on their compatibility with the current selections. For instance, a POI may not be compatible due to its opening hours, or because its distance to the rest of the selected POIs is too big and thus it is not feasible to be visited in the specified time constraints. The user is, of course, able to also de-select POIs in order to for instance make time for other selections.

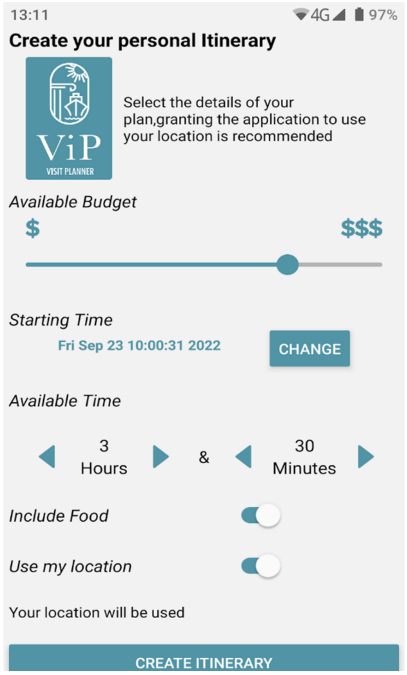


Fig. 7. Specify constraints screen

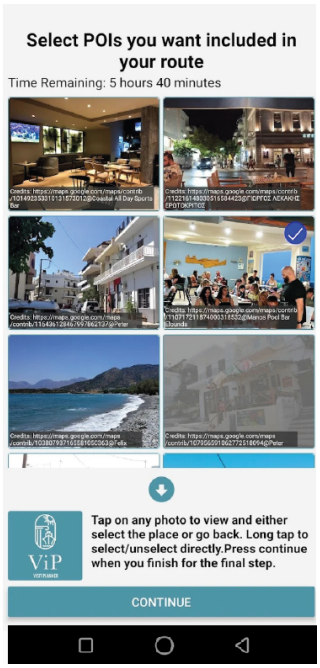


Fig. 8. POI selection



Fig. 9. Itinerary

After all selections have been made, the user will be presented with the final itinerary, using two different modes, as seen in Figs. 9 and 10. The first mode presents the sequence of the selected POIs in the most efficient order, as specified by the "Itinerary creation" module of the system, which is described in detail in Section 3.1. The mode also specifies to the user the arrival and departure times from each POI, in order for the user to be able to follow the created itinerary. Travel times by foot have been included in the computed times. Finally, the "Map" mode of the itinerary presents the user with the map of the area, where the location of the POIs that comprise the itinerary have been indicated in a numbered fashion, depending on their position in the itinerary sequence. The user is able to select each one of them and be transferred to the "Google Maps" application on their device in order to obtain the path to each POI.

It is important here to emphasize that the user is also able to rate any POIs that have been visiting by them, in order to further refine their preferences profile and obtain even more accurate POI recommendations.

4. System Architecture and Implementation

The architectural design and development of the platform should incorporate and implement the following key features:

1. Hardware should be available as a service (PaaS/SaaS).
2. Graded access depending on the type of services and the identity of the users.
3. Relational database

They should also ensure:

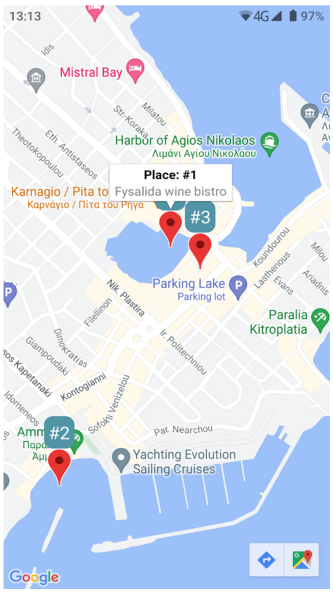


Fig. 10. Itinerary map

- An open application development environment.
- Standard communication protocols.
- Open environment in terms of database technology.
- Management of the content and services of the web portal and applications should be done by users through a simple web browser, allowing easy management & supervision.

The ViP pilot platform has been implemented with "Kubernetes" technology to be able to offer the software/platform as a service (SaaS/PaaS).

The architecture of the implemented prototype consists of four main components, illustrated in Fig. 11. These are:

1. The Front-end which is implemented as a mobile device application which interacts with the user.
2. The Back-end system which stores all necessary information.
3. The Middleware, whose purpose is to orchestrate the cooperation of the various components.
4. The Recommendations component which provides accurate and personalized POI recommendations, as well as specify the optimal selected POIs visit sequence through the Itinerary creation algorithm.
5. The Itinerary creation components that extracts the final tour trajectory.

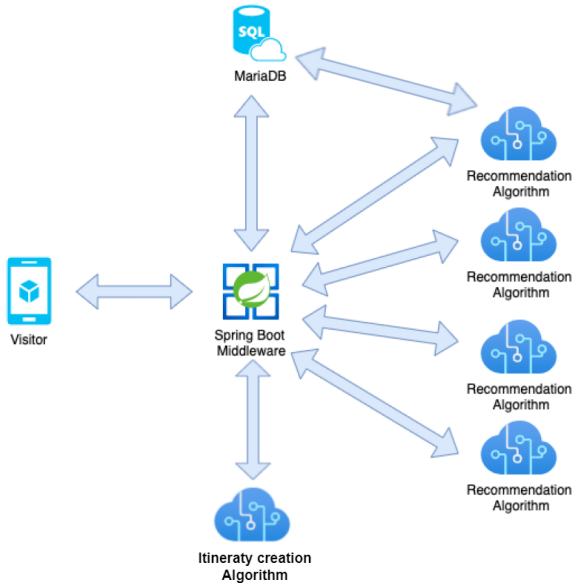


Fig. 11. System Architecture

4.1. Front-end

The front-end of the system is comprised of an android application which is available free of charge on Google Play[6].

4.2. Back-end

The Back-end of the system is comprised of a MariaDB relation database. The purpose of the database is to store all the necessary information for the smooth operation of the offered functionalities. It stores the following information, for several entities of the system:

1. User (see Listing 1.1 of the Appendix).
   The information stored for each user contains an id, their demographic information (nationality, gender, ageLevel, tripType), email as well as their specified preferences which were generated during the "Specifying preferences" procedure (see Section 3.1) as well as their ratings on specific POIs.
2. POI (see Listing 1.2 of the Appendix)
   The information stored for each POI contains its id, name, location, average rating, the categories it belongs to (see Listing 1.3), its opening hours, price Level, as well as the values of the POI's features and its users' ratings. The list of features is presented in the Listing 1.4 of the Appendix. For each POI, each one of the featured has been assigned a value between 0.0 and 1.0 which indicates the correspondence of the POI to that feature.

[6] https://play.google.com/store/apps/details?id=com.netmechanics.vip

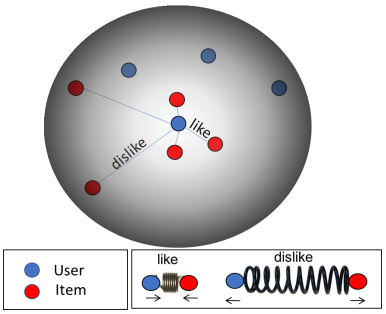


Fig. 12. A synthetic example after the execution of Synthetic Coordinates that shows the position of nodes (users and items) in $\mathbb{R}^2$. The distance between user u and item i corresponds to the prediction for the preference of user u for item i. The preference at each point of the graph for the user located in the center of the graph corresponds to the color brightness of the graph which varies from light grey (like) to dark gray (dislike)

4.3. Middleware

The Middle-ware is based on the well-known Spring Boot framework. Its main goal is to orchestrate the flow of control between the rest of the system's components, in order to implement the required functionality. It receives requests from the Front-end and serves them by contacting the rest of the modules, such as the various Recommendation algorithms that have been implemented as well as the database. All communication between components is performed through REST calls.

4.4. Recommendations component

The Recommendations module is comprised of four distinct, novel, recommender algorithms, which have been researched, developed and implemented for the needs of the system. These algorithms are summarily described in this Section. Each user is assigned one algorithm to be used in their case at all times. This is done in order to be able to evaluate the performance of each algorithm, based on the satisfaction of each user.

SCoR SCoR [13, 12] uses a Model-based Collaborating Filtering approach, which is dependent on a known set of user-to-item ratings, in order to train a preference prediction model. Thus, a number of preferences (ratings) of each user for some items (POIs) must be already known. These are provided in the form of triplets (u, i, r), where r is the scalar rating of user u for item i (POI).

In the core of SCoR lies the spring metaphor which inspired the Vivaldi synthetic network coordinate algorithm [5]. Essentially, the basis of SCoR is a Synthetic Euclidean Coordinate system, which randomly assigns a position in an N-dimensional Euclidean space to each element in the user U and the item I sets. The algorithm iteratively updates the positions of all elements (users and items) until, for every known rating (u, i, r), the Euclidean distance between user u and item i corresponds to the value r. The positions

are updated using (1), as follows:

$$p(x) = p(y) + \delta \cdot (dd(x,y) - d(x,y)) \cdot b(x,y) \tag{1}$$

$$b(x,y) = \frac{p(x) - p(y)}{d(x,y)} \tag{2}$$

where $p(x)$, $p(y)$ are the positions of a user-item pair, $d(x,y)$ is their current Euclidean distance, $dd(x,y)$ is their desired distance (based on the rating value r). The unit vector $b(x,y)$ provides the direction towards which node x should move, and δ controls the method's convergence, since it is the fraction of distance node x is allowed to move toward its ideal position. Upon algorithm conversion, the Euclidean distance between user u and an unrated (by user u) item i provides a prediction for the preference of user u for item i. Thus, after the training phase, SCoR is able to provide a recommendation $\hat{r}(u,i)$ for any given user-item pair (u,i) in $O(1)$ based on the Euclidean distance between u and i. More details about SCoR can be found in [13].

A synthetic example, after the computation of Synthetic Coordinates, is depicted in Figure 12 that shows the position of nodes (users and items). It depicts the preferences of the user located in the center of the graph and each item node of the graph via changes in the brightness of the background color varying from light gray (like) to dark grey (dislike).

Content-based Recommendations Using a Hierarchy Similarity Measure The second recommendation algorithm implemented in our system is a content-based one, that employs a hierarchy similarity measure on a well-defined hierarchy structure of POIs. Our hierarchy contains 90 generic POIs (i.e., each generic POI corresponds to one category) 430 real POIs as leafs, which belong to 90 categories, and correspond to touristic attractions of Agios Nikolaos. The categories were carefully selected given local and expert knowledge, and also the results of a survey we conducted and involved 150 real tourists visiting Agios Nikolaos. Moreover, we include in the hierarchy 90 "virtual" POIs corresponding to 90 generic images (one per category); these virtual POIs are employed for capturing user interests as follows.

In the beginning of the recommendation process, 15 out of 90 generic images (i.e., images that correspond to the generic POIs contained in our hierarchy tree) are presented to the user so as to be classified by her as "liked" or "disliked" (see Section 3.1 above). Our algorithm then computes the similarity between the "selected" generic images and all POIs, by using a modified version of a hierarchy similarity measure termed *extended Wu-Palmer similarity (XWP)* [17], which itself is an extended version of the so-called *Wu-Palmer similarity* [24]. In some detail, *XWP* takes into account the number of edges between the compared objects (i.e., POIs) and the hierarchy tree's root node, as well as the distance of the objects' Least Common Ancestor (or LCA) from the root node. Subsequently, we append the most similar POIs to the user's preferences from each generic image, and sample out a set of 20 POIs. This is the set provided to the user as the algorithm's final recommendations. The details of this approach can be found in [3].

Content-based Recommendations Using a Hybrid Similarity Measure This content-based approach takes as input data provided by the aforementioned algorithm in Section 4.4, and combines it with a novel *non-hierarchical* similarity metric, termed the

Weighted Extended Jaccard Similarity (WEJS) and introduced in [25]. As such, this approach constitutes a *hybrid* semantic similarity-based recommendation technique, which is employed to generate recommendations with respect to the user's interests.

In some detail, the WEJS metric takes into account only features of POIs which are equal in value. Each such feature is weighted based on the features of the user's preferences. WEJS is formally described in Eq. 3 below:

$$WEJS(X,Y) = \frac{\sum_{i \in X \cap Y} a * w_i}{\sum_{j \in X \cup Y} w_j}, a = \begin{cases} 1, & \text{if } v_i^X = v_i^Y \\ 0, & \text{otherwise} \end{cases}, \; with \; i \in X \cap Y \quad (3)$$

Here X and Y are the sets of the features of the two compared POIs; w_i and w_j are the weights of the ith member of the intersection of two compared sets and the jth member of their union, respectively; and v_i^X and v_i^Y denote the values of a given feature i for each set X and Y respectively, with $i \in X \cap Y$.

Assuming a maintained user profile built via the "likes" of generic POIs provided by the user (see Section 3.1 above), the use of *WEJS* on top of *XWP*-processed data, enables the focus to be placed on the actual user interests. Intuitively, *WEJS* considers a POI to be relevant to one matching the user profile if they *(a)* share the exact same characteristics (i.e., share features with the exact same values); and *(b)* these characteristics are deemed important (i.e., they are highly "weighted" by the user). The algorithm then recommends POIs that achieve high *WEJS* scores (with some exploration). The details of this approach can be found in [3].

Moreover, in [3] we extended this approach via putting forward a novel *hybrid recommender algorithm* that combines this semantic similarity-based recommender with a Bayesian Inference process to elicit user preferences. The latter will be further analyzed in the next Section.

A Bayesian Recommender Algorithm A fourth algorithm implemented in our system, is a Bayesian recommender method. To this end, every user and POI are modeled as multivariate normal distributions, based on the values of the features where POIs are concerned, whereas for users, the distributions are calculated based on questionnaires that were filled by actual tourist of Agios Nikolaos city. Notably, the preferences of each user are represented by a multivariate normal distribution and our recommender system does not posses such information. Thus, the goal of our system is to learn the preferences of each user, i.e., to generate a distribution that describes recommender's belief regarding each user's interests. For such purpose, we adopt the "*You are what you consume*" idea [1] in order to construct a representative model for each user's preferences.

To give some details, when a user enters our system, the recommender employs *Normal-Inverse Wishart (NIW)* conjugate priors to model its *beliefs* about the user's underlying parameters. NIW is a multivariate four-parameter family of continuous probability distributions, which is the *conjugate* prior of a multivariate Gaussian (normal) distribution with unknown mean and covariance matrix. Conjugacy ensures a closed-form computation of the posterior distribution. In our system, users—and also items—are modelled as *multivariate Gaussians*, hence the use of NIW priors ensures a computationally efficient Bayesian updating procedure for the user model, which is updated (in a Bayesian fashion) each time new preference information becomes available (for instance, whenever a user rates a POI that they visited) [1, 19].

Once our system has generated a distribution that describes the preferences of a specific user, it employs the Kullback-Leibler (KL) Divergence metric (see Eq. 4) in order to recommend the POIs that have the highest expected rating of a specific user.

Specifically, since both users and items share a common representation, as they are both modelled as *multivariate Gaussians*— we can employ the KL-divergence criterion in order to find "how similar" their distributions are. Formally, the KL-divergence between a Gaussian x and a Gaussian y, of dimension D each, is given by:

$$KL(y \parallel x) = \frac{1}{2} log|S_y^{-1} S_x| + \frac{1}{2} tr((S_y^{-1} S_x)^{-1}) - \frac{D}{2} + \frac{1}{2}(m_x - m_y)^T S_x^{-1}(m_x - m_y) \tag{4}$$

where S_y, m_y, S_x and m_x are the distributions' parameters, and $tr(\cdot)$ is the trace of the corresponding matrix [10]. In principle, a small KL-divergence between Gaussian x and Gaussian y implies similarity among the two, while a large KL-divergence implies they not similar. Thus, we assume that the more similar the distributions of a user u and an item i are, the higher user u rates item i. For more detailed description of the algorithm, please see [20, 19].

4.5. Itinerary creation

The proposed itinerary creation method is based on an user-interactive system that is depicted in Figs. 8, 9 and 10. The system also includes an expectation maximization technique that computes the best trajectory for the user, by maximizing an objective function that takes into consideration time, spatial, location, categories of POIs, and other constraints.

The itinerary creation is done in an iterative process so that in each step, the user selects the POI to be included in the visited set of the itinerary (see Fig. 8), from a set of unvisited and legal (valid) POIs that are automatically computed by the proposed method taking into account the user time budget and POIs opening hours and spatial constraints (user and POIs locations, etc). The ordering of the legal pre-mentioned POIs is provided to the user in a personalized way. Their sorting is done in descending order by its rating value according to the used recommender system described in Section 4.4. In each step of the method, the selected POI by the user is included to current itinerary in the position that maximizes a suitable objective function (see Eq. 5), reducing the system computational complexity. This objective function and the same methodology has been also successfully used in our recent work [11], where the itinerary was automatically created via a iterative generation process. The method terminates when the user time budget is exhausted, resulting the proposed itinerary as a timetable (Fig. 9). The user has also the option to view the itinerary in Google maps as depicted in Figure 10, respectively.

Next, we define preliminaries concerning the proposed methodology. We assume a graph (e.g. city map) with n POIs $P = \{p_1, ..., p_n\}$. Let T be the traveling time matrix ($n \times n$) of the pair-wise distances for all POI. Additionally, for each POI p_i the visit duration d_i and the opening time window o_i is known. According to the problem definition, the user provides the starting time st and the time budget B of the tour. This means that the tour itinerary should end at $st + B$ or earlier. s_i defines the gained user satisfaction per hour by visiting POI p_i. In our framework, s_i is computed offline by any of the three proposed recommendation methods described in Section 4.4. The output of our

approach is an itinerary c, which defines the visited POIs as well as the corresponding temporal information. Therefore, an itinerary c is defined by a sequence of triples, where each triple (p_i, at_i, dt_i) is comprised by the visited POI p_i with the corresponding arrival at_i and departure dt_i times. Thus, we denote by $v(c)$, the sequence of triples (p_i, at_i, dt_i) of itinerary c, for which it holds that $dt_i > at_i$.

Hereafter, we describe the objective function that is used to evaluate an itinerary in order to find the best position of the new inserted POI. The proposed objective function F that has the following properties in order to to achieve the highest user satisfaction, while respecting the given problem constraints:

- For each POI (p_i) of itinerary c, F linearly increases with the corresponding gained user satisfaction per hour s_i that is multiplied by the visit duration $dt_i - at_i$. Intuitively, the larger gained satisfaction, the more preferable the itinerary c.
- The number of visited points $|v(c)|$ slightly increases the value of the objective function, so that when two legal itineraries yield almost the same user satisfaction, the larger itinerary will be more preferable.

The aforementioned properties are well captured by defining the objective function[7] $F(c)$ as following:

$$F(c) = (1 + log(|v(c)|)) \cdot \sum_{(p_i, at_i, dt_i) \in c} s_i \cdot (dt_i - at_i) \tag{5}$$

According to the proposed methodology, the expected value of the objective function $\overline{F}(c)$ is selected to be maximized as was done in [11] yielding more robust results.

$$\overline{F}(c) = \frac{B}{dt_n - at_1} \cdot F(c) \tag{6}$$

The expected value of the objective function of the itinerary c shows the upper limit of its current value $F(c)$ taking into account that the maximum time duration of an itinerary is at most B.

5. Conclusions

The Visit Planner (ViP), a mobile application prototype for personalized trip recommendations, has been presented. ViP recommends short tour itineraries of POIs based on the personalized preferences defined by tourists and their past interaction with the application. The proposed mobile application consists of several innovative modules, including the front-end web based interface and the back-end system architecture with its various components: the recommendation and the itinerary creation modules and the relational database. The trip design system is implemented based on an expectation maximization of a suitable objective function that takes into account time, spatial, location, categories of POIs, and other constraints. The itinerary creation process allows via a user-interactive system the flexibility to select from the recommended POIs those to be included in the final itinerary. The mobile application prototype has been implemented for Android-based

[7] The value of the objective function for non legal itineraries (according to the problem constraints) is $-\infty$ [11].

smartphones, on an open application environment, using standard communication protocols and open database technology.

The ViP prototype has been successfully implemented for the city of Agios Nikolaos in Crete, and is now available for download from Google Play. This mobile application prototype is the first of its kind for the tourist trip design problem applied to a city in Greece. Thanks to its modular and adaptable design, the proposed system can be effortlessly customized for any other city without requiring any changes to the system architecture. By simply incorporating new points of interest (POIs) and relevant city information such as opening hours, travel times, and city maps, the system can be tailored to fit any city. Apart from its direct application to other cities, we aim to focus on further developing and evaluating the Visit Planner App. Our plan is to extend the main functionality of the proposed system to include group itinerary recommendations [4]. The proposed system can be further enhanced by integrating information on public transportation such as buses, taxis, and metros into the itinerary creation process. Additionally, the system can incorporate details on accommodation and additional transportation options to create comprehensive trip recommendations for longer durations spanning several days across the entire region or sub-regions of Crete.

Acknowledgments. This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH - CREATE - INNOVATE B cycle (project code: T2EDK-03135).

Appendix

Listing 1.1. User schema

```
{
  "id": 0,
  "nationality": "string",
  "gender": "string",
  "ageLevel": "string",
  "tripType": "string",
  "email": "string",
  "unencryptedPassword": "string",
  "password": "string",
  "preferences": {
    "additionalProp1": 0,
    "additionalProp2": 0,
    "additionalProp3": 0
  },
  "reviews": [
    {
      "id": 0,
      "timestamp": "2023-02-14T11:25:38.933Z",
      "rating": 0,
      "userId": "string",
      "placeId": 0
    }
  ],
  "feedback": {
    "feedbackText": "string"
  }
}
```

Listing 1.2. POI schema

```
{
  "id": 0,
  "url": "string",
  "name": "string",
  "nameEn": "string",
  "latitude": 0,
  "longitude": 0,
  "rating": 0,
  "visitorsRating": 0,
  "categories": [
    "string"
  ],
  "openingHours": "string",
  "priceLevel": 0,
  "visitDuration": 0,
  "photoReference": "string",
  "photoAttribution": "string",
  "feature": {
    "culture": 0,
    "sunAndSea": 0,
    "historyArchaeology": 0,
    "adventureSports": 0,
    "affordablePrices": 0,
    "familyFriendlyActivitiesFacilities": 0,
    "ruralTourism": 0,
    "luxuryAccommodationLeisure": 0,
    "nightlife": 0,
    "gastronomyCuisine": 0,
    "generalShopping": 0,
    "localProductsShopping": 0
  },
  "visitorReviews": [
    {
      "id": 0,
      "timestamp": "2023-02-14T11:53:51.148Z",
      "rating": 0,
      "userId": "string",
      "placeId": 0
    }
  ]
}
```

Listing 1.3. POI categories

```
Bookstore
Library
Artwood Gallery
Artwork Gallery
Ceramics Gallery
Indoor Movie Theater
Outdoor Movie Theater
Archaeological Site
Historical Site
Monument
Archaeological Museum
Vintage Vehicle Museum
Historical Museum
Heritage Museum
Plant Museum
Chapel
Church
Historic Church
Monastery
Medieval Byzantine
Park
Square
Cape
Gorge
```

Grove
Hill
Mountain Village
Mountaintop
Bay
Non Organized Beach
Organized Beach
Bridge
Canal
Harbor
Island
Lake
Marina
River
Cafe Bar
Beach Bar
Cocktail Bar
Music Bar
Nightlife Club
Pool Bar
Pub
Rock Bar
Sports Bar
Wine Bar
Classic Cafeteria
Playground Cafeteria
Take Away Cafeteria
Tourist Cafeteria
Traditional Cafeteria
Bakery
Creperie
Ice Cream
Patisserie
Street Food
Asian Restaurant
Casual Greek Restaurant
Fine Greek Restaurant
Italian Restaurant
Greek Tavern
Mezedopoleio
Seafood Tavern
Flowers Shop
Cosmetic Store
Liquor Store
Diving Shop
Furs Store
Men Casual
Men Sports
Men Shoes
Women Casual
Women Sports
Women Shoes
Kids Clothing
Kids Shoes
Handmade Jewellery
Jewelry
Gift Shop
Traditional Products Shop
Ceramic Workshop
Cosmetics Workshop
Wood Workshop
Farm Tour
Boat Tour
Sports Stadium
Water Sports
Spa

Listing 1.4. POI Features

```
Culture
Sun&Sea
History/Archaeology
Adventure/Sports
Affordable Prices
Family Friendly Activities/Facilities
Rural Tourism
Luxury Accommondation/Leisure
Nightlife
Gastronomy/Cuisine
General Shopping
Shopping Local Products
```

References

1. Babas, K., Chalkiadakis, G., Tripolitakis, E.: You are what you consume: A bayesian method for personalized recommendations. In: Proceedings of the 7th ACM Conference on Recommender Systems. p. 221–228. RecSys '13, Association for Computing Machinery, New York, NY, USA (2013)
2. Barbieri, N., Costa, G., Manco, G., Ortale, R.: Modeling item selection and relevance for accurate recommendations: a bayesian approach. In: Mobasher, B., Burke, R.D., Jannach, D., Adomavicius, G. (eds.) 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011. pp. 21–28. ACM (2011)
3. Chalkiadakis, G., Ziogas, I., Koutsmanis, M., Streviniotis, E., Panagiotakis, C., Papadakis, H.: A novel hybrid recommender system for the tourism domain. Algorithms submitted for review (2023)
4. Chen, L., Cao, J., Chen, H., Liang, W., Tao, H., Zhu, G.: Attentive multi-task learning for group itinerary recommendation. Knowledge and Information Systems 63(7), 1687–1716 (2021)
5. Dabek, F., Cox, R., Kaashoek, F., Morris, R.: Vivaldi: A decentralized network coordinate system. ACM SIGCOMM Computer Communication Review 34(4), 15–26 (2004)
6. Kashevnik, A.M., Mikhailov, S., Papadakis, H., Fragopoulou, P.: Context-driven tour planning service: An approach based on synthetic coordinates recommendation. In: 24th Conference of Open Innovations Association, FRUCT 2019, Moscow, Russia, April 8-12, 2019. pp. 140–147. IEEE (2019)
7. Kenteris, M., Gavalas, D., Economou, D.: An innovative mobile electronic tourist guide application. Pers. Ubiquitous Comput. 13(2), 103–118 (2009)
8. Konstantakis, M., Christodoulou, Y., Aliprantis, J., Caridakis, G.: ACUX recommender: A mobile recommendation system for multi-profile cultural visitors based on visiting preferences classification. Big Data Cogn. Comput. 6(4), 144 (2022)
9. Missaoui, S., Kassem, F., Viviani, M., Agostini, A., Faiz, R., Pasi, G.: LOOKER: a mobile, personalized recommender system in the tourism domain based on social media user-generated content. Pers. Ubiquitous Comput. 23(2), 181–197 (2019)
10. Nielsen, F., Nock, R.: Emerging Trends in Visual Computing: LIX Fall Colloquium, ETVC 2008, Palaiseau, France, November 18-20, 2008. Revised Invited Papers. Springer Berlin Heidelberg, Berlin, Heidelberg (2009)
11. Panagiotakis, C., Daskalaki, E., Papadakis, H., Fragopoulou, P.: Personalized itinerary recommendation via expectation-maximization. In: 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT). pp. 210–213. IEEE (2022)
12. Panagiotakis, C., Papadakis, H., Papagrigoriou, A., Fragopoulou, P.: Improving recommender systems via a dual training error based correction approach. Expert Systems with Applications 183, 115386 (2021)
13. Papadakis, H., Panagiotakis, C., Fragopoulou, P.: Scor: A synthetic coordinate based system for recommendations. Expert Systems with Applications 79, 8–19 (2017)

14. Papadakis, H., Papagrigoriou, A., Panagiotakis, C., Kosmas, E., Fragopoulou, P.: Collaborative filtering recommender systems taxonomy. Knowl. Inf. Syst. 64(1), 35–74 (2022)
15. Pasquale Lops, M.d.G., Semeraro, G.: Content-based recommender systems: State of the art and trends. In: Recommender Systems Handbook, pp. 73–106. Springer-Verlag, Berlin, Heidelberg (2010)
16. Podsukhina, E., Smith, M.K., Pinke-Sziva, I.: A critical evaluation of mobile guided tour apps: Motivators and inhibitors for tour guides and customers. Tourism and Hospitality Research 22(4), 414–424 (2022)
17. Shenoy, M.K., Shet, K., Acharya, U.D.: A new similarity measure for taxonomy based on edge counting. International Journal of Web & Semantic Technology 3(4), 23 (2012)
18. Shi, Y., Larson, M., Hanjalic, A.: Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. ACM Comput. Surv. 47(1), 3:1–3:45 (May 2014), http://doi.acm.org/10.1145/2556270
19. Streviniotis, E., Chalkiadakis, G.: Multiwinner election mechanisms for diverse personalized bayesian recommendations for the tourism domain. In: Neidhardt, J., Wörndl, W., Kuflik, T., Goldenberg, D., Zanker, M. (eds.) Proceedings of the Workshop on Recommenders in Tourism (RecTour 2022) co-located with the 16th ACM Conference on Recommender Systems (RecSys 2022), Seattle, WA, USA and Online, September 22, 2022. CEUR Workshop Proceedings, vol. 3219, pp. 65–82. CEUR-WS.org (2022)
20. Streviniotis, E., Chalkiadakis, G.: Preference aggregation mechanisms for a tourism-oriented bayesian recommender. In: Aydogan, R., Criado, N., Lang, J., Sánchez-Anguix, V., Serramia, M. (eds.) PRIMA 2022: Principles and Practice of Multi-Agent Systems - 24th International Conference, Valencia, Spain, November 16-18, 2022, Proceedings. Lecture Notes in Computer Science, vol. 13753, pp. 331–346. Springer (2022)
21. Tenemaza, M., Luján-Mora, S., de Antonio, A., Ramírez, J.: Improving itinerary recommendations for tourists through metaheuristic algorithms: An optimization proposal. IEEE Access 8, 79003–79023 (2020)
22. Tripolitakis, E., Chalkiadakis, G.: Probabilistic topic modeling, reinforcement learning, and crowdsourcing for personalized recommendations. In: Pacheco, N.C., Carrascosa, C., Osman, N., Inglada, V.J. (eds.) Multi-Agent Systems and Agreement Technologies - 14th European Conference, EUMAS 2016, and 4th International Conference, AT 2016, Valencia, Spain, December 15-16, 2016, Revised Selected Papers. Lecture Notes in Computer Science, vol. 10207, pp. 157–171. Springer (2016)
23. W. Wörndl, D.H.: Mobile applications for e-tourism. Handbook of e-Tourism, Springer (2020, Springer)
24. Wu, Z., Palmer, M.: Verb semantics and lexical selection. arXiv preprint cmp-lg/9406033 (1994)
25. Ziogas, I.P., Streviniotis, E., Papadakis, H., Chalkiadakis, G.: Content-based recommendations using similarity distance measures with application in the tourism domain. In: 12th SETN Conference on Artificial Intelligence. pp. 1–10 (2022)

Harris Papadakis is an Assistant Professor of the Department of Electrical and Informatics Engineering of the Hellenic Mediterranean University in the field of "Distributed Services and Networks". He holds a PhD from the Department of Computer Science of the University of Crete, a degree in Computer Science from the Department of Computer Science of the University of Crete and a postgraduate MSc diploma from the Department of Computer Engineering of the University of Patras. He has participated as researcher and head in several projects.

Costas Panagiotakis is Associate Professor, Director of Data Science, Multimedia and Modelling Laboratory (DataLab) and Head of the Department of Management Science and Technology, HMU. He has been involved in several R&D projects funded by the EC during the last 23 years in the fields of computer vision, pattern recognition and recommender systems. He is also a researcher at the Computational Vision and Robotics Laboratory, Institute of Computer Science, Foundation for Research and Technology-Hellas.

Paraskevi Fragopoulou obtained her BA from the Computer Science Department of the University of Crete in 1989 and M.Sc. and Ph.D. in computer science in 1991 and 1995, respectively, from the Department of Computing and Information Sciences, Queen's University, Kingston, Ontario, Canada. She is a Professor at the Department of Electrical and Informatics Engineering of the Hellenic Mediterranean University (EL.ME.PA). She has worked in research positions at various European Research Institutes. She is a collaborating faculty member at the Institute of Technology and Research (ITE), Heraklion, Crete.

Georgios Chalkiadakis joined the Technical University of Crete in March 2011. Before coming to TUC, he was a Research Fellow at the School of Electronics and Computer Science, University of Southampton, UK, from December 2006 to January 2011. Georgios received his PhD from the University of Toronto, Canada, in 2007. His PhD thesis was a nominee for the IFAAMAS Victor Lesser Distinguished Dissertation Award (2007). Georgios has also worked as a software engineer at the Institute of Computer Science of the Foundation for Research and Technology - Hellas.

Errikos Streviniotis completed his undergraduate studies and acquired a diploma degree at the School of Electrical and Computer Engineering at the Technical University of Crete (December 2020). He acquired an M. Sc. in Electronic and Computer Engineering at the Technical University of Crete (December 2022) under the supervision of Professor Georgios Chalkiadakis. Currently, Errikos is a PhD candidate at the same institution under the supervision of Assistant Professor Nikos Giatrakos. In parallel to his studies, Errikos works as a research associate at the 'Athena' Research Center. His primary research focuses on Artificial Intelligence, Machine Learning, and Game Theory.

Ziogas Ioannis Panagiotis completed his undergraduate studies and acquired a diploma degree at the School of Electrical and Computer Engineering at the Technical University of Crete (February 2020). Acquired the M. Sc. in Research in Electronic and Computer Engineering at the Technical University of Crete (March 2023) under the supervision of Professor Georgios Chalkiadakis.

Michail Koutsmanis is currently a Ph.D. student at the School of Electrical and Computer Engineering in Technical University of Crete (since August 2022) under the supervision of Associate Professor Georgios Chalkiadakis. He earlier completed his undergraduate studies and acquired a diploma degree at the Department of Electrical and Computer Engineering in Aristotle University of Thessaloniki (July 2021). His research interests lie mainly on Bounded Rationality and Game Theory along with their implementation in Multi-agent Systems.

Panagiotis Bariamis received his BSc from the Technological and Educational Institute of Athens, in Information Technology. He is a full-time software engineer since 2003, currently working for Netmechanics S.A.

Received: June 10, 2023; Accepted: February 15, 2024.

Computer Science and Information Systems 21(3):947–969 https://doi.org/10.2298/CSIS231206022X

ECW-EGNet: Exploring Cross-Modal Weighting and Edge-Guided Decoder Network for RGB-D Salient Object Detection

Chenxing Xia[1,2,3], Feng Yang[1,⋆], Songsong Duan[4], Xiuju Gao[5], Bin Ge[1], Kuan-Ching Li[6], Xianjin Fang[1,7], Yan Zhang[8], and Ke Yang[2,⋆]

[1] College of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, 232001, China
2021201237@aust.edu.cn
[2] The First Affiliated Hospital of Anhui University of Science and Technology, (Huainan First People's Hospital), China
cxxia@aust.edu.cn
[3] Anhui Purvar Bigdata Technology Co. Ltd, Huainan, 232001, China
[4] State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China
[5] College of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan, Anhui, China
[6] Department of Computer Science and Information Engineering, Providence University, Taiwan
[7] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
[8] The School of Electronics and Information Engineering, Anhui University, Hefei, Anhui, China

Abstract. Existing RGB-D salient object detection (SOD) techniques concentrate on combining data from multiple modalities (e.g., depth and RGB) and extracting multi-scale data for improved saliency reasoning. However, they frequently perform poorly as a factor of the drawbacks of low-quality depth maps and the lack of correlation between the extracted multi-scale data. In this paper, we propose a Exploring Cross-Modal Weighting and Edge-Guided Decoder Network (ECW-EGNet) for RGB-D SOD, which includes three prominent components. Firstly, we deploy a Cross-Modality Weighting Fusion (CMWF) module that utilizes Channel-Spatial Attention Feature Enhancement (CSAE) mechanism and Depth-Quality Assessment (DQA) mechanism to achieve the cross-modal feature interaction. The former parallels channel attention and spatial attention enhances the features of extracted RGB streams and depth streams while the latter assesses the depth-quality reduces the detrimental influence of the low-quality depth maps during the cross-modal fusion. Then, in order to effectively integrate multi-scale features for high-level and produce salient objects with precise locations, we construct a Bi-directional Scale-Correlation Convolution (BSCC) module in a bi-directional structure. Finally, we construct an Edge-Guided (EG) decoder that uses the edge detection operator to obtain edge masks to guide the enhancement of salient map edge details. The comprehensive experiments on five benchmark RGB-D SOD datasets demonstrate that the proposed ECW-EGNet outperforms 21 state-of-the-art (SOTA) saliency detectors in four widely used evaluation metrics.

Keywords: cross-modality fusion, depth-quality, edge-guided, RGB-D images, salient object detection.

⋆ Corresponding authors

1. Introduction

Salient object detection (SOD), a crucial component of computer vision, tries to mimic the visual learning mechanism of the human body to react fast to visual stimuli and produce visually appealing or fascinating objects or regions by exploring and segmenting particular objects. As a valuable preprocessing tool, it has been used extensively in various computer vision applications, including semantic segmentation [1], image retrieval [2] , visual tracking [3], and remote sensing image segmentation [4]. The technique can perform a series of image processing operations and speed up the processing of data by intelligently, effectively, and precisely recognizing the salient object regions. It can also effectively distribute the limited computing resources to some extent.

In traditional RGB-based SOD, the color and texture cues captured by RGB images play a crucial role in identifying salient objects. However, their performance suffers when faced with difficult situations such complex backgrounds and poor illumination. As a result of the development of sensors like smartphones and Microsoft Kinect depth camera, it is now possible to quickly gather both RGB maps and equivalent depth maps from a scene. The spatial information and geometry contained in the depth map can be used to offset the shortcomings of the RGB image, which provides detailed information. As a result, it is a reasonable choice to combine RGB images and depth maps for SOD tasks (called RGB-D SOD), which can handle more complicated scenarios and satisfy the needs of advanced detection. However, directly integrating RGB images and depth images is not a good countermeasure and introduces unique challenges. These challenges include: (1) Depth noise: Depth maps acquired from depth sensors can be affected by noise and inaccuracies, which can impact the quality and accuracy of saliency maps. (2) Complementary fusion: Integrating RGB and depth information to capture complementary cues and avoid redundancy is crucial. Developing effective fusion strategies that fully utilize the strengths of each modality is a challenge.

The depth map is generally used as prior knowledge to assist handcrafted features for SOD in the early stages of RGB-D SOD approaches. Unfortunately, early handcrafted feature techniques were really crude, frequently unable to represent data with rich high-level semantics, and their performance was subpar. In past decades, feature representation in convolutional neural networks (CNNs) has grown functional because CNNs can acquire features from the geometry, color, and spatial information of images, causing RGB-D SOD technology to pay greater attention to deep learning. Existing approaches focus on how to design effective ”interaction/fusion” modules that bridge the gap between the two modes. Wei *et al*. [5] sufficiently melded the bidirectional attention interaction module achieves bidirectional interaction between cross-modality features. Wang *et al*. [6] converted the original 4-dimensional RGB-D cycle into DGB, RDB, and RGD to fuse RGB and depth before extracting depth features, which helps achieve the best channel complementary fusion state between RGB and depth. Chen *et al*. [7] adaptively fusion of multi-modal information makes use of complimentary RGB and depth cues to address the issue of inaccurate depth maps. Xia *et al*. [8] introduced a bi-directional interactive architecture in order to improve RGB and depth features through a circular interaction. Bi *et al*. [9] proposed a feedback mechanism that receives two input features from RGB and depth branches to exchange would exchange the new information into existing interactive information. Zeng *et al*. [10] propose a compensated attention feature fusion and hierarchical multiplication decoder network. Although these models have achieved remarkable

results in exploring feature fusion between modalities, they ignore the possible negative effects caused by the introducing low-quality depth maps in the fusion process.

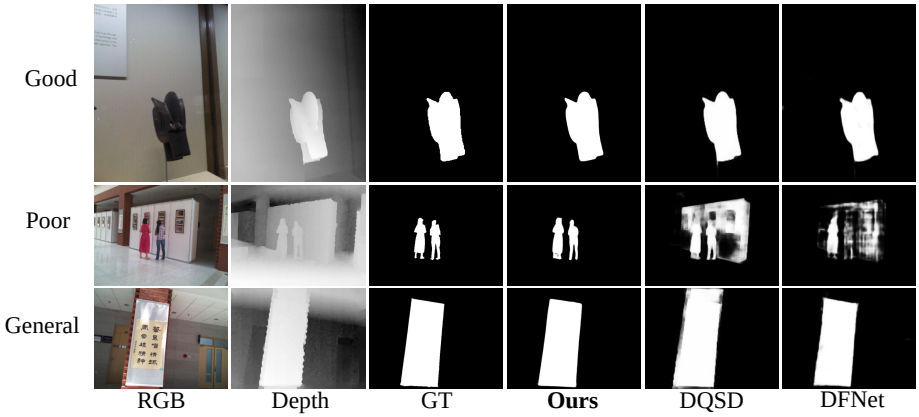


Fig. 1. Visual examples of our method, DQSD [11] and DFNet [12] of depth maps with Good, Poor, and General (top-to-bottom)

In order to avoid the contamination of low-depth maps, some works [11,13,14] develop depth awareness, a guidance method of fusing cross-modal data. On the other hand, some researchers [12,15] enhance depth maps and suppress noise from depth maps using feature enhancement modules. In order to improve the accuracy, a growing number of techniques [16,17] are devoted to presenting elaborate feature fusion modules, which include well-designed alternate interactions between features to filter out noise in RGB and depth features with the help of other modal data. Furthermore, several methods [18,19] exploit estimates supplementary meaning depth maps to lessen the impact of low-quality depth maps. However, there are some problems with these approaches: (1) Feature enhancement and elaborate feature fusion mainly focus on complementary and ignore the specificity between different modalities; (2) Depth-aware and depth estimation methods undoubtedly increase the computational cost by estimating the depth map; (3) The performance of the SOD is significantly impacted by effective mechanisms for assessing the quality of depth maps. Some visual examples are shown in Figure 1, where the depth map quality is classified as good, poor, and general. In these cases, features extracted from low-quality input images reduce the discriminative ability of fused features. Therefore, it is important to consider the quality of depth maps in the RGB-D SOD task.

Furthermore, it is important to note that SOD is fundamentally a pixel-wise dense prediction work, which utilizes the categorization of pixels on feature maps by relying on high-resolution and multi-scale features. However, the latest advances are mainly based on atrous convolution techniques, which enhance high-level features through the integration of semantic segmentation modules. One of the most well-known modules is the ASPP and its variant versions [20,21,22,23], which adjust the receptive field size and enhance the expressive power of the feature map by controlling the atrous rate without introducing additional parameters. Although these models have achieved remarkable results, they

still have two limitations: (1) ***The loss of semantic information***. The calculation of atrous convolutions is similarly to a checkerboard format, where the convolution kernel only focuses on certain positions of the pixels in the feature map during computation, while other positions are replaced by far "zero elements". This means that the output results are determined only by pixels in a few positions, which may lead to the loss of semantic information. (2) ***The lack of information correlation***. These methods capture a wide range of contextual information by varying the atrous convolution. However, in practical applications, the feature representations across different scales often exhibit correlations. Separate multi-scale pooling operations on the outputs of the deep network do not adequately consider this correlation, which may also be detrimental to the generation of high-quality salient maps.

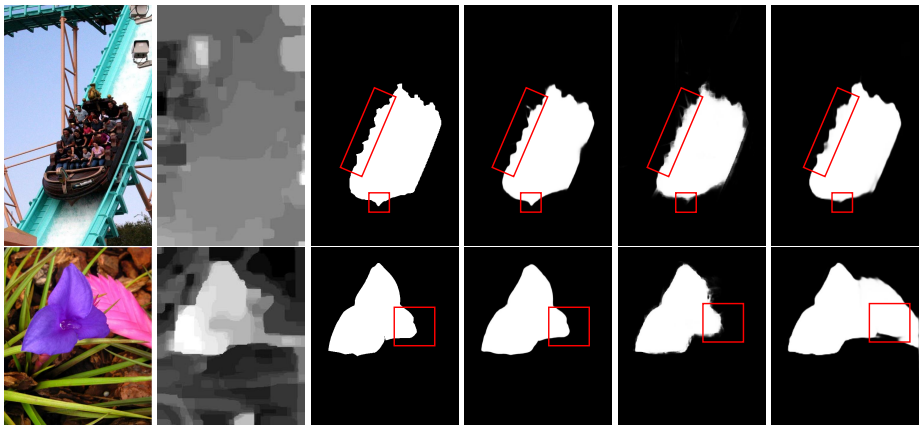


Fig. 2. Compare the results of our model with JL-DCF [24] and SMEG [25]. The red anchor box represents the comparison area

Moreover, edge features refer to the information around objects in an image, which can provide important clues about object shape and structure. In SOD, these edge features are crucial for distinguishing salient objects from the background, as salient objects typically have distinct contours and edge features that differ from the surrounding background. By effectively extracting and analyzing the edge features of an image, salient objects can be more accurately identified and separated them from the background. However, most of the past research has primarily focused on the structural integrity of SOD results rather than edge quality [26]. As a result, many existing approaches suffer from poor edge quality in their output saliency maps. As shown in the first row of Figure 2, when performing SOD tasks in simple scenes, JL-DCF [24] and SMEG [25] can effectively extract objects from the background. However, the edge details captured by these models appear slightly rougher compared to those produced by the proposed ECW-EGNet method. This observation highlights the need for accurate edge localization and the incorporation of

edge attention features to optimize the performance of RGB-D SOD algorithms, thereby mitigating the issue of edge blurring, as shown in the red box in the figure.

To solve the above problems, a Exploring Cross-Modal Weighting and Edge-Guided Decoder Network (ECW-EGNet) is proposed for RGB-D SOD, which is equipped with a cross-modal weighting fusion (CMWF) module, a bi-directional scale-correlation convolution (BSCC) module, and an edge-guided decoder. Specifically, the CMWF module is designed to enhance the complementary between the cross-modal features and better integrate the features, where Channel-Spatial Attention Feature Enhancement (CSAE) mechanism and depth-quality assessment (DQA) mechanism are embedded to enhance the discriminant feature and adaptively weights and refines the RGB and depth branches, respectively. In addition, to exploit the property of fused multi-scale features, the BSCC module is constructed to better capture scale-correlation information and integrate multi-scale features of high-level features by employing depth-wise separable convolutions to compensate for the atrous convolution, where each feature branch is connected in series to fill the holes and further improve the quality of saliency maps. Finally, to enhance the edge information, an Edge-Guided (EG) decoder is proposed to synthesize more accurate boundary masks by means of an edge detection operator, thus guiding the decoder to achieve more accurate image segmentation.

The main contributions of this paper include the following:

- We suggest a Exploring Cross-Modal Weighting Fusion and Edge-Guided Decoder Network (ECW-EGNet) for RGB-D SOD that not only assesses the quality of the depth map but also utilizes the feature attributes of different scales and edge-enhancement features to improve the efficiency of generating high-quality saliency maps.
- We propose a Cross-Modal Weighting Fusion (CMWF) module to facilitate the learning of complementary information across modalities, in which a Channel-Spatial Attention Feature Enhancement (CSAE) mechanism is employed to enhance the features of extracted RGB streams and depth streams, and a Depth-Quality Assessment (DQA) mechanism is embedded to assess the quality of the feature of the depth map by calculating the weighting factors to reduce the interference problem caused by the low-quality of the depth map.
- We construct a Bi-directional Scale-Correlation Convolution (BSCC) module based on a well-designed bi-directional structure to capture the complementary and correlated information of high-level features from different scales and reduce the loss of local information.
- We propose an Edge-Guided (EG) decoder that guides the upgrading of salient object edges by using an edge detection operator to synthesize more accurate edge masks.

2. Materials and Methods

2.1. Network Overview

The overall framework of the proposed ECW-EGNet is shown in Figure 3. We adopt the Swin Transformer [27] as the backbone of a two-stream architecture to acquire multi-level feature representation via RGB and depth maps. Swin Transformer has a powerful ability to model far range dependencies and capture global context. Formally, the generated from the RGB and depth streams are named as $\{f_i^r\}_{i=1}^4$ and $\left\{f_i^d\right\}_{i=1}^4$, respectively. The

discriminant representation of the CMWF with enhanced fusion features is proposed to obtain multi-levels of fusion features by integrating RGB and depth features. Notably, F_3^f and F_4^f are sent to a bi-directional scale-correlation convolution (BSCC) module to gather high-level semantic features and enhance the inference of the result. For the decoder, we use the edge detection operator (i.e., Sobel [28]) to generate edge-enhanced features for subsequent edge-guided (EG) decoders to retain a clearer edge

2.2. Cross-Modality Weighting Fusion (CMWF) module

The interaction between RGB features and depth features is essential for RGB-D SOD. However, if low-quality depth maps are chosen to be used in the cross-modal fusion process, interference and non-salient objects may be included, and resulting in suboptimal performance. CMWF is deployed to leverage the unique properties of RGB and depth information in order to enhance the fusion process. By assigning appropriate weights to these modalities, the mechanism aims to effectively capture and combine the complementary information they provide. As illustrated in Figure 4, CMWF consists of three main components: channel-spatial attention feature enhancement, depth-quality assessment mechanism and depth-weighted cross-modality fusion.

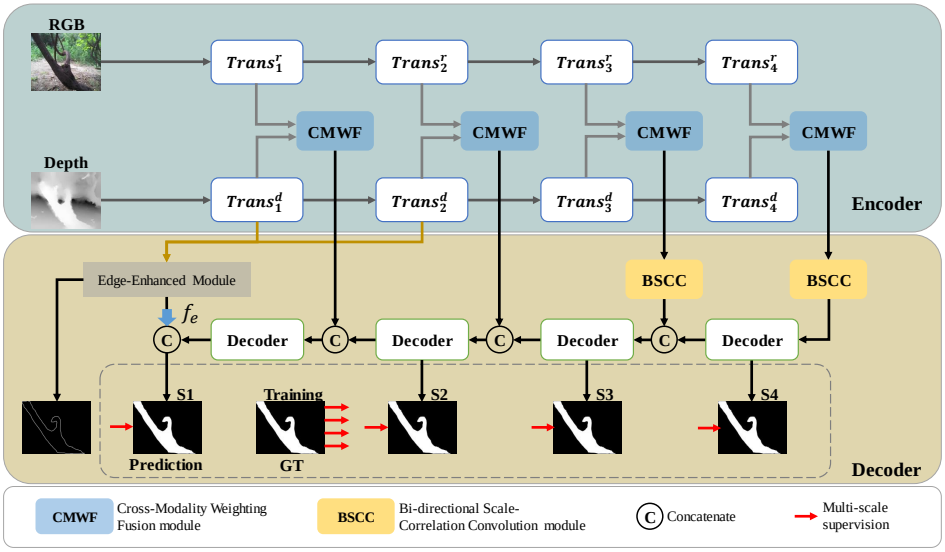


Fig. 3. Architecture of the proposed method. The feature encoding, cross-modal weighting fusion, and saliency inference are depicted as three essential processes in the part

Channel-Spatial Attention Feature Enhancement(CSAE). It is increasingly acknowledged that RGB images and depth maps are essentially two distinct kinds of information representations in RGB-D SOD. RGB information contains rich color and texture details, which are particularly useful for capturing surface appearance and object bound-

aries. On the other hand, depth information provides valuable depth cues and shape information, which are crucial for understanding object geometry and spatial relationships. Therefore, the features of each modality hold different importance in the channel. A Simple fusion of these features may introduce noise, thereby requiring feature enhancement to increase the discrimination between modalities, facilitating the subsequent feature assessment. Attention mechanisms have been introduced to the computer vision system, which is motivated by the human visual system. Across the previous years, it has seen an increase in the significance of attention mechanisms in computer vision. In general, there are three distinct categories of attention mechanisms: channel attention mechanisms, spatial attention mechanisms, and combination mechanisms that combine the former two classes. Different from previous works [29,30], which combined channel and spatial attention mechanisms serially, we use a parallel approach to infer attention mapping for each branch (spatial and channel). As illustrated in Figure 4, this parallel channel-spatial attention could simultaneously concentrate on "what" and "where", which captures the importance of different feature channels, emphasizes common salient objects, and removes unnecessary noise.

Specifically, we calculate channel attention maps and spatial attention maps of RGB feature f_i^r and depth feature f_i^d, which are defined as follows:

$$CA_i^c = Sigmoid(Conv_1(GAP(f_i^c))), \tag{1}$$

$$SA_i^c = Sigmoid(Conv_3(CGAP(f_i^c))), \tag{2}$$

where $c \in (r, d)$, $i \in (1, 2, 3, 4)$, CA_i^c and SA_i^c present the channels and spatial attention maps at the i-th level, respectively. $GAP(\cdot)$ represents the global average pooling operation, $CGAP(\cdot)$ means the global average pooling operation along channel direction. $Conv_k(\cdot)$ represents the convolution operation with the kernel size $k \times k (k = 1, 3)$, and $Sigmoid(\cdot)$ represents the sigmoid activation function. Next, the original features with the attention are combined to create the enhanced features, which are described as follows:

$$F_i^c = f_i^c \times (CA_i^c \times SA_i^c), \tag{3}$$

where $c \in (r, d)$, F_i^c is the feature enhanced by channel-spatial attention, and $\times$ represents the operations of element-wise multiplication.

Depth-Quality Assessment (DQA) Mechanism. RGB images and depth maps are essential components of RGB-D SOD, and the fused features often serve as the key breakthroughs in RGB-D SOD performance during cross-modal data fusion. However, low-quality depth images resulting from factors such as dark lighting, random noise, and human interference can lead to information loss. Directing the fusion of features extracted from RGB and depth streams without discrimination will lead to adverse outcomes. Therefore, it is necessary to develop quality assessment mechanisms and feature screening for depth maps. Inspired by [11,31,32] that high-quality depth map should be at the same level of quality as the RGB image and have similar distinctive features, we design a depth-quality assessment (DQA) mechanism to compute depth weighting factors. As shown in Figure 4, we weight the enhanced feature F_i^d to prevent excessive noise

introduction. Specifically, we assess the similarity between the depth feature-level attention map and the RGB feature-level attention map to reduce the importance of these depth maps. The quality assessment mechanism is a crucial part of CMWF, which minimizes the negative effects of low-quality depth maps on SOD.

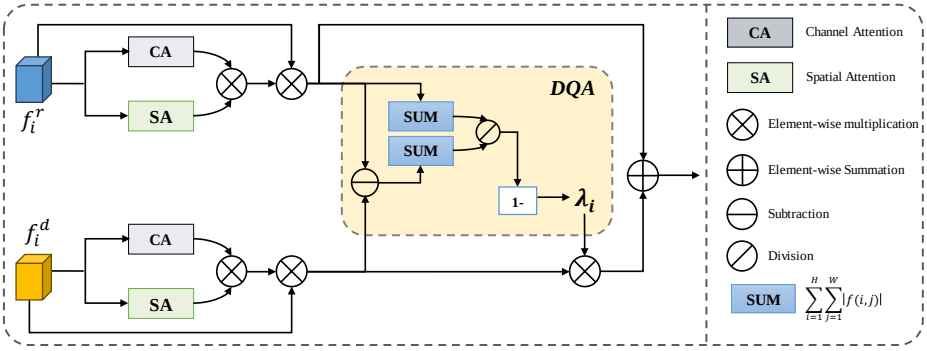


Fig. 4. The structure of the cross-modality weighting Fusion (CMWF) module."1-" means 1 subtract the input of the equation

Specifically, calculating the difference between the enhanced RGB features and the depth features is a subtraction operation, which generates the absolute value difference between the features of the two modalities. Then, the resulting difference of the absolute value of the RGB feature pixel value is utilited to get the i-th level weighting factor λ_i, which can be defined as follows:

$$S_i^{rd} = Conv_1(F_i^r) - Conv_1(F_i^d), \tag{4}$$

$$\lambda_i = 1 - \frac{SUM(S_i^{rd})}{SUM(F_i^r)}, \tag{5}$$

where $SUM(f) = \sum_{i=1}^{H}\sum_{i=1}^{H}|f(i,j)|$, $|\cdot|$ represent the absolute value operation, H and W are the height and width of feature f, respectively.

Depth-weighted Cross-Modality Fusion. This process is shown in Figure 4, we combine RGB and depth features in a weighting factors manner to obtain fusion features F_i^f, The process can be expressed as:

$$F_i^f = F_i^r + F_i^d \times \lambda_i, \tag{6}$$

this module uses the depth information to weigh the contribution of depth modality during the fusion process, thus improving accuracy and reducing noise.

2.3. Bi-directional Scale-Correlation Convolution (BSCC)

Generally, low-level features focus on the fine-grained details of the learning object, while high-level features provide more semantic information. Several approaches [21,33,34,35]

often use atrous convolution with different dilation rates to aggregate multi-scale features and inference context information. However, this approach leads to the loss of some details and local semantic information. Moreover, in the process of multi-scale information fusion, features across different scales lack correlation, which may not be conducive to generating high-quality saliency maps. To leverage the advantages of high-level features and long-distance information correlation, we design a bi-directional scale-correlation convolution (BSCC) module. The module extracts features step-by-step via top-down and bottom-up bi-directional paths. As shown in Figure 5, the BSCC contains four branches and connects them in series to fill in the holes. Each branch passes through a series of fine-grained convolutional operations to extract higher-level semantic information.

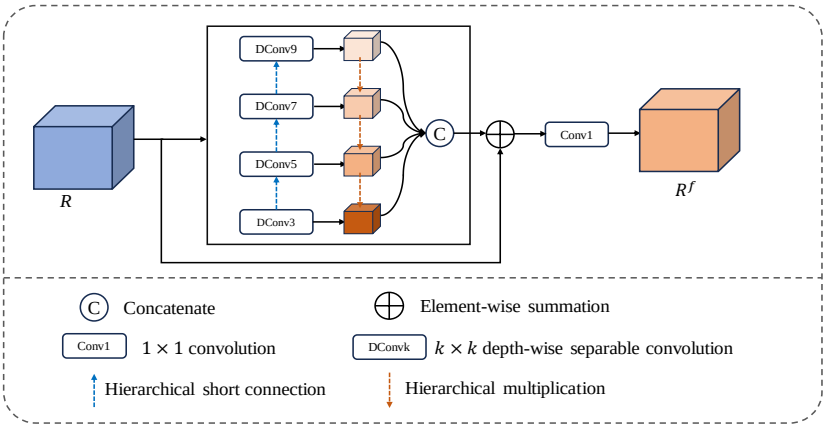


Fig. 5. The structure of the BSCC for multi-scale information fusion

Specifically, hierarchical short connections are first used to successively obtain four-scale features from the depth-wise separable convolution layers of the convolution kernel $(3, 5, 7,$ and $9)$. In addition, through hierarchical multiplication, the proposed bi-directional scale-correlation convolution (BSCC) module enhances both high-level features and multi-scale feature learning. Suppose the input feature is denoted R:

$$\begin{cases} R_1 = DConv_3(R) + R \\ R_i = DConv_{2\times i+1}(R_{i-1}) + R, i \in (2, 3, 4), \\ \tilde{R}_i = R_{i+1} \times R_i, i \in (1, 2, 3) \end{cases} \tag{7}$$

where $DConv_3$ represent a 3×3 depth-wise separable convolution, and $DConv_{2\times i+1}$ represent depth-wise separable convolution with a convolution kernel of $2 \times i + 1$. Finally, we cascade the obtained multi-scale features by concatenation, and transform the feature map through a 1×1 convolutional layer. In addition, residual concatenation is designed inside the BSCC to prevent the loss of high-level semantic information, which can be expressed as:

$$R^f = Conv_1(Cat(\tilde{R}_1, \tilde{R}_2, \tilde{R}_3, R_4)) + R. \tag{8}$$

BSCC learns scale-correlation features efficiently by increasing the receptive field using progressive scale features, thereby greatly improving the ability to effectively infer salient objects from complex real-world scenes.

2.4. Edge-Guided (EG) decoder

This section focuses on the edge-guided decoder, which includes feature optimisation based on the edge-enhanced mechanism and information fusion strategies based on edge-guided fusion.

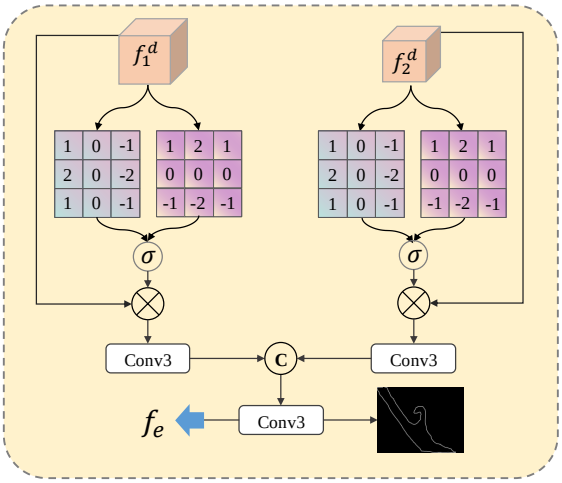


Fig. 6. The architecture of the Edge-Enhanced Module

Edge-Enhanced Mechanism. Edges are one of the important features in saliency object detection, and typically RGB image based edge enhancement methods are not as effective as they could be, mainly because RGB image low-level features only contain simple representations of surface information such as colors and textures, which are susceptible to interference when the background texture of some application scenes is complex. In contrast, depth maps are able to directly capture the difference in distance between the surface of an object and the background, better preserving detailed information such as the edge of the object. Figure 6 shows how EEM extracts edge information by using depth low-level features f_1^d and f_2^d to filter out unimportant edge information. To accomplish this, we build the gradient map using the Sobel operator [28] in both the horizontal G_x and vertical G_y direction. Convolution procedures are performed with two fixed $3\times$ 3 parameters with a stride of 1. One definition of these convolutions is:

$$M_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \qquad M_y = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 1 \\ -1 & -2 & -1 \end{bmatrix}. \tag{9}$$

The two convolutions described above are then applied to the input feature map to create the gradient maps I_x andI_y. The edge-enhanced feature map f_e is produced by fusing these gradient maps with the input feature map after normalizing them with the sigmoid function:

$$f_e = f_d \times Sigmoid(Cat(I_x, I_y)), \tag{10}$$

where $Cat(\cdot)$ denotes concatenation along the channel dimension. Next, we fuse the edge-enhanced feature maps of f_1^d and f_2^d. Frist, the feature map of f_2^d is first applied to a 1×1 convolution operation, and a feature map with the same size as f_1^d is generated using bilinear upsampling. Then, to equalize the number of channels between the two feature maps, a 3×3 convolution is performed on each of the two feature maps. The final edge feature map f_e is created by concatenating the two feature maps along the channel dimension and applying two layers of convolution, with the result being supervised by the ground truth edge map.

Edge-Guided Fusion. Low-level features can provide more detailed and localized information that can be used to locate small objects well, while high-level features can capture global information and enable the detection of large objects. Inspired by the decoding idea widely used in the Unet framework [36]. According to [36], the edge-guided decoder is designed to combine the obtained multi-scale cross-modal weighting fusion features with edge-enhanced features in a progressive manner to make full use of the multi-scale features. As shown in Figure 3, multi-scale feature fusion results are obtained by channel concatenating using cross-modal weighting fusion features and edge-enhanced features as input. The methods are as follows:

$$\overline{f}_i^c = \begin{cases} Cat(F_i^f, Up(Conv_3(F_{i+1}^f))), i = 1, 2, 3 \\ F_i^f. \qquad\qquad\qquad\qquad\quad i = 4 \end{cases} \tag{11}$$

Finally, the edge features from the edge-enhanced module are combined with the fused features to generate the edge-guided salient map *S*.

$$S = Up(Cat(\overline{f}_i^c, f_e)). \tag{12}$$

2.5. Loss Function

As shown in Figure 2, saliency maps of intermediate predictions are produced during the network testing phase in each decoder block, denoted as *S*(*t*). The ground truth *GT* supervises each *S*(*t*). We adopt the Binary Cross-Entropy (BCE) loss function and the Intersection-Over-Union (IOU) loss function during the supervised training procedure to enhance the content representation. The total loss function, denoted *TLoss*, is defined as follows:

$$TLoss = \sum_{t=1}^{5} L_{bce}^{(t)}(up(S(t)), GT) + L_{iou}^{(t)}(up(S(t)), GT), \tag{13}$$

where *GT* represents the ground truth, $up(\cdot, \cdot)$ represents bilinear upsampling, and $S(t)$ is upsampled to the same resolution as GT; $L_{bce}^{(t)}(\cdot, \cdot)$ represents BCE loss, $L_{iou}^{(t)}(\cdot, \cdot)$ represents the IOU loss, and its calculation formula is defined as:

$$L_{bce} = -y \times \lg(\hat{y}) + (1 - y)\lg(1 - \hat{y}), \tag{14}$$

$$L_{iou} = \frac{|A \cap B|}{|A \cup B|}, \tag{15}$$

where y represents the analogy to which it belongs, $\hat{y}$ represents the prediction map; A represents the $S(t)$ area and B represents the GT area.

3. Experiments

3.1. Datasets

We evaluate the performance of the proposed model on five popular RGB-D SOD benchmarks, including RGBD135 dataset [37], NLPR dataset [38], NJU2K dataset [39], SSD dataset [40], and STERE [41] dataset. RGBD135 dataset [37] consists of 135 indoor RGB-D images obtained by Kinect. NLPR dataset [38] is a collection of 1000 RGB images and depth maps pairs that were acquired using Microsoft Kinect and include both indoor and outdoor scenes. NJU2K dataset [39] is a collection of 1985 RGB images and matching depth maps gathered from stereoscopic photographs, the internet, and 3D movies. The SSD [40] is made up of 80 RGB-D images, each of which is part of stereoscopic film production. The STERE [41] integrates 1000 pairs of binocular images collected from the Internet and is the first stereo image dataset in this field. For unbiased algorithm evaluation and comparison, following [42,43], the same training dataset comprising 1485 RGB-D images from the NJU2K dataset and 700 RGB-D images from the NLPR dataset was used. The remaining images were used for subsequent testing.

3.2. Evaluation Metrics

To evaluate the performance of the proposed approach and other approaches, we select four common evaluation metrics, including Precision-Recall (PR) curve, adaptive E-measure (E_{ξ}^{adp}) [44], adaptive F-measure (F_{β}^{adp}) [45], S-measure (S_{α}) [46], and Mean Absolute Error (MAE) [47].

The Precision and Recall of a saliency map *S* and a ground truth *G* are calculated by the PR curve, which are defined as follows:

$$Precision = \frac{|M \cup G|}{|M|}, \tag{16}$$

$$Recall = \frac{|M \cap G|}{|G|}, \tag{17}$$

where M is the conversion of S into a binary image, specifically, a set of thresholds are used to divide the saliency map S, and the threshold variation range is 0-255. First, the recall and precision are calculated for each threshold, then, they are combined to represent the PR curves of the model at different thresholds. The PR curves reflect the mean recall and precision of different saliency maps at various thresholds.

The F-measure generates the harmonic mean of precision P (precision) and recall R (recall), which is expressed as:

$$F_{\beta} = (1 + \beta^2)\frac{P \times R}{\beta^2 \times P + R}, \tag{18}$$

where β is set to emphasize the accuracy [48], and based on [48], F_{β}^{adp} is obtained under various thresholds (0-255).

The S-measure calculates the structural similarity between the predicted image and the ground-truth map, which is defined as:

$$S_{\alpha} = \alpha * S_o + (1 - \alpha) * S_r, \tag{19}$$

where S_o represents object-aware structural similarity, S_r represents region-aware structural similarity, and α is a balance parameter, set to 0.5.

The E-measure measures both image-level statistics and local pixel-matching information, which is defined as:

$$E_{\phi} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi_{FM}(i, j), \tag{20}$$

where ϕ_{FM} represents the enhanced alignment matrix [44], W and H stand for the width and height of an image, respectively, and we can get E_{ξ}^{adp} from the basics.

MAE calculates the pixel-level errors between the ground truth G and the predicted saliency map S, which is defined as:

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S_{(i,j)} - G_{(i,j)}|, \tag{21}$$

where $S(i, j)$ and $G(i, j)$ present the value of the pixel (i, j) from predicted saliency map and ground truth, respectively.

3.3. Implementation Details

We implemented our proposed methodology with one NVIDIA GeForce RTX 3090 GPU founded on the Pytorch framework. No preprocessing techniques are used to support the model during training and debugging to maximize performance. All images have been resized to 384 $\times$ 384, and three channels are copied with the depth maps. We use a pre-trained model from ImageNet [49] to initialize the parameters of the backbone network. Then, we enforce augmentation on all the training dataset to avoid overfitting, including random flipping and rotating, such results show that we can achieve better results with enhancement training in most cases. Furthermore, the learning rate initiates at a value of $1e^{-5}$, which degrades by 10 every 100 epochs. With a batch size of 3, we train our network for 150 epochs till convergence.

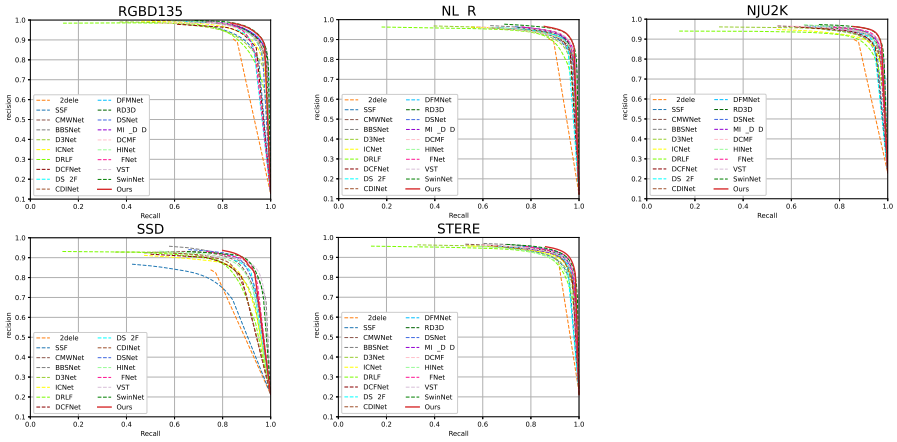


Fig. 7. PR curves of the proposed model and other salient object detection techniques on RGBD135 [37], NLPR [38], NJU2K [39], SSD [40], and STERE [41] datasets

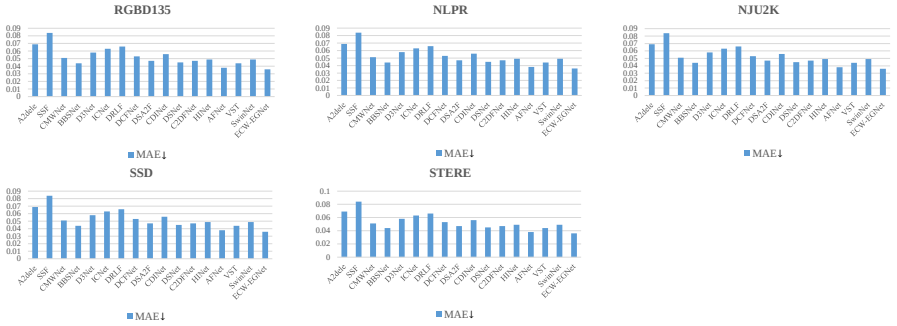


Fig. 8. MAE visualization result of the proposed model and other salient object detection techniques on RGBD135 [37], NLPR [38], NJU2K [39], SSD [40], and STERE [41] datasets

3.4. Comparison with SOTA Methods

To validate the effectiveness of the proposed ECW-EGNet, we compare with 21 SOTA methods, including A2dele [50], SSF [51], CMWNet [52], BBSNet [53], D3Net [54], ICNet [55], DRLF [6], DCFNet [56], DSA2F [57], CDINet [58], DFMNet [59], RD3D [60], DSNet [61], MIA_DPD [62], DCMF [63], C2DFNet [64], HINet [9], AFNet [7], VST [65], SwinNet [30], CATNet [66] and FCFNet [67]. For fair comparisons, the authors or providers of these methods provided or gave the default parameter settings for implementation of the original code, which was used to produce the saliency maps.

Quantitative evaluation. To validate the effectiveness of the ECW-EGNet proposed in this paper, our model is compared with 21 SOTA methods in a comprehensive manner

in terms of four evaluation metrics, with higher E_{ξ}^{adp}, S_{α} and F_{β}^{adp} values indicating better performance and the opposite for MAE values, as shown in Table 1. We can see that our method achieves the best results in almost all five test datasets. This can be attributed to three main aspects: Firstly, our CMWF module can perform a weighted fusion of low-quality depth maps. Secondly, BSCC can effectively integrate high-level semantic information and highlight object regions. Finally, EG synthesizes more accurate edge masks to guide the upgrading of salient object edges. In addition, we also provide the Precision-Recall (PR) curves in Figure 7 and visualize the MAE from Table 1 (as depicted in Figure 8) to further show the effectiveness of the proposed ECW-EGNet method. As can be seen from the results, our model achieves higher precision and recall scores on the five datasets compared to other methods. Thus, all quantitative measures demonstrate the effectiveness of the proposed model.

Table 1. Quantitative comparisons of the proposed method against the other 21 state-of-the-art RGB-D SOD methods. ↑/↓ indicates that a larger/smaller is better. The top three results are highlighted in red, blue, and green, respectively

Dataset	Metric	A2dele	SSF	CMWNet	BBSNet	D3Net	ICNet	DRLF	DCFNet	DSA2F	CDINet	DFMNet	RD3D	DSNet	MALDPD	DCMF	C^{2}DFNet	HINet	AFNet	FCFNet	VST	SwinNet	CATNet	ECW-EGNet
		CVPR20	CVPR20	ECCV20	ECCV2020	TNNLS20	TIP21	TIP21	CVPR21	CVPR21	MM21	MM21	AAAI21	TMM21	NP22	TIP22	TMM22	PR23	NP23	TSCVT23	ICCV21	TCSVT22	TMM 23	Ours
		CNN-based																			Transformer-based			
RGBD135	$E_{\xi}^{adp}\uparrow$	0.922	0.948	0.967	0.967	0.951	0.959	0.954	0.960	0.957	0.972	0.972	0.975	0.970	0.973	0.967	0.967	0.969	0.954	0.978	**0.979**	**0.980**	**0.980**	**0.977**
	$S_{\alpha}\uparrow$	0.886	0.905	0.934	0.933	0.897	0.920	0.895	0.918	0.916	0.937	0.938	0.935	0.928	0.936	0.932	0.924	0.927	0.925	**0.939**	**0.943**	**0.945**	**0.945**	**0.939**
	$F_{\beta}^{adp}\uparrow$	0.865	0.876	0.900	0.906	0.870	0.889	0.868	0.895	0.898	0.913	0.907	0.917	0.910	0.917	0.896	0.907	0.907	0.894	0.908	0.917	**0.926**	**0.923**	**0.933**
	$MAE\downarrow$	0.028	0.025	0.022	0.021	0.031	0.027	0.030	0.022	0.021	0.020	0.019	0.019	0.021	0.018	0.023	0.018	0.022	0.022	**0.017**	**0.017**	**0.016**	**0.016**	**0.015**
NLPR	$E_{\xi}^{adp}\uparrow$	0.945	0.951	0.940	0.953	0.945	0.944	0.936	0.956	0.952	0.955	0.954	0.959	0.957	0.958	0.940	0.961	0.950	0.960	0.949	0.956	**0.969**	**0.971**	**0.973**
	$S_{\alpha}\uparrow$	0.896	0.914	0.917	0.930	0.911	0.922	0.903	0.921	0.917	0.927	0.925	0.930	0.926	0.931	0.922	0.927	0.922	**0.936**	0.924	0.931	**0.941**	**0.940**	**0.941**
	$F_{\beta}^{adp}\uparrow$	0.878	0.875	0.859	0.882	0.861	0.869	0.844	0.893	0.896	0.883	0.880	0.892	0.886	0.887	0.854	0.899	0.877	0.886	0.879	0.886	**0.908**	**0.917**	**0.924**
	$MAE\downarrow$	0.028	0.026	0.029	0.023	0.029	0.028	0.032	0.023	0.024	0.024	0.024	0.022	0.024	0.022	0.029	0.021	0.026	**0.020**	0.024	0.023	**0.018**	**0.018**	**0.016**
NJU2K	$E_{\xi}^{adp}\uparrow$	0.916	0.935	0.922	0.942	0.915	0.912	0.903	0.941	0.937	0.945	0.937	0.942	0.947	0.944	0.925	0.941	0.939	0.949	0.929	0.943	**0.954**	**0.958**	**0.960**
	$S_{\alpha}\uparrow$	0.869	0.899	0.903	0.921	0.901	0.894	0.886	0.903	0.903	0.919	0.912	0.916	0.921	0.914	0.913	0.908	0.915	0.926	0.918	0.922	**0.935**	**0.932**	**0.934**
	$F_{\beta}^{adp}\uparrow$	0.874	0.886	0.880	0.902	0.865	0.867	0.849	0.898	0.901	0.907	0.894	0.901	0.907	0.898	0.881	0.899	0.896	0.909	0.891	0.900	**0.922**	**0.929**	**0.931**
	$MAE\downarrow$	0.051	0.042	0.045	0.035	0.046	0.052	0.055	0.038	0.039	0.035	0.039	0.037	0.034	0.036	0.043	0.038	0.039	0.032	0.034	0.034	**0.027**	**0.026**	**0.024**
SSD	$E_{\xi}^{adp}\uparrow$	0.870	0.873	0.902	0.920	0.904	0.879	0.880	0.906	0.912	0.906	-	-	0.923	-	-	0.920	0.916	**0.932**	-	0.922	**0.925**	-	**0.938**
	$S_{\alpha}\uparrow$	0.808	0.790	0.875	0.882	0.856	0.848	0.835	0.852	0.876	0.853	-	-	0.885	-	-	0.872	0.865	**0.897**	-	0.889	**0.892**	-	**0.891**
	$F_{\beta}^{adp}\uparrow$	0.790	0.761	0.820	0.849	0.813	0.799	0.801	0.829	0.852	0.827	-	-	0.853	-	-	0.848	0.837	**0.855**	-	0.842	**0.863**	-	**0.880**
	$MAE\downarrow$	0.069	0.084	0.051	0.044	0.058	0.063	0.066	0.053	0.047	0.056	-	-	0.045	-	-	0.047	0.049	**0.038**	-	0.044	**0.040**	-	**0.036**
STERE	$E_{\xi}^{adp}\uparrow$	0.935	0.935	0.930	0.941	0.923	0.925	0.916	0.945	0.949	0.942	0.939	0.944	0.947	0.942	0.930	0.946	0.927	**0.952**	0.927	0.942	0.950	**0.954**	**0.958**
	$S_{\alpha}\uparrow$	0.885	0.893	0.905	0.908	0.898	0.905	0.888	0.908	0.903	0.905	0.908	0.911	0.915	0.909	0.910	0.905	0.892	0.918	0.906	0.913	**0.919**	**0.921**	**0.922**
	$F_{\beta}^{adp}\uparrow$	0.884	0.880	0.869	0.885	0.859	0.864	0.845	0.897	**0.898**	0.890	0.875	0.886	0.894	0.882	0.866	0.892	0.859	**0.898**	0.868	0.878	0.893	**0.904**	**0.910**
	$MAE\downarrow$	0.043	0.044	0.043	0.041	0.046	0.045	0.050	0.037	0.036	0.041	0.040	0.037	0.036	0.037	0.043	0.038	0.049	0.034	0.034	0.038	**0.033**	**0.030**	**0.028**

Qualitative evaluation. Figure 9 shows the qualitative comparison of our proposed approach and other SOTA methods in in various challenging scenarios, including low-quality depth maps (1st and 2nd rows), small objects (3rd and 4th rows), multiple objects (5th and 6th rows), low contrast (7th and 8th rows), complex scenes (9th and 10th rows).

In the 1st and 2nd rows of the example, the depth map information is poor, but our method can still accurately detect the saliency object, which proves the effectiveness of the CMWF module. The 3rd and 4th rows show some small object samples, and the 3rd row includes a small flower. Despite the small size of these objects, our method is also able to detect them accurately. Examples with multiple objects are shown in the 5th and 6th rows. In the 6th row, most SOTA methods fail to detect three complete windows, while our method is able to segment salient objects accurately. The 7th and 8th rows demonstrate examples of low contrast between background and object regions. In the 8th row, most SOTA methods fail to separate the owl from the branch, but our method is able to accurately separate the salient object from the background. Finally, the 9th and 10th rows show examples of complex scenarios where our method accurately detects

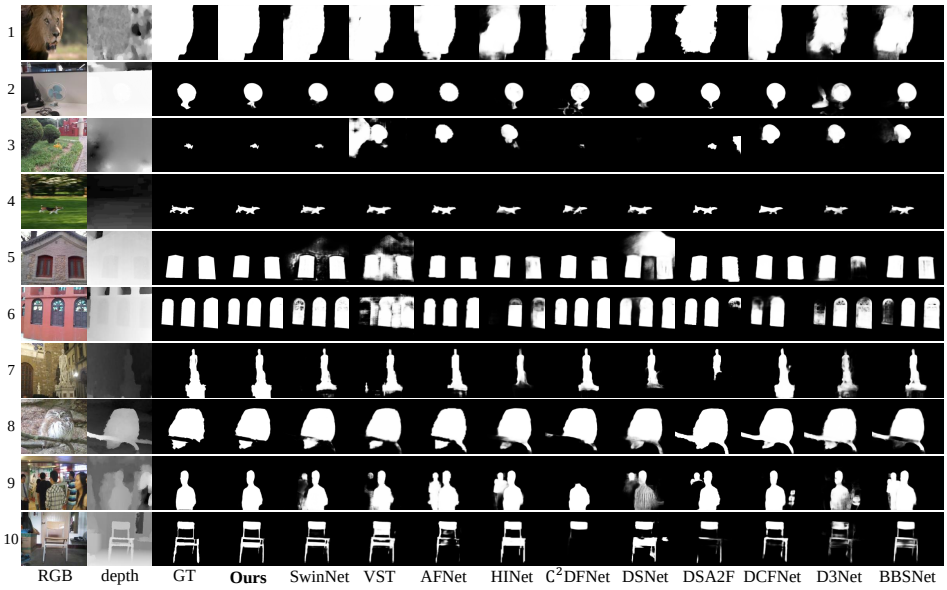


Fig. 9. Visual comparison of different RGB-D SOD methods. Including SwinNet [30], VST [65], AFNet [7], HINet [9], C2DFNet [64], DSNet [61], DSA2F [57], DCFNet [56], D3Net [54], BBSNet [53]

salient objects by effectively suppressing background interference, while other methods produce spurious results. On the 10th row, our method completely segmented the objects by highlighting foreground objects, while the other methods output incomplete results. The above examples show that our approach can effectively detect meaningful objects in a variety of scenarios.

4. Discussion

In this section, we perform ablation studies to verify the effectiveness of each component of our method on the NLPR and STERE datasets. We focus on (1) the importance of the CMWF module, (2) the effect of BSCC on RGB-D SOD, (3) the necessity of Edge-Guided (EG) decoders, and (4) the impact of multi-scale supervision (MS). As shown in Table 2, where $M1$ indicates that the model does not use any components and only contains MS. $M2$ added CMWF and MS, and removed BSCC and EG. $M3$ means that BSCC and MS are added to the model, and CMWF and EG are removed. $M4$ indicates that the model contains both CMWF, EG, and MS, but does not use BSCC. $M5$ indicates that the model contains both CMWF, BSCC, and EG, but does not use MS. $M6$ means that the model uses all components simultaneously (i.e., CMWF, BSCC, EG, and MS).

The effectiveness of CMWF. The CMWF module merges RGB and depth images in our network in a weighted manner. By comparing $M1$ and $M2$, we can see a significant improvement in performance for both datasets. This phenomenon is even more pronounced in the STERE dataset: the F-measure and MAE of the CMWF module im-

Table 2. Ablation studies for our ECW-EGNet on two datasets. " $\surd$ " denote the model contains the corresponding component

Model	CMWF	BSCC	EG	MS	NLPR				STERE			
					$E_\xi^{adp}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{adp}\uparrow$	$MAE\downarrow$	$E_\xi^{adp}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{adp}\uparrow$	$MAE\downarrow$
M 1				$\surd$	0.965	0.936	0.907	0.020	0.942	0.917	0.887	0.035
M 2	$\surd$			$\surd$	0.969	0.937	0.917	0.018	0.952	0.920	0.902	0.031
M 3		$\surd$		$\surd$	0.972	0.940	0.918	0.017	0.953	0.921	0.902	0.030
M 4	$\surd$		$\surd$	$\surd$	0.970	0.941	0.917	0.017	0.951	0.922	0.902	0.030
M 5	$\surd$	$\surd$	$\surd$		0.969	0.941	0.921	0.017	0.955	0.921	0.904	0.030
M 6	$\surd$	$\surd$	$\surd$	$\surd$	**0.973**	**0.941**	**0.924**	**0.016**	**0.958**	**0.922**	**0.910**	**0.028**

prove performance by 1.7% and 11.4% respectively. This is because STERE contains a high volume of noisy depth images, whereas our CMWF reduces the interference information from low-quality images. To demonstrate the superiority of CMWF, we have experimentally compared the performance of three variants of the CMWF module. Specifically, the model $w/o\ CMWF$ express our full model without the CMWF module; The model $w/o\ CSAE$ illustrate our full model without the CSAE mechanism; and the model $w/o\ DQA$ displays our full model without the DQA mechanism. The results are shown in Table 3 and it can be seen that all three steps in CMWF are helpful in improving the performance of SOD, especially after removing the DQA module there is a clear decrease in performance, which confirms the potential of our DQA to evaluate depth images and reduce the negative impact of low-quality depth maps.

The effectiveness of BSCC. In ECW-EGNet, multi-scale feature fusion is achieved by using depth-wise separable convolutions of various convolutional kernel sizes. In Table 2, the comparison between $M1$ and $M3$ shows that the addition of the BSCC module to integrate global features improves the performance of the network on these two datasets. Specifically, when evaluating the NLPR dataset, the E-measure and F-measure of the model increase by 0.7% and 1.2% respectively upon incorporating the BSCC module.

Table 3. Comparison of quantitative indicators between CMWF and variant

Model	NLPR				STERE			
	$E_\xi^{adp}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{adp}\uparrow$	$MAE\downarrow$	$E_\xi^{adp}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{adp}\uparrow$	$MAE\downarrow$
w/o CMWF	0.962	0.930	0.904	0.020	0.948	0.918	0.895	0.033
w/o CSAE	0.972	0.940	0.918	0.017	0.953	0.918	0.901	0.031
w/o DQA	0.971	0.940	0.921	0.017	0.951	0.921	0.901	0.031
CMWF	**0.973**	**0.941**	**0.924**	**0.016**	**0.958**	**0.922**	**0.910**	**0.028**

The effectiveness of EG. It refines the contour details of salient objects by extracting salient edge features and integrating the edge features into the decoding process. In Table 2, by looking in $M2$ and $M4$, we can see that adding EG also wins edge segmentation on NLPR, e.g., E_ξ^{adp}: 0.969$\longrightarrow$ 0.970, S_α: 0.937$\longrightarrow$ 0.941, MAE: 0.018$\longrightarrow$ 0.017.

Effectiveness of multi-scale supervision (MS). During the training of the network, we uses multi-scale supervision to achieve an accurate representation of the salient objects. To verify the effectiveness of multiple supervision, we remove all other losses in the network, keeping only the last layer of loss L1. The results for this experiment are presented in row $M5$ of Table 2. It can be observed that the removal of multi-scale supervision results in a decrease in performance, which confirms that the introduction of supervision in the output of each side facilitates performance improvement.

5. Conclusions

In this paper, we propose an efficient ECW-EGNet framework to implement RGB-D SOD. We propose a cross-modality weighting fusion (CMWF) module that assesses the quality of the depth map and fuses cross-modal features by calculating the difference between RGB modal and depth modal in order to compensate for the difference between RGB and depth modes. In CMWF, features extracted from the backbone network are enhanced to generate discriminative features, which are then used to generate weighting factors for the enhanced discriminative features to determine the weights for depth feature fusion. Additionally, we introduce a bi-directional scale-correlation convolution (BSCC) module to learn high-level semantic information to better capture contextual information for effective guided saliency prediction. Moreover, the accuracy of SOD is further improved with the use of an Edge-Guided (EG) decoder. Experimental tests on five representative datasets show that the proposed method outperforms 21 SOTA methods in four evaluation metrics. In the near future, we are especially interested in the prospect of reducing model parameters without compromising model performance. We design a more lightweight backbone network to replace the current Swin Transformer for feature extraction.

Acknowledgments. This work was supported by the Anhui University of Science and Technology Medical Special Training Project(YZ2023H2B003), Huainan City Science and Technology Plan Project(2023A316), National Natural Science Foundation of China (62102003), Anhui Postdoctoral Science Foundation (2022B623), Natural Science Foundation of Anhui Province (2108085QF258), the University Synergy Innovation Program of Anhui Province (GXXT-2021-006, GXXT-2022-038), Academic Support Program for Top Talents in Academic Disciplines (Majors) of Anhui Province (gxbjZD2021050), the Institute of Energy, Hefei Comprehensive National Science Center under (21KZS217), University-level general projects of Anhui University of science and technology (xjyb2020-04), Central guiding local technology development special funds (202107d06020001).

References

1. Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7223–7233, 2019.
2. Wei Xiong, Yafei Lv, Yaqi Cui, Xiaohan Zhang, and Xiangqi Gu. A discriminative feature learning approach for remote sensing image retrieval. *Remote Sensing*, 11(3):281, 2019.
3. Junyang Yu, Mengle Zuo, Lifeng Dong, Huanlong Zhang, and Xin He. The multi-level classification and regression network for visual tracking via residual channel attention. *Digital Signal Processing*, 120:103269, 2022.

4. Fenglei Chen, Haijun Liu, Zhihong Zeng, Xichuan Zhou, and Xiaoheng Tan. Bes-net: Boundary enhancing semantic context network for high-resolution image semantic segmentation. *Remote Sensing*, 14(7):1638, 2022.
5. Kang Yi, Jinchao Zhu, Fu Guo, and Jing Xu. Cross-stage multi-scale interaction network for rgb-d salient object detection. *IEEE Signal Processing Letters*, 29:2402–2406, 2022.
6. Xuehao Wang, Shuai Li, Chenglizhao Chen, Yuming Fang, Aimin Hao, and Hong Qin. Data-level recombination and lightweight fusion scheme for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:458–471, 2020.
7. Tianyou Chen, Jin Xiao, Xiaoguang Hu, Guofeng Zhang, and Shaojie Wang. Adaptive fusion network for rgb-d salient object detection. *Neurocomputing*, 522:152–164, 2023.
8. Chenxing Xia, Songsong Duan, Xianjin Fang, Xiuju Gao, Yanguang Sun, Bin Ge, Hanling Zhang, and Kuan-Ching Li. Efgnet: Encoder steered multi-modality feature guidance network for rgb-d salient object detection. *Digital Signal Processing*, 131:103775, 2022.
9. Hongbo Bi, Ranwan Wu, Ziqi Liu, Huihui Zhu, Cong Zhang, and Tian-Zhu Xiang. Cross-modal hierarchical interaction network for rgb-d salient object detection. *Pattern Recognition*, 136:109194, 2023.
10. Zhihong Zeng, Haijun Liu, Fenglei Chen, and Xiaoheng Tan. Compensated attention feature fusion and hierarchical multiplication decoder network for rgb-d salient object detection. *Remote Sensing*, 15(9):2393, 2023.
11. Chenglizhao Chen, Jipeng Wei, Chong Peng, and Hong Qin. Depth-quality-aware salient object detection. *IEEE Transactions on Image Processing*, 30:2350–2363, 2021.
12. Hao Chen, Yongjian Deng, Youfu Li, Tzu-Yi Hung, and Guosheng Lin. Rgbd salient object detection via disentangled cross-modal fusion. *IEEE Transactions on Image Processing*, 29:8407–8416, 2020.
13. Xiaolong Cheng, Xuan Zheng, Jialun Pei, He Tang, Zehua Lyu, and Chuanbo Chen. Depth-induced gap-reducing network for rgb-d salient object detection: an interaction, guidance and refinement approach. *IEEE Transactions on Multimedia*, 2022.
14. Chenxing Xia, Songsong Duan, Xianjin Fang, Bin Ge, Xiuju Gao, and Jianhua Cui. Dast: Depth-aware assessment and synthesis transformer for rgb-d salient object detection. In *Pacific Rim International Conference on Artificial Intelligence*, pages 473–487, 2022.
15. Qian Chen, Keren Fu, Ze Liu, Geng Chen, Hongwei Du, Bensheng Qiu, and Ling Shao. Ef-net: A novel enhancement and fusion network for rgb-d saliency detection. *Pattern Recognition*, 112:107740, 2021.
16. Gongyang Li, Zhi Liu, Minyu Chen, Zhen Bai, Weisi Lin, and Haibin Ling. Hierarchical alternate interaction network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:3528–3542, 2021.
17. Yang Yang, Qi Qin, Yongjiang Luo, Yi Liu, Qiang Zhang, and Jungong Han. Bi-directional progressive guidance network for rgb-d salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5346–5360, 2022.
18. Chenglizhao Chen, Jipeng Wei, Chong Peng, Weizhong Zhang, and Hong Qin. Improved saliency detection in rgb-d images using two-phase depth estimation and selective deep fusion. *IEEE Transactions on Image Processing*, 29:4296–4307, 2020.
19. Wen-Da Jin, Jun Xu, Qi Han, Yi Zhang, and Ming-Ming Cheng. Cdnet: Complementary depth network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:3376–3390, 2021.
20. Wujie Zhou, Qinling Guo, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Irfr-net: Interactive recursive feature-reshaping network for detecting salient objects in rgb-d images. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
21. Yu Qiu, Yun Liu, Yanan Chen, Jianwen Zhang, Jinchao Zhu, and Jing Xu. A2sppnet: attentive atrous spatial pyramid pooling network for salient object detection. *IEEE Transactions on Multimedia*, 2022.

22. Nianchang Huang, Yi Liu, Qiang Zhang, and Jungong Han. Joint cross-modal and unimodal features for rgb-d salient object detection. *IEEE Transactions on Multimedia*, 23:2428–2441, 2020.
23. Wenbo Zhang, Yao Jiang, Keren Fu, and Qijun Zhao. Bts-net: Bi-directional transfer-and-selection network for rgb-d salient object detection. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2021.
24. Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Rrecognition*, pages 3052–3062, 2020.
25. Zhengyi Liu, Kaixun Wang, Hao Dong, and Yuan Wang. A cross-modal edge-guided salient object detection for rgb-d image. *Neurocomputing*, 454:168–177, 2021.
26. Zhengyi Liu, Song Shi, Quntao Duan, Wei Zhang, and Peng Zhao. Salient object detection for rgb-d image by single stream recurrent convolution neural network. *Neurocomputing*, 363:46–57, 2019.
27. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10012–10022, 2021.
28. Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of Solid-State Circuits*, 23(2):358–367, 1988.
29. Zhengyi Liu, Yuan Wang, Zhengzheng Tu, Yun Xiao, and Bin Tang. Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network. In *Proceedings of the International Conference on Multimedia*, pages 4481–4490, 2021.
30. Zhengyi Liu, Yacheng Tan, Qian He, and Yun Xiao. Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4486–4497, 2021.
31. Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for rgb-d saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13756–13765, 2020.
32. Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate rgb-d salient object detection via collaborative learning. In *Proceeding of the European Conference on Computer Vision*, pages 52–69. Springer, 2020.
33. Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3684–3692, 2018.
34. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.
35. Yu-Huan Wu, Yun Liu, Le Zhang, Ming-Ming Cheng, and Bo Ren. Edn: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing*, 31:3125–3136, 2022.
36. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceeding of the Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.
37. Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *Proceedings of International Conference on Internet Multimedia Computing and Service*, pages 23–27, 2014.
38. Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *Proceeding of the European Conference on Computer Vision*, pages 92–109, 2014.

39. Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *IEEE International Conference on Image Processing*, pages 1115–1119, 2014.
40. Chunbiao Zhu and Ge Li. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3008–3014, 2017.
41. Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 454–461, 2012.
42. Yingjie Zhai, Deng-Ping Fan, Jufeng Yang, Ali Borji, Ling Shao, Junwei Han, and Liang Wang. Bifurcated backbone strategy for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:8727–8742, 2021.
43. Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving rgb-d saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4681–4691, 2021.
44. Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018.
45. Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.
46. Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4548–4557, 2017.
47. Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–740, 2012.
48. Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. *IEEE Transactions on Cybernetics*, 48(11):3171–3183, 2017.
49. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
50. Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, pages 9060–9069, 2020.
51. Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for rgb-d saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3472–3481, 2020.
52. Gongyang Li, Zhi Liu, Linwei Ye, Yang Wang, and Haibin Ling. Cross-modal weighting network for rgb-d salient object detection. In *Proceeding of the European Conference on Computer Vision*, pages 665–681, 2020.
53. Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *Proceeding of the European Conference on Computer Vision*, pages 275–292, 2020.
54. Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2075–2089, 2020.
55. Gongyang Li, Zhi Liu, and Haibin Ling. Icnet: Information conversion network for rgb-d based salient object detection. *IEEE Transactions on Image Processing*, 29:4873–4884, 2020.

56. Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated rgb-d salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9471–9481, 2021.
57. Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, and Xi Li. Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1407–1417, 2021.
58. Chen Zhang, Runmin Cong, Qinwei Lin, Lin Ma, Feng Li, Yao Zhao, and Sam Kwong. Cross-modality discrepant interaction network for rgb-d salient object detection. In *Proceedings of the International Conference on Multimedia*, pages 2094–2102, 2021.
59. Wenbo Zhang, Ge-Peng Ji, Zhuo Wang, Keren Fu, and Qijun Zhao. Depth quality-inspired feature manipulation for efficient rgb-d salient object detection. In *Proceedings of the International Conference on Multimedia*, pages 731–740, 2021.
60. Qian Chen, Ze Liu, Yi Zhang, Keren Fu, Qijun Zhao, and Hongwei Du. Rgb-d salient object detection via 3d convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 2, pages 1063–1071, 2021.
61. Hongfa Wen, Chenggang Yan, Xiaofei Zhou, Runmin Cong, Yaoqi Sun, Bolun Zheng, Jiyong Zhang, Yongjun Bao, and Guiguang Ding. Dynamic selective network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:9179–9192, 2021.
62. Yanhua Liang, Guihe Qin, Minghui Sun, Jun Qin, Jie Yan, and Zhonghan Zhang. Multi-modal interactive attention and dual progressive decoding network for rgb-d/t salient object detection. *Neurocomputing*, 490:132–145, 2022.
63. Fengyun Wang, Jinshan Pan, Shoukun Xu, and Jinhui Tang. Learning discriminative cross-modality features for rgb-d saliency detection. *IEEE Transactions on Image Processing*, 31:1285–1297, 2022.
64. Miao Zhang, Shunyu Yao, Beiqi Hu, Yongri Piao, and Wei Ji. C^2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection. *IEEE Transactions on Multimedia*, 2022.
65. Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4722–4732, 2021.
66. Fuming Sun, Peng Ren, Bowen Yin, Fasheng Wang, and Haojie Li. Catnet: A cascaded and aggregated transformer network for rgb-d salient object detection. *IEEE Transactions on Multimedia*, pages 1–14, 2023.
67. Qiang Zhang, Qi Qin, Yang Yang, Qiang Jiao, and Jungong Han. Feature calibrating and fusing network for rgb-d salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023.

Chenxing Xia received the Ph.D. degree from Hunan University in 2019. He is currently an Associate Professor at the College of Computer Science and Engineering, Anhui University of Science and Technology. His research interests include saliency detection, computer vision, image processing, and depth prediction.

Feng Yang received the bachelor degree from Anhui Sanlian University. She is currently pursuing the master degree with Anhui University of Science and Technology. Her research interests include computer vision, machine learning, salient objection detection.

Songsong Duan received the M.S. degree with Anhui University of Science and Technology. He is currently pursuing the Ph.D. degree with the School of Telecommunications Engineering, Xidian University, China. His research interests include computer vision, salient object detection, and weakly supervised learning.

Xiuju Gao received the M.S. degree in Computer Science and Electronic Engineering from Hunan University, Changsha China, in 2016. She is currently an assistant at the College of Electrical and Information Engineering, Anhui University of Science and Technology. Her research interests include image/video processing and computer vision.

Bin Ge received the Ph.D. degree in computer application technology from Hefei university of technology, Hefei China, in 2016. He began to teach at Anhui University of Science and Technology in 1999. Currently, he is a professor of information security at Anhui University of Science and Technology. His research interests include information security, Internet of things technology and artificial intelligence.

Kuan-Ching Li is a Distinguished Professor in the Dept of Computer Science and Information Engineering (CSIE) at Providence University, Taiwan, where he also serves as the Director of the High-Performance Computing and Networking Center. He received the Licenciatura in Mathematics, and MS and Ph.D. degrees in electrical engineering from the University of Sao Paulo (USP), Brazil, in 1994, 1996, and 2001, respectively. He published more than 250 scientific papers and articles and is author, co-author or editor of more than 25 books published by Taylor & Francis, Springer, and McGraw-Hill. Professor Li is the Editor in Chief of the Connection Science and serves as an associate editor for several leading journals. Also, he has been actively involved in many major conferences and workshops in program/general/steering conference chairman positions and has organized numerous conferences related to computational science and engineering. He is a Fellow of IET and a Senior Member of the IEEE. His research interests include parallel and distributed computing, Big Data, and emerging technologies.

Xianjin Fang received the Ph.D. degree in computer application technology from Anhui University in 2010. He is a professor and Ph.D. supervisor at Anhui University of Science and Technology. His research interests include information security, data mining, etc.

Yan Zhang received the Ph.D. degree in computer application technology from Anhui University in 2010. She is a professor and Ph.D. supervisor at Anhui University of Science and Technology. Her research interests include information security, data mining, etc.

Ke Yang graduated from Huainan University of Technology with a bachelor's degree in 2001, and received the M.S. degree and Ph.D. degree from Anhui University of Science and Technology in 2004 and 2007. He is currently the dean of the School of Energy and Safety at Anhui University of Science and Technology, a second-level professor, and a doctoral supervisor. His research interests include data Analysis, mine pressure and control, special mining methods, and coal and gas co-mining.

Received: December 06, 2023; Accepted: March 01, 2024.

Computer Science and Information Systems 21(3):971–988 https://doi.org/10.2298/CSIS231111021V

MakeCode Arcade Platform for Game-Based Learning

Patrik Voštinár

Department of Computer Science, Tajovského 40
974 01 Banská Bystrica, Slovakia
patrik.vostinar@umb.sk

Abstract. Currently, there is an increasing demand for specialists in the field of computer science. However, for many children computer science is not as interesting as it could be. They learn many topics in a very short time and usually do not have the motivation to continue their studies in Computer Science. The solution could be teaching through games. In this paper, we describe our research into whether the MakeCode Arcade environment is suitable for teaching programming through games (game-based learning platform). In the research, we compare the responses of 315 male and female students and the responses of elementary and secondary school students. The research showed that boys liked the MakeCode environment more than girls and also that programming games in this environment were more meaningful to them as the girls. In this contribution, we describe also our experiences with using this environment for teaching in extracurricular activity at our university, in normal lessons in primary schools, as a motivational tool for distance workshops and also for workshops for teachers.

Keywords: MakeCode Arcade, children education, game-based learning, teaching programming.

1. Introduction

For the successful structuring of students' educational processes, the system of education necessitates the employment of new teaching technology. In today's digital era, smart education is a form of education which is becoming increasingly generally recognized and enthusiastically embraced adopted by younger generations [1].

We are currently facing technology all around us. More and more programmers are in demand to acquire, develop, and work with these technologies. Therefore, it is important that our education system produces more students who are able to program and think algorithmically. One possibility could be to use various types of computer games to motivate and also teach algorithm thinking.

Computational thinking (CT) has been actively promoted in schools in an integrative way or as a part of computer science subject, as it addresses concepts and learning goals of computer science. We may notice that interest in teaching computer science in primary school has increased during the last decade [2]. As recommended for computer science education in the report by the Committee on European Computing Education, "All students must have access to ongoing education in computer science in the school system. Computer science teaching should preferably start in primary school. . . " [3, p. 3]. Dagienė et al. [4] did research, about the worldwide tendencies of teaching CT through computing in primary education and primary school. The results of the quantitative study had shown

that CT and concepts related to computer science subject are introduced in the majority of surveyed countries (83%) in primary education. However, there are a lot of differences in the level of its implementation.

Computer games have a huge impact on a child's development. Actually, recent educational neuroscience researches show that the best way to teach children is through playing, getting their attention, their engagement, receiving feedback and consolidating their skills [5].

A special category are games that could also be used in the educational process in schools. It is important to select appropriate games that allow them to develop their personality. Examples of such games are Roblox and Minecraft Education. The authors of Minecraft Research: A Game as an Education and Scientific Learning Tool [6] examines how Minecraft can help teach different areas. They state that the diversity of biomes and terrain in Minecraft's world can help students understand ecology and geology. They say that the behavior of objects in Minecraft is similar to that in the real world. For example, trees need light, water, and space to grow, or sugar cane must live close to water [7]. Redstone logic uses the principles of logic gates (OR, AND, NOR, and XOR). For example, this simple logic allowed players to create a working computer in the game that could perform simple mathematical operations. The use of Redston is very easy for new players and it can be used to explain the basics of electrical engineering and logic circuits. With Minecraft: Education Edition, it is also possible to teach mathematics, chemistry, biology, history, art and design, English, and programming. Han, Liu, and Gao [8] focused in their research on the positive influence of using Roblox during the teaching process.

2. Gamification, Game-Based Learning and Serious games

There exist special terms when computer games are used in the educational process: gamification, game-based learning (GBL) education, and serious gameplay. The term gamification is generally used to denote the application of game mechanisms in non-gaming environments, such as earning points, levels, or badges. Gamification aims to transform systems, activities, organizations, and services to a system that uses the characteristics of game elements [9]. Gamification is a well known technique in education, organization engagement, crowdsourcing, commerce, information retrieval, and so on [10,11]. Gamification appears to be an emerging trend in many sectors: health, social policy, business, education [12]. Garcia-Iruela and Hijón-Neira in their study [13] focused on the perception that students have of the different elements of gamification in three different experiences, after which students were asked to answer a survey about the gamified elements of the course in which they participated.

Game-based learning is to present the instruction by games in learning, with the main purpose of triggering learners' motives instead of instructing the courses. GBL has been used to teach programming, any other topic, in a way that engages learners [14-16]. Hussain et al. [17] did research focusing on digital game-based learning methods to improve students' critical thinking skills in elementary science. The outcome of their study may help researchers, teachers, and practitioners in the educational sector to evaluate the effectiveness of serious games in facilitating students' critical thinking. Many researchers show that digital game-based learning is becoming a powerful tool in education, making learning more enjoyable, easier and efcient [18,19].

Edutainment means educating children using methods and tools that are traditionally associated with entertainment. Examples of this type of education are movies, television shows, and game consoles. Edutainment was popular during the 1990s with its growing multimedia PC market [20].

Serious games encompass the same goals as education but extend far beyond teaching facts and rote memorization, and instead include all aspects of education, teaching, training, and informing, and for all ages. Serious games can be applied to a broad spectrum of application areas, e.g. military, government, educational, corporate, and healthcare [21]. Christinaki et al. [22] did research, that serious games can be very effective in the areas of therapy and education for children with autism. Alamri et al. [23] created and used a serious game focused on problem of obesity.

Many other researchers are not only using Roblox and Minecraft, but they teach programming through many other games or game environments. Jesus and Silveira in their research [22] used the collaborative learning concept in secondary school students through a digital games programming workshop. Many types of research [24-29] show that teaching programming games for students could be motivating for students. It is also very good for the imagination, for developing children's computational thinking.

Another interesting environment for game-based learning is the MakeCode Arcade environment. Siever and Rogers conducted research [30] about this interesting environment, which can create games, that can be deployed to lowcost consoles, comparable to Nintendo's Gameboy. The camp used MakeCode Arcade as the programming language while still focusing on game development. Brown et al. [31] conducted a survey of interface representations in visual programming language environments for children's physical computing kits that shows that MakeCode Arcade is a suitable environment for teaching programming. Begel et al. [32] developed instructional materials to teach computer programming, video game design, and communication and teaming skills through the MakeCode Arcade environment. Students used the MakeCode Arcade development environment to build their games and Zoom to remotely collaborate with their teammates. In summative interviews, students reported improved programming skills, increased confidence in communication, and better experiences working with others. They found that students valued the opportunity to practice teamwork, such as being more vocal in expressing ideas to their peers and working out differences of opinion with their teammates. We described our pilot research and experience with MakeCode Arcade in conference Edunine [33].

3. MakeCode environment for teaching programmings

The Microsoft company started developing the MakeCode Arcade environment in 2018. The first beta version appeared in 2019 and the official version started in 2020. Microsoft MakeCode is an online web environment that offers visual programming language, JavaScript, and Python language to program educational board micro:bit, robots Lego EV3 and Cue, and the game Minecraft: Educational Edition, MakeCode Arcade, and Adafruit Circuit Playground. The environment is available at https://arcade.makecode.com/

MakeCode is not a universal programming platform, but an extensible framework to create any number of MakeCode based experiences customized to a specific target. These

targets are typically makerstyle or educational singleboard computers, but could also be a simulated only 'soft' target.

The concept of each device, that provides the possibility to program in the MakeCode environment, is the same - it offers many tutorials on the main page, the essentials of a MakeCode editor consist of switching buttons between programming languages, an interactive simulator, target code button, and blocks/language editor. The only difference is in the visual site (different colors for each device).

Fig. 1 shows the MakeCode Arcade page, which contains the simulator on the left side, under the simulator is the target code button (after pressing this button a popup window will appear with a list of available hardware devices, consoles), next to the simulator is a palette of blocks, and on the right side is the editor of visual code programming.

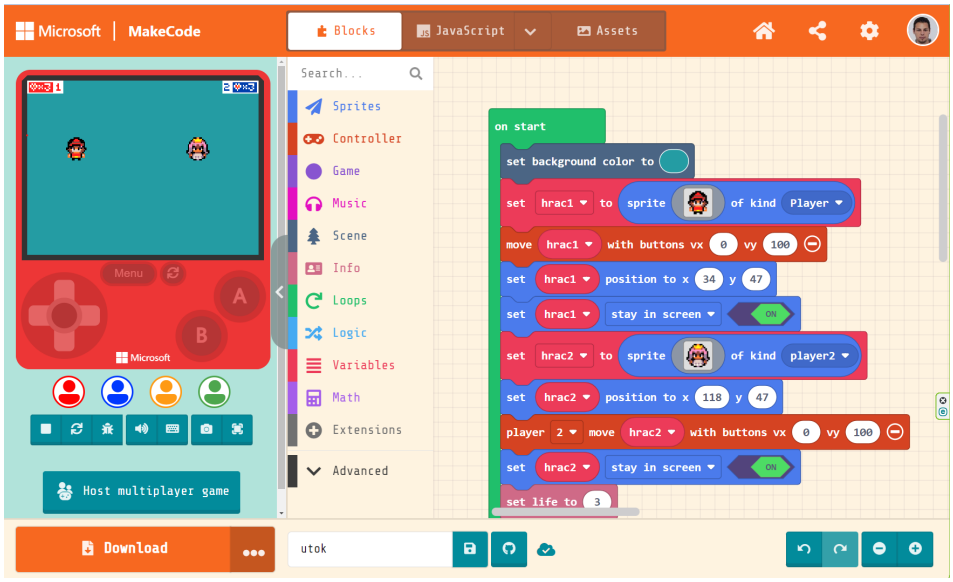


Fig. 1. The MakeCode Arcade environment

The advantage of this environment is the ability to test the created 2D games in the simulator, on a special game console (approx. 35 €), and on mobile devices. The interactive simulator provides students immediate feedback on how their program is running and makes it easy to test and debug your code. The simulator can be controlled with its buttons, with the computer keyboard keys W, S, A, D and space, or with a gamepad connected to the computer (e.g. Xbox gamepad). After pressing the download button, located under the simulator, it is possible to choose from many types of game consoles intended for MakeCode Arcade. Fig. 2 shows multiple game consoles for MakeCode Arcade games.

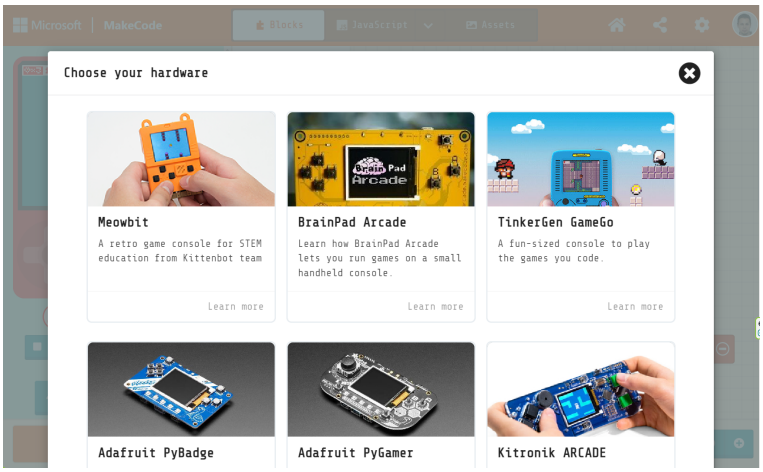


Fig. 2. The MakeCode Arcade – a hardware for testing

The third option is to test the game on a mobile device by pressing the share button (on the orange top panel, the third button). It will create a unique URL, which can be copied to the mobile browser (through QR or just through the URL link). A screen with a game console (buttons, display) will appear on the mobile device. This screen looks and behaves exactly like a game console.

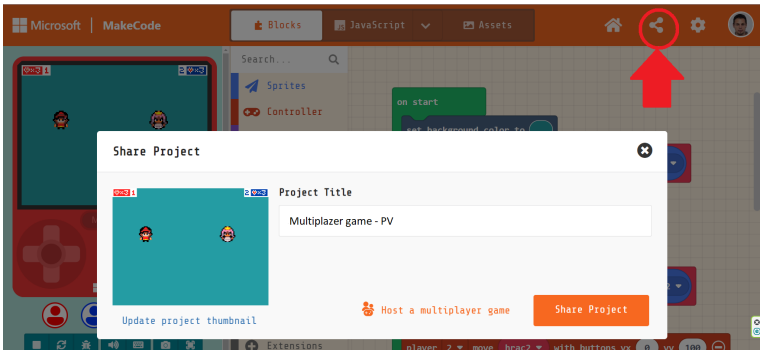


Fig. 3. The MakeCode Arcade – sharing a project for mobile devices

The MakeCode Arcade environment contains many instructions for creating different types of games, but also for teaching programming. Compared to the Scratch environment, MakeCode Arcade focuses more on game principles (velocity, status bar or health bar of sprites, working with scenes, connection to physical game consoles) than Scratch. Scratch is a more universal environment. Compared to other game development environments, this environment is best suited for primary school teaching (it uses block programming that students are used to). Another advantage is that we don't have to log in or

install to computers. Other environments for teaching game programming, e.g. Unity, Unreal Engine, Godot Engine require a more advanced level of programming (more suitable for secondary schools or universities). The environments eBlock, Scratux, Code Kit are based on the Scratch environment and, like Scratch, are more universal environments for teaching programming than for programming games. Another advantage of MakeCode Arcade is that Microsoft created a curriculum, a full academic year course designed to be taught for introductory programming classes in secondary schools. The curriculum was adapted as a core teaching platform using a game-based learning approach to engage a diversity of students. The original TEALS course has been successfully implemented in hundreds of secondary schools. The course is divided into a series of units which successively introduces new concepts and reinforces those learnt previously. The course starts with teaching programming using visual drag-and-drop blocks and then JavaScript or Python. During the course, students learn basic principles of games, programming concepts, creating various types of biomes, working with coordinates, and advanced graphics techniques. Students work in teams (collaboration). The course is available on the official website https://arcade.makecode.com/courses/introcs.

3.1. Examples of educational games

On the official website of MakeCode Arcade, there are several games that we can use for educational purposes. Creating educational games in this environment is quite easy, so it is possible to create different types of educational games.

Educational games - we can create different educational games for geography (different biomes), create history (prehistoric period - etc.), math coordinates/examples, practice language learning, observation, plays focused on significant works of literature (e.g. Hamlet), etc. We can also create different games for multiplayer cooperation (competition for button pressing speed, who gets clovers first, etc.). Games aimed at explaining game design concepts, explaining the principles of various graphic algorithms, fibonacci tiles, arts, etc. Fig. 4 shows existing games for game design concepts and graphic (Setting the Scene, Tale of Talagron, Fibonacci Tiles, etc.).

We can run games created on the MakeCode Arcade platform on a computer, mobile device, specialised game consoles, or we can create our own game console according to the instructions (basics of electrical engineering). Fig. 5 shows tutorials for DIY (Do it yourself) hardware. In the Shoebox Controller tutorials, you can make a game console from a shoebox, Arcade cabinet - make a game console from wood, etc.

3.2. Programming games

If the created games are not enough for us, we can simply create our own game. Creating games is very easy even for non-programmers. We have tutorials for each type of game. We can create games in blocks, JavaScript or Python.

We will explain the principle of creating a simple game for two players in a game where two players will throw fruit at each other. Figure 6 shows the code blocks for creating two characters. The picture shows a code block without the principles of object-oriented programming. A more suitable version is if we divide the creation of characters into several functions (one function is for one character). The students themselves will understand that object-oriented programming is easier to read the code.

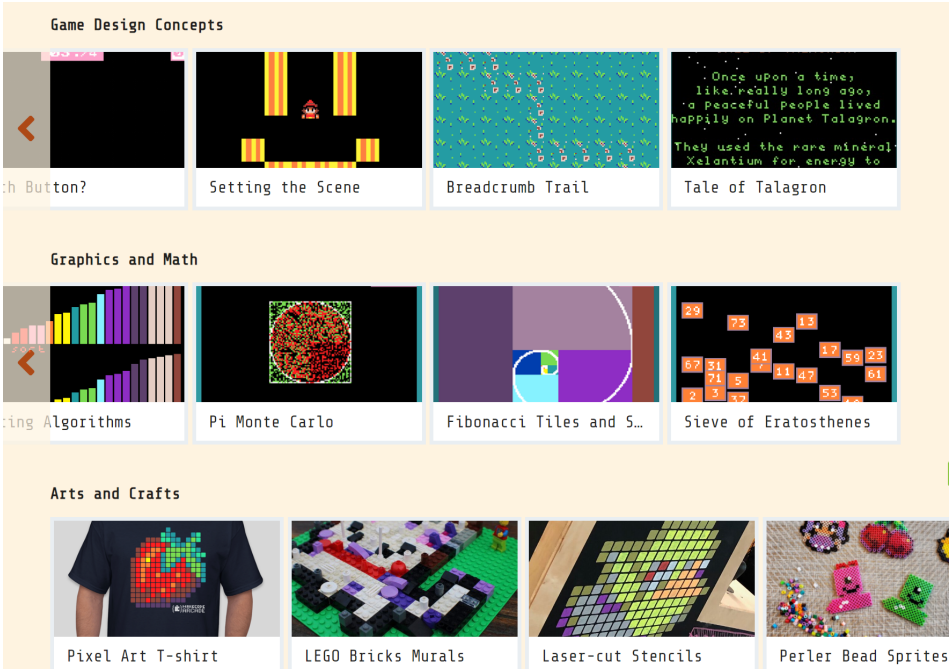


Fig. 4. Games for game design concepts and graphic

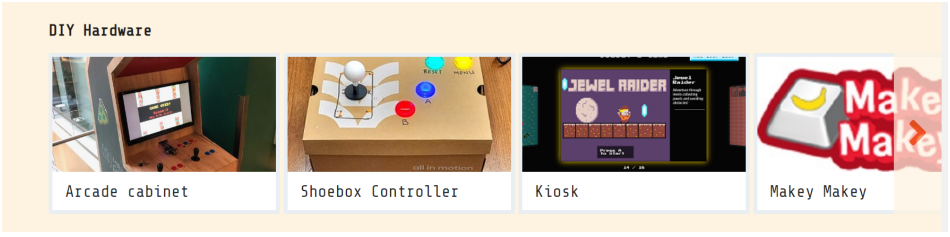


Fig. 5. DIY hardware



Fig. 6. Blocks code - creating two characters

Figure 7 shows a block of code that, after pressing the A button creates a projectile, which will appear on player 1 (boy) and has the velocity x (vx set to 100) and velocity y (vy set to 0), projectile will fly horizontally with speed 100 pixels per second.

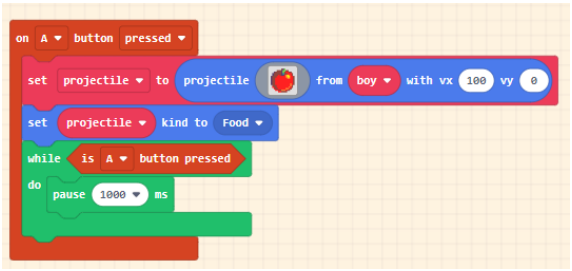


Fig. 7. Blocks code - events

The last step is to create behavior, what will happen when Player 1 (boy) overlaps the projectile Food2 (which was thrown by Player 2-girl) and the same when Player 2 (girl) overlaps the projectile Food (which was thrown by Player 1-boy). After a player overlaps one of the projectiles, we will destroy the projectile and increase the score of the player who touched it. Figure 8 shows objects overlaping behaviour.

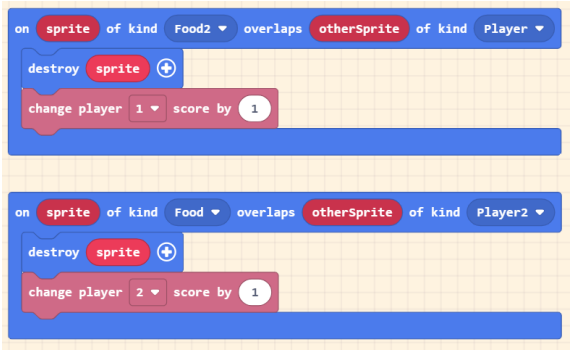


Fig. 8. Blocks code - overlaping behaviour

In this step, it is very important to remove the projectile from the game. If we do not remove it, it will still exist in the memory of the game device, and after a certain time the game console (or computer) will start to freeze.

4. Teaching through the MakeCode Arcade environment

We started using the MakeCode Arcade environment for teaching in an extracurricular activity for primary schools, which we organized in our department at the university. After testing the environment in the extracurricular activity, we started teaching with this environment in primary school in Banská Bystrica, secondary school in Žiar nad Hronom and Zvolen and later then on special online activity propagating computer science called IT tea room. After that we also showed this environment in an activity called The club of computer science teachers in Banská Bystrica (training for teachers).

4.1. Computer science extracurricular activity

In our Department of Computer Science FNS UMB, we are organizing free extracurricular activities for primary school students since the academic year 2017/2018. The activity is organized twice a week for 1,5 hours. The students of the Teaching trainee of the Computer Science program and Applied Computer Science program are involved as lecturers or tutors in this activity. Each of our students teaches 1-2 children from primary school. The main focus of this activity is to obtain additional teaching experience for our students and increase motivation to study computer science (programming). Each year we have approximately 10-13 for each group (one group is on Wednesday and another group on Thursday), so together 20-26 students.

In our extracurricular activity, we are using many types of educational aids such as micro:bit, Ozobot, mBot, lego EV3, programming drones, games, webpages, etc. We also use many types of tasks. Since February 2020, schools have been closed due to the Covid-19 pandemic. We had to switch to online teaching through Microsoft Teams, and we decided to use the online environment MakeCode Arcade for teaching game programming, mainly because students do not need to install any hardware on their computers.

First we started to program according to tutorials from official webpage:

- Chase the Pizza
- Happy flower
- Lemon Leak
- Galga
- Barrel Dodger
- Maze

The tutorials were quite easy for the students who had previous experience programming many educational robots. After each tutorial was finished, we started improvising our lectures after each game we asked students to brainstorm on how we could improve. We also added to our tutorials some additional enemies, some items that cannot be touched and some items (for example life) that appear only a few seconds in random positions, after some score we added boss (alien) enemy, which have to be shot more times, etc.

After that we started to create our own story for games, a multiplayer game, and many other games.

students were happy with the classes, visual block programming was nothing new for them. The observation shows that this environment is suitable for teaching programming and also for developing your creativity. We also sent them an anonymous questionnaire after we finished coding in this environment - we used this environment 6 months. Fig. 9 shows teaching the extracurricular activity.

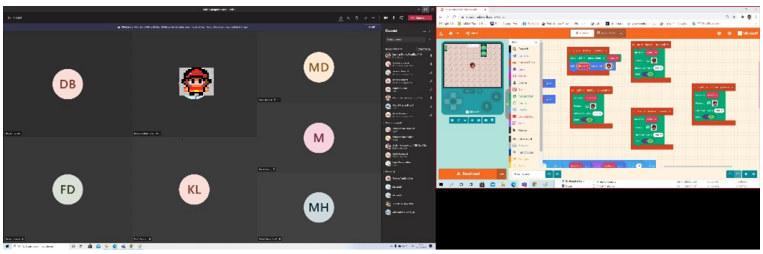


Fig. 9. The extracurricular activity – online teaching

4.2. Primary school

After our experience with programming in the environment MakeCode Arcade in our extracurricular activities, we decided to use it in the normal teaching hours of the primary school Ďumbierska 17 in Banská Bystrica. We started teaching with this environment from April 2020 in 6 classes (13-15 years old children, 45 minutes for each class weekly). During three months of online teaching in this environment we can say, based on our observations, that the students liked this environment because it was something different for them than programming in Scratch. For them, the previous experience with subject Computer Science was mainly programming in Scratch. Since the academic year 2019/2020, they started using educational board micro:bit for teaching of programming, so they were familiar with the MakeCode environment.

For these lessons, we had selected the most interesting games from the extracurricular activity (Chase the Pizza, Lemon Leak, Galga, Maze, Multiplayer game, Shark attacing,

Simon says, etc.). After they familiarized with MakeCode Arcade, they had to work in 4 person groups to create their own game. They worked in MS Teams breakout rooms, so during the lessons we could check their progress. Most of the groups also worked in their spare time and created interesting games.

In the end, the teams had to present their games to their classmates. After that they were evaluated.

4.3. Secondary schools

As part of the special project called Harpuna (the project focuses on increasing children's soft skills in the Computer Science subject), we also had the opportunity to teach in two secondary schools, where we taught computer science during the academic year 2021/2022 for 6 months. We had two classes available at the Milan Rúfus Secondary School in Žiar nad Hronom and one class at the Ľudovít Štúr Secondary School in Zvolen. Both schools are eight-year secondary schools that prepare students from 11 to 18 years of age. In both schools, we taught programming using micro-bit educational boards, game programming in MakeCode Arcade, and mobile application programming in App Inventor (each topic for about 2 months). At the beginning of the lesson, we showed the environment and some examples of the application. Then the students worked in groups of 4 on their own projects. During the lessons, we continuously helped them if necessary.

4.4. "IT čajovňa" and "Veda v meste" ("IT teahouse" and "Science in the city")

In January 2021the Gymnázium Sama Chalupku in Brezno requested that our department teaches a motivational online workshop for computer science. At this time, we were also requested by the primary school Trieda SNP in Banská Bystrica to teach the activity called "Veda v meste" ("Science in the city").

In both schools, we conducted two-hour motivational workshops for children 14-15 years of age with topic programming games in the MakeCode Arcade. In Brezno we had 30 children and in Trieda SNP we had 29 children. We showed them simple games similar to Space Invaders. Based on comments from children during the workshops and after the workshops from their teachers, the students were excited. In February 2021, we were approached by the national IT Academy project (activity called "IT čajovňa" – "IT tea house") to do IT workshops/lectures for primary and secondary school students. As part of these workshops, we conducted a series of four workshops for primary school students and a series of two workshops for secondary school students, focusing on game programming in the MakeCode Arcade environment. The workshops lasted 2 to 2.5 hours and they were recorded so that the students who could not attend, could watch them afterwards. The difference between the workshop content for primary and secondary school was in the level of difficulty: in the workshops for primary school we programmed simpler games then in the workshop for secondary school where we programmed more advanced games with animations. A total of 47 secondary school students and 126 primary school students participated in the workshops. There were more different topics available for secondary schools, so fewer students participated than in workshops from primary schools. From the organizers, we received continuous feedback in comments and after the workshops. We also asked the students to fill out a questionnaire on how they liked the workshop. The

students were also interested in the workshops because we held six different workshops where they programmed six different games.

Fig. 10 shows the online workshops in the MakeCode environment.

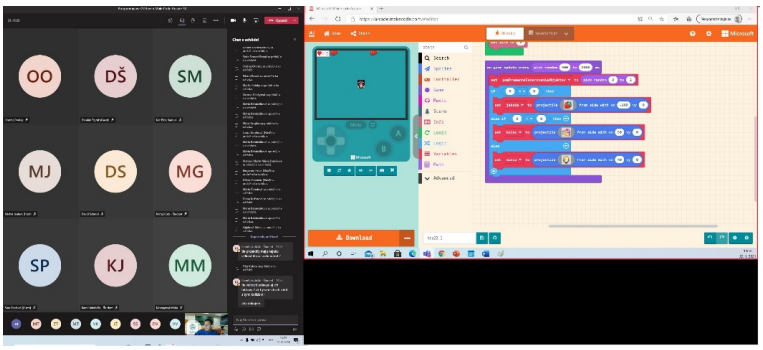


Fig. 10. The online workshops teaching

4.5. Club of computer science teachers

Our Department of Computer Science, FNS, UMB started organizing in January 2020 a training with new technologies for teachers. In this activity, we focus on new popular technologies, ICT devices and innovative software, to teach computer science. Based on previous positive feedback from the MakeCode Arcade, we decided to show teachers this environment, which is suitable for presence and distance learning.

In November 2020 we organized the first four hours of the online training for primary school teachers in the Banská Bystrica area. We showed them the environment and together we programmed three simple games. After the training, we asked them how did they like this environment and if they want to have another training on the same topic, but with new games.

Based on the positive feedback from the questionnaire, we organized in March 2021 another training with MakeCode Arcade with new games. Both workshops were attended by 10 teachers. The teachers were very enthusiastic about the MakeCode Arcade platform. They liked that they could use it to explain the principles of programming, graphics, and also that they could very easily create their own educational games that they could use in teaching various subjects.

5. Methodology and Research

We started our research in the academic year 2020/21. In our research, we used the MakeCode Arcade environment for learning programming in our extracurricular activity at our university during the Covid-19 pandemic, in classic primary school lectures and as a topic in motivational workshops. We also verified this environment by discussing the advantages and disadvantages of this environment with primary school teachers.

We used a questionnaire to measure the data of our research. 315 children - 178 boys and 137 girls, filled out the anonymous questionnaire using Google Forms. The questionnaire consisted of 12 questions. The first three questions focused on gender, the name of the school, and the grade of the pupil visiting the school at the time when filling out the questionnaire. The questionnaire was filled out by 176 students from secondary schools and 139 students from primary schools. All students first completed the same introduction to the MakeCode Arcade environment. Secondary school students programmed more challenging games than primary school students in the same environment (we even tried Python with some groups). A questionnaire of our design was used to obtain empirical data, which were processed in Microsoft Excel using descriptive statistics methods. In evaluating the responses to the chosen items, we utilized the Chi-Square test.

In our research, we set several hypotheses, which we investigated, e.g., the differences between the answers of boys and girls and the differences between the answers of students from primary and secondary schools.

The first question that interested us was whether they had ever programmed games (e.g. in Scratch). 202 students (64.1%) answered that they did and 113 students (35.9%) answered that they had never programmed any games. In the next question, we asked if they had ever programmed games in the MakeCode Arcade environment - 297 students (94.3%) answered that they had never programmed in this environment. Only 18 students answered that they had already tried this environment.

Using the questionnaire item, *Were these activities meaningful to you?* we investigated the difference between the answers of boys and girls, students from primary and secondary schools. Table 1 shows the hypotheses for this questionnaire item.

Table 1. Hypotheses for the questionnaire item *Were these activities meaningful to you?*

H_0	H_1	Probability of error	The null hypothesis
There are no significant differences in boys' and girls' responses about whether they thought the activities with MakeCode Arcade were meaningful.	There are significant differences in boys' and girls' responses about whether they thought the activities with MakeCode Arcade were meaningful.	0,62%	reject
There are no significant differences in the responses of primary and secondary school students as to whether they thought the activities with MakeCode Arcade were meaningful.	There are significant differences in the responses of primary and secondary school students as to whether they thought the MakeCode Arcade activities were meaningful.	24,62%	accept

Based on the arithmetic mean, boys think that the activities in the MakeCode Arcade environment were more meaningful to them than girls. There are no differences in the answers of students from primary and secondary schools to the question whether the activities were meaningful to them. The answers were positive, both groups think that these activities were meaningful for them.

Using the questionnaire item, *Did you like programming games in MakeCode Arcade?* we investigated the difference between the answers of boys and girls, students from primary and secondary schools. Table 2 shows the hypotheses for this questionnaire item.

Table 2. Hypotheses for the questionnaire item *Did you like programming games in MakeCode Arcade?*

H_0	H_1	Probability of error	The null hypothesis
There are no significant differences in the answers of boys and girls about whether they liked programming in MakeCode Arcade.	There are significant differences in the answers of boys and girls about whether they liked programming in MakeCode Arcade.	0,07%	reject
There are no significant differences in the answers of students from primary and secondary schools, whether they liked programming in MakeCode Arcade	There are significant differences in the answers of students from primary and secondary schools, whether they liked programming in MakeCode Arcade	38,52%	accept

Based on the arithmetic mean, boys liked game programming more than girls. There are no differences in the answers of students from primary and secondary schools to this question. The answers were positive, both groups like programming in this environment.

Using the questionnaire item, *Were the activities difficult for you?* we investigated the difference between the answers of boys and girls, students from primary and secondary schools. Table 3 shows the hypotheses for this questionnaire item.

Table 3. Hypotheses for the questionnaire item *Were the activities difficult for you?*

H_0	H_1	Probability of error	The null hypothesis
There are no significant differences in the answers of boys and girls, whether programming in MakeCode Arcade was difficult.	There are significant differences in the answers of boys and girls about whether programming in MakeCode Arcade was difficult.	42,09%	accept
There are no significant differences in the responses of students from primary and secondary schools as to whether programming in MakeCode Arcade was difficult.	There are significant differences in the answers of students from primary and secondary schools, whether programming in MakeCode Arcade was difficult.	63,89%	accept

There were no differences in the responses of primary and secondary school students, boys and girls, as to whether the activities were difficult for them. The answers showed that the activities were easy for them.

Using the questionnaire item, *Would you like to continue with MakeCode Arcade?* we investigated the difference between the answers of boys and girls, students from primary and secondary schools. Table 4 shows the hypotheses for this questionnaire item.

Table 4. Hypotheses for the questionnaire item *Would you like to continue with MakeCode Arcade?*

H_0	H_1	Probability of error	The null hypothesis
There are no significant differences in the answers of boys and girls as to whether they would like to continue teaching programming in MakeCode Arcade.	There are significant differences in the answers of boys and girls as to whether they would like to continue teaching programming in MakeCode Arcade.	00,00%	reject
There are no significant differences in the answers of students from primary and secondary schools, whether they would like to continue teaching programming in MakeCode Arcade.	There are significant differences in the answers of students from primary and secondary schools, whether they would like to continue teaching programming in MakeCode Arcade.	9,37%	accept

Based on the arithmetic mean, boys would like to continue programming games in the MakeCode Arcade environment more than girls. There are no differences in the answers of students from primary and secondary schools to this question. The answers were positive, both groups would like to continue in this environment.

In another open question, we asked what games they would like to program. Most of the answers were I don't know or I don't care, in addition to these answers, students would like to program games in the style of Super Mario Bros and Pac Man, single and multiplayer games, RPG games. To the last question, whether they wanted to say something to the authors, the students again mostly answered no, or that they were thankful for the opportunity to try game programming.

6. Discussion

Programming environment MakeCode is a very popular programming environment for teaching with educational board micro:bit, Lego EV3, and also for programming games in MakeCode Arcade. The main advantage of this environment is that it is free and that MakeCode developers keep adding new features. We used this environment for teaching programming and our small research confirms that this environment is suitable for teaching programming in Slovak schools. Unfortunately, there is not so much research on using this environment, but our small research confirmed similar results to Siever, Rogers, and Brown, the students liked this environment (more than Scratch), it is suitable for promotional activities for computer science, it is suitable for explaining the creation of scenes, game design principles. Comparing the answers of our students with the answers of the students from the research Siever and Rogers and Brown the students liked this environment and would like to continue in it.

We also verified the suitability of using the MakeCode Arcade environment in the teaching process at a free training for computer science teachers. 10 teachers from different schools participated in this training. After the 8-hour lecture, we asked the teachers for their opinion and had them fill out a short questionnaire. When asked whether the training met expectations, 9 teachers answered yes and one teacher rather yes. In the next question, we asked if they would use the tasks created by us for this environment in the teaching process - 5 teachers answered yes and 5 teachers rather yes. When asked if they think this environment is suitable for teaching, they all answered yes.

7. Conclusion

In this contribution, we describe our experience with the online environment MakeCode Arcade, which, in our opinion, is a suitable environment for teaching programming and for motivational activities to increase the interest in programming. This environment contains many created games that can be used for game-based learning. We tested this environment in the education process on selected students who are interested in programming (extracurricular activity and motivational online workshops), but also in normal lessons for primary school children. The suitability of this environment was confirmed by a questionnaire from participants (students) and by teachers from primary schools who completed a training at our department for this environment. Based on our experiences we will prepare more MakeCode Arcade tasks and we will continue with this environment in our activities to motivate children to study computer science. Teachers like this environment, as they can explain different game principles, the basics of graphics, animations. They can use one of the created educational games, or create their own game as needed.

Acknowledgments. This contribution has been processed as part of grant project no. 001UMB-4/2023 Implementation of blended learning in the training of professional bachelors and teachers of mathematics and computer science.

References

1. Omonayajo, B., Al-Turjman, F., Cavus, N.: Interactive and Innovative Technologies for Smart Education. Computer Science and Information Systems, Vol. 19, No. 3, 1549-1564. (2022), https://doi.org/10.2298/CSIS210817027O
2. Juškevičienė, A., Passey, D.: Outcomes of Computing Education. In: Tatnall A. (eds) Encyclopedia of Education and Information Technologies. Springer, Cham (2019).
3. The Committee on European Computing Education (CECE): Informatics Education in Europe: Are We All in the Same Boat? 2017 Technical Report. ACM, New York, NY, United States (2017). DOI: https://doi.org/10.1145/3106077
4. Dagienė, V., Jevsikova, T., Stupurienė, G., Juškevičienė, A.: Teaching Computational Thinking in Primary Schools: Worldwide Trends and Teachers' Attitudes. Computer Science and Information Systems, Vol. 19, No. 1, 1-24. (2022), https://doi.org/10.2298/CSIS201215033D
5. Sigman, M., Peña, M., Goldin, A., Ribeiro, S.: Neuroscience and education: prime time to build the bridge. Nat. Neurosci. 17, 497–502 (2014)
6. Sălceanua, C.: The Influence of Computer Games on Children's Development. Exploratory Study on the Attitudes of Parents. In Procedia – Social and Behavioral Sciences, 149, 837-841. (2014)

7. Ekaputra, G.,Lim, Ch., Eng, K., I.: Minecraft: A Game as an Education and Scientific Learning Tool. In Information Systems International Conference (ISICO). (2013)
8. Han, J., Liu, G., Gao, Y.: Learners in the Metaverse: A systematic review on the use of roblox in learning,". In Education Sciences. Vol. 13, No. 3. (2023)
9. The Children of the Cost, Quality, and Outcomes Study Go To School, Available : http://fpg.unc.edu/sites/fpg.unc.edu/files/resources/reports-and-policy-briefs/NCEDLCQOtechnicalreport.pdf. (current May 2019)
10. Huotari, K., Hamari, J.: A definition for gamification: Anchoring gamification in the service marketing literature. In: Electronic Markets, 27(1), 21–31, (2017).
11. Warmelink, H., Koivisto, J., Mayer, I., Vesa, M., Hamari, J.: Gamification of production and logistics operations: Status quo and future directions. In: Journal of Business Research. (2018).
12. Caponetto, I., Earp, J., Ott, M.: Gamification and education: A literature review. European Conference on Games Based Learning. In Academic Conferences International Limited. Vol. 1. (2014)
13. Garcia-Iruela, M., Hijón-Neira, R.: What Perception Do Students Have About the Gamification Elements?. In IEEE Access. Vol. 8, 134386-134392, 2020)
14. Lai, C. H. , Lin, Y. C., Jong, B. S., Hsia, Y. T.: Adding Social Elements to Game-Based Learning. In International Journal of Emerging Technologies in Learning (iJET). Vol. 9, 12–15. (2014)
15. Schaffer, D., W., Squire, K., D., Halverson, R., Gee, J. P.: Video games and the future of learning. In Phi Delta appan. Vol. 87, pp. 104-111. (2005)
16. Lamrani, R., Abdelwahed, E. H.: Game-based learning and Gamification to improve skills in early years education. Computer Science and Information Systems, Vol. 17, No. 1, 339-356. (2020), https://doi.org/10.2298/CSIS123456789 94833333
17. Hussein, M. H., Ow, S. H., Cheong, L. S., Thong, M. -K.: A Digital Game-Based Learning Method to Improve Students' Critical Thinking Skills in Elementary Science. In IEEE Access. Vol. 7, 96309-96318. (2019)
18. Boyle,, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., Lim, T., Ninaus,M., Ribeiro, C., Pereira, J.: An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. Computers and Education, Vol. 94, 178–192. (2016)
19. Hafeez, M.: Efects of game-based learning in comparison to traditional learning to provide an efective learning environment: A comparative review. Contemporary Educational Researches Journal, Vol. 12, No. 2, 89–105. (2022)
20. Michael, D. Chen, S.: Serious games Games that educate, train, and inforn. Boston, MA.: Thomson Course Technology. (2005)
21. Susi, T., Johannesson, M., Backlund, P.: Serious Games : An Overview. In Retrieved from Institutionen for kommunikation och information. (2007)
22. Christinaki, E., Vidakis, N., Triantafyllidis, G.: A Novel Educational Game for teaching Emotion Identification Skills to Preschoolers with Autism Diagnosis. Computer Science and Information Systems, Vol. 11, No. 2, 723–743. (2014), https://doi.org/10.2298/CSIS140215039C
23. Alamri, A., Hossain, M. A., Hassan, M. M., Hossain, M. S., Alnuem, M., Ahmed, D. T., Saddik, A. E.: A Cloud-Based Pervasive Serious Game Framework to Support Obesity Treatment. Computer Science and Information Systems, Vol. 10, No. 3, 1229-1246. (2013), https://doi.org/10.2298/CSIS120717046A
24. Malliarakis, C., Satratzemi, M., Xinogalos, S.: Optimization of server performance in the CMX educational MMORPG for Computer Programming. Computer Science and Information Systems, Vol. 11, No. 4, 1537–1553. (2014), https://doi.org/10.2298/CSIS131220078M
25. de Jesus, Â., M., Silveira, I., F.: A Collaborative Learning Framework for Computational Thinking Development through Game Programming. In Informatics in Education. Vol. 21, No. 2, 253-281. (2022)

26. Dele-Ajayi, O., Strachan, R., Pickard, A. J., Sanderson, J. J.: Games for Teaching Mathematics in Nigeria: What Happens to Pupils' Engagement and Traditional Classroom Dynamics?. In IEEE Access. Vol. 7, 53248-53261. (2019)
27. Kian, T. W., Sunar, M. S., Su, G. E.: The Analysis of Intrinsic Game Elements for Undergraduates Gamified Platform Based on Learner Type. In IEEE Access. Vol. 10, 120659-120679. (2022)
28. Kazimoglu, C.: Enhancing Confidence in Using Computational Thinking Skills via Playing a Serious Game : A Case Study to Increase Motivation in Learning Computer Programming. In IEEE Access. Vol. 8, 221831-221851. (2020)
29. Montes, H., Hijón-Neira, R., Pérez-Marin, D., Montes, S.: Using an Online Serious Game to Teach Basic Programming Concepts and Facilitate Gameful Experiences for High School Students. In IEEE Access. Vol. 9, 12567-12578. (2021)
30. Siever, B., Rogers, M., P.: Game On! Inspired CS Education with MakeCode Arcade. Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE '21). In Association for Computing Machinery. New York, NY, USA. (2021)
31. Brown, S. A., Chu, S. L., Yin, P.: A Survey of Interface Representations in Visual Programming Language Environments for Children's Physical Computing Kits. Interaction Design and Children (IDC '21). In Association for Computing Machinery, New York, NY, USA, 268–275 . (2021)
32. Begel, A., Dominic, J., Phillis, C., Beeson, T., Rodeghero, P.: How a Remote Video Game Coding Camp Improved Autistic College Students' SelfEfficacy in Communication. In Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE '21). Association for Computing Machinery, New York, NY, USA, 142–148. (2021)
33. Voštinár, P.: MakeCode Arcade: Interesting environment for programming 2D games. In 2021 IEEE World Conference on Engineering Education (EDUNINE). 1-6. (2021).

Patrik Vostinar is a young researcher who received his Ph.D. with a dissertation thesis "Multimedia as a motivational instrument in preparation for a Math teacher". His research mainly focuses on training pre-service computer science teachers with innovative robotics aids, e-learning, mobile learning, microcontrollers, 3D modelling, 3D printing and developing virtual reality environments. Since September 2017, he has been an Assistant Professor in the Department of Computer Science, where he gives lectures and seminars in courses on programming, virtual reality, and many subjects related to teacher trainee students. He is the author and co-author of more than 78 publications in scientific journals and proceedings of scientific conferences and the author of two textbooks. He organises free extracurricular activities for students in primary schools, which aim to teach children programming for various types of robotics, programming, 3D printing, and creating virtual reality games. Moreover, he leads a training program for in-service teachers focusing on new popular technologies, ICT devices, e-learning and innovative software applicable to computer science education. http://www.fpv.umb.sk/en/pvostinar/

Received: November 11, 2023; Accepted: March 19, 2024.

Computer Science and Information Systems 21(3):989–1012 https://doi.org/10.2298/CSIS231225025Y

News Recommendation Model Based on Encoder Graph Neural Network and Bat Optimization in Online Social Multimedia Art Education

Jing Yu[1], Lu Zhao[1,⋆], Shoulin Yin[2], and Mirjana Ivanovic[3]

[1] Lu Xun Academy of Fine Arts
Shenyang 110004, China
ohyou@sina.com
zzll_201@foxmail.com
[2] Software College, Shenyang Normal University
Shenyang 110034, China
yslin@hit.edu.cn
[3] Faculty of Sciences, University of Novi Sad
21000 Novi Sad, Serbia
mira@dmi.uns.ac.rs

Abstract. At present, the existing news recommendation system fails to fully consider the semantic information of news, meanwhile, the uneven popularity of news will also cause the phenomenon of long tail. Therefore, we propose a novel news recommendation model based on encoder graph neural network and Bat optimization in online social networks. Firstly, Bat optimization algorithm is used to improve the effect of news clustering. Secondly, the concept of metadata is introduced into the graph neural network, and the ontology of learning resources based on knowledge points is established to realize the correlation between news resources. Finally, the model combining Convolutional Neural Network (CNN) and attention network is used to learn the representation of news, and Gate Recurrent Unit (GRU) is used to learn the short-term preferences of users from their recent reading history. We carry out experiments on real news datasets, and compared with other advanced methods, the proposed model has better evaluation indexes.

Keywords: news recommendation system, encoder graph neural network, Bat optimization, online social networks, GRU.

1. Introduction

With the rapid development of network environment, information technology has been widely used, and people's news reading habits have gradually shifted from the traditional media such as newspaper and TV to the Internet. However, people's ability to read news information is often lower than the speed of news production, and the news information environment is becoming increasingly dense, eventually leading to "information overload". Therefore, how to efficiently, quickly and effectively recommend news text resources of interest to users has become increasingly important [1,2].

⋆ Corresponding author

In recent years, personalized news recommendation has gradually become the hot spot of Web technology, and the recommendation system has a large number of applications in real-time information, news, microblog, movie, music, blog, e-commerce and so on [3]. Through the recommendation system, problems such as information overload can be quickly, efficiently and accurately dealt with, users' interest rating results on news content and user behavior characteristics can be analyzed, and users' interest model can be automatically constructed by using historical data. There is no need to deliberately conduct interest surveys on users, which greatly and effectively reduces the burden on users and greatly improves the popularity of the recommendation system [4].

Personalized recommendation function is also very common in the mobile Internet industry, today's headlines, Meituan public point rating network have personalized recommendations [5]. Although people have carried out in-depth technical research and development of recommendation algorithms for many times, the existing recommendation algorithms still have limitations: (1) News has strong timeliness. News platforms automatically generate a lot of new news every day, and old news will quickly disappear. This causes a large degree of cold start problem, resulting in many user behavior dependent recommendation methods such as collaborative filtering can not be used; (2) News articles have a rich body, which contains important information that is not used, and cannot be directly represented by ID and other features simply and intuitively; (3) Accurately model user interest. Users' interest is diverse and changes over time, which requires a lot of mining and modeling based on user feedback behavior. But there is often no explicit user feedback on news platforms, and even implicit feedback is sparse. Therefore, personalized news recommendation has become a research hotspot of many scientific research institutions at home and abroad, and has received high attention from domestic and foreign academic conferences in the fields of information retrieval, data mining and artificial intelligence [6,7].

In order to alleviate information overload and meet users' reading needs, many news platforms, such as Google News [8], apply recommendation system technology to recommend news and push personalized article content to users, which can significantly improve users' article click rate and reading satisfaction, and improve user experience. The potential value brought by news recommendation is multifaceted. In addition to intuitively increasing the user click rate and reading probability, the recommendation system can continuously meet the needs of users' preferences, increase user stickiness, and cultivate user loyalty for media platforms. Compared with other recommendation systems [9], news recommendation is not only affected by problems such as cold start and sparse data, but also faces unique problems in this field, such as rapid iteration of news content, unstructured data generated by news text, and uneven popularity of news. Therefore, researchers need to apply the right medicine and propose effective solutions to the actual problems in the field of news recommendation [10].

To solve the above problems, this paper divides the user's interest into multiple dimensions for learning when dealing with the user's long-term reading history, combines the user's own attribute information, and uses the attention mechanism to distinguish the influence of each dimension, so as to extract the user's interest in different dimensions in the reading history over a long period of time as the user's long-term preference. In addition, this paper uses the Convolutional Neural Network (CNN) [11] and the attention network [12] to learn the news representation model, and uses Gate Recurrent Unit

(GRU) [13] to learn the short-term preferences of users from the news reading sequence of users in the recent period, and integrates the long-term and short-term preferences to form a complete user representation. Therefore, this paper proposes a news recommendation model based on encoder graph neural network and Bat optimization. To a certain extent, the imbalance between the head and the tail can be alleviated, and the unpopular news can be accurately recommended to the required users as far as possible. Our main contributions are as follows:

1. Firstly, Bat optimization algorithm is used to improve the effect of news clustering.
2. Secondly, the concept of metadata is introduced into the graph neural network, and the ontology of learning resources based on knowledge points is established to realize the correlation between news resources.
3. Finally, the model combining Convolutional Neural Network (CNN) and attention network is used to learn the representation of news, and Gate Recurrent Unit (GRU) is used to learn the short-term preferences of users from their recent reading history.

2. Related Works

By analyzing and processing the original news data and user behavior data, the personalized news recommendation system combines different news recommendation methods to model news and users, fully extracts the features of news content, and mining user preferences to generate news and user embedded representation. Its architecture is shown in Figure 1. When the user enters the personalized news recommendation system, the recommendation engine will select the news that meets the user's needs and preferences from the candidate news set according to the user's reading history, location, preference and other factors, and sort the candidate news according to the prediction model to generate a recommendation list and display it to the user. For example, if the user has read football news before, the recommendation system may recommend the latest World Cup news to the user. If the user is located in Beijing, then the recommendation system may recommend news about the Beijing area. In addition, the user interface will display news on different topics for each user, collect user feedback and update the recommendations accordingly, thus enabling personalized news recommendations.

Although PNR technology has made remarkable progress, it still needs to further improve the level of personalized recommendation, including more comprehensive mining of news semantics, more granular extraction of user preferences and the construction of more efficient personalized news recommendation models. With the vigorous development of mobile Internet technology, personalized news recommendation based on mobile terminal has become the mainstream trend. Personalized mobile news recommendation system can provide news information to users anytime and anywhere, with good interactivity, and bring more convenient and comfortable experience for users to obtain news information in real time. However, the small screen size of mobile devices, the unstable network quality and the variability of usage scenarios may affect the effect and efficiency of personalized news recommendation system, which is still an important issue to be solved in the future personalized mobile news recommendation research.

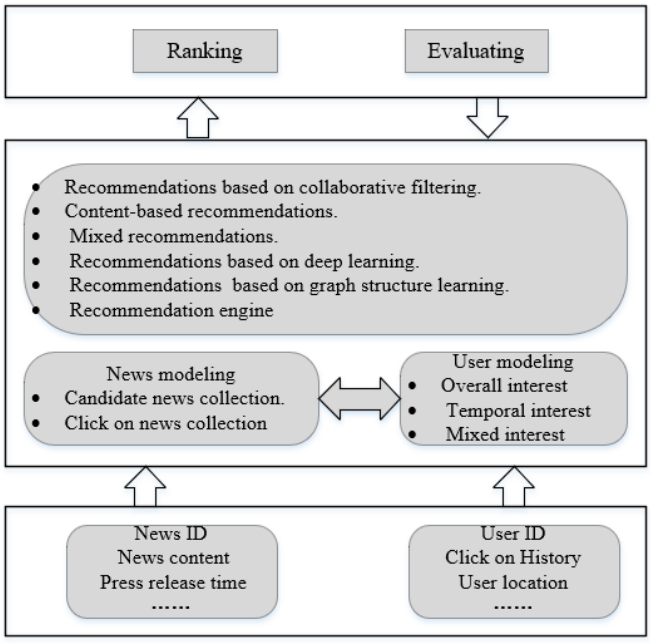


Fig. 1. Personalized news recommendation system architecture

2.1. Traditional News Recommendation Methods

Early news recommendation usually relied on the correlation between news and semantic correlation [14]. Both of these methods are to model news and calculate the similarity between news. However, neither of these two methods can accurately describe the user's interest degree, they only represent the similarity between news and lack the expression of user's interest degree. Therefore, how to mine and express users' interest in news reading in personalized news recommendation task is a very important issue.

Sequential recommendation attempts to predict a users next behavior by exploiting their historical behavior-sequences [15], which has been widely adopted in modern online information systems, such as news, video, advertisements, etc. Differ from traditional recommendation tasks that model user preferences in a static fashion, sequential recommendation is capable of capturing users evolved and dynamic preferences. For example, a user may prefer to watch soccer news only during the period of World Cup, which can be regarded as a kind of short-term preference.

Collaborative filtering (CF) [16] calculates the behavior similarity between the target user and the neighbor, predicts the target user's possible score on a certain item based on the nearest neighbor, and finally pushes the Top-N score to the target user. However, because the news recommendation platform produces a large number of new news every day, the news recommendation system has a serious cold start problem. However, the collaborative filtering method is difficult to effectively solve the cold start problem of the system, so it is not suitable for news recommendation. For this reason, many news recommendations are based on news content, such as Koning et al. [17] combine the

semantics of TFIDF and domain ontology to make recommendations. Yang et al. [18] introduced a method of using word embedding to build user interest model and realize personalized news recommendation. Both of these methods only encode news from a single perspective and cannot fully and accurately express the semantic information of news.

2.2. News Recommendation Based on Deep Learning Methods

In 2017, Yahoo News used a deep learning network as a feature extractor to study news content, so as to find out users' preferences, which effectively improved the accuracy of the recommendation system compared with traditional methods [19]. However, this method only vectored news from the perspective of news words. Huang et al. [20] proposed a new deep hybrid recommendation model (DMFL), it vectored news from multiple perspectives, and finally summed and averaged the vectors from each perspective to obtain the feature vectors of users, which could not accurately model users.

In order to realize news recommendation, Google proposed Wide and Deep model [21], which could train both linear model and deep network, but had low learning efficiency in real scenes. Tiwari et al. [22] proposed personalized news recommendation for users based on their location and interest. This method mainly vectored users and could not accurately consider news semantic content. Liang et al. [23] proposed that knowledge perception neural network learned news representations from news article titles, and then learned user representations from news representations according to the similarities between candidate news and each browsed news. This method only quantified news from the perspective of news headlines, and could not accurately model news content.

Cold-start Recommendation. Making recommendations for new users or items with limited interaction samples is a challenging problem in recommender systems, also named cold-start problem. According to the number of interaction samples, the cold-start problem can be divided into two phases: cold-start phase (zero sample) and warm-up phase (a few samples).

For the cold-start phase, it is common to use auxiliary information, e.g., user attributes, item attributes, knowledge graph, samples in auxiliary domain.

The warm-up phase is a dynamic process that gradually improves the recommendation performance with the increase of samples. DropoutNet, MetaEmb and MeLU are able to solve the cold-start problem in warm-up phase, and they have been introduced above. MetaHIN and MAMO share similar idea with MeLU. These methods can be categorized into three groups: (1) MeLU, MAMO, MetaHIN try to personalize the parameters of the deep model. (2) MetaEmb exploits a good pre-trained embedding. (3) DropoutNet learns a more robust item embedding. Note that the proposed MWUF is completely different from these methods, and does not belong to the three groups. MWUF learns to warm up cold items by using meta networks to predict scaling and shifting functions which can transform the cold ID embeddings to fit the model better.

Multimodal-based news recommendation. Most of the existing news representation methods usually only learn the news representation from the news text, and ignore the visual information in the news (such as pictures, animations, etc.). In fact, users click on news not only because they are interested in the headlines, but also because they are likely to be attracted by multimodal features such as images, audio, and video. Therefore,

integrating visual and text information to learn multimodal features is particularly important for news modeling and predicting news click rates. Guo et al. [24] proposed a news recommendation method based on deep reinforcement learning by integrating multimodal features to learn news representations and representing users' interests to multi-modal information. Wu et al. [25] adopted a pre-trained visual language model to encode news text and the region of interest images extracted from news images, and proposed a multimodal news recommendation method. Xun et al. [26] adopted the method of visual semantic modeling to capture the visual impression information perceived by users when browsing news, so as to further understand the process of users reading news. Experiments showed that news recommendation with multi-modal features could describe news content more comprehensively and improve the effect and accuracy of news recommendation.

Reference [27] proposed a method to predict news click rate through deep knowledge perception network and improved news representation learning by using knowledge graph. Reference [28] proposed a deep reinforcement learning framework, which could dynamically complete the modeling of news and users, and improve the diversity of news recommendation while ensuring the accuracy of recommendation. Reference [29] proposed a deep neural network model combined with attention mechanism to complete the task of news recommendation. The model included three extractors, which learned news representation, sequence information features and user interests respectively, effectively improving the accuracy of news matching. References [30,31] respectively proposed a neural network news recommendation model based on personalized attention mechanism and a neural network news recommendation method based on attention multiple views. The former focused on mining the impact of news on users at the word and document levels by applying attention mechanism, while the latter focused on mining different types of news information by integrating multi-view news recommendation methods for better news recommendation. It can be seen that most of the mainstream methods in the field of news recommendation are based on neural network and deep learning related technologies. In the future, neural network and deep learning will be important directions in this field.

3. Proposed News Recommendation Model

Firstly, Bat optimization algorithm is used to improve the effect of news clustering. Secondly, the concept of metadata is introduced into the graph neural network, and the ontology of learning resources based on knowledge points is established to realize the correlation between news resources. Finally, the model combining Convolutional Neural Network (CNN) and attention network is used to learn the representation of news, and Gate Recurrent Unit (GRU) is used to learn the short-term preferences of users from their recent reading history. The proposed scheme process is shown in figure 2.

3.1. Bat Optimization-based News Clustering

In the process of data modeling, project type characteristics are set in this paper, which can accurately feedback news preferences and provide accurate recommendation content. For personalized recommendation system, there are many types of characteristics in each project. This results in fewer project types than project numbers. In addition, different

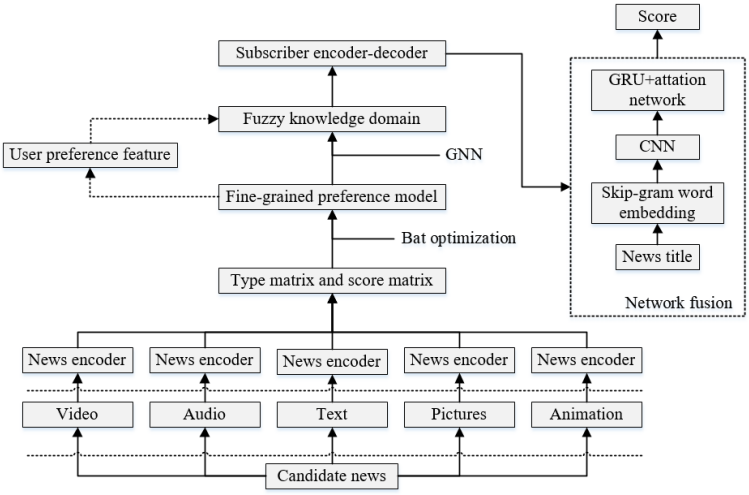


Fig. 2. Proposed news recommendation model framework

users will have different news interests and preferences for each project type. According to the above characteristics, this paper chooses the method combining item-scoring matrix and item-type matrix to design the ration-proportion-item-preference (RPIP) algorithm. Based on this, the user's preferences for various items are predicted, and the fine-grained user-project type preference matrix is constructed.

Rating proportion (RP) represents the proportion of the total rating R_u of a user u to the total rating of an item of type $R_{u,i}$. The calculation process of the score ratio is shown in equation (1).

$$Q_{u,i} = \frac{\sum_{i \in I_{u,e}} R_{u,i}}{R_u}. \tag{1}$$

Where $Q_{u,i}$ is the score ratio. i is a self-scoring variable.

Item preference (IP) represents the proportion of the item number $C(I)$ of type e in the total number $C(I_e)$, which can effectively avoid the difference in user preference caused by popular items of type e. The matrix $M_{i,e}$ is compiled by calculating the items in equation (2). To avoid the problem that the number of items of type e is 0, this article uses $C(I_e) + 1$ as the denominator.

$$M_{i,e} = ln \frac{C(I)}{C(I_e) + 1}. \tag{2}$$

User's preference for project type P_u is:

$$P_u = Q_{u,i} \times M_{i,e}. \tag{3}$$

Formula (3) is used to calculate the degree of user preference for project type, and a fine-grained user-project type preference matrix is constructed. In this paper, the real

evaluation data are selected in the process of analysis and calculation. For the case that the evaluation data is empty, part of the data is represented as 0.

The processing process of bat-optimized user fuzzy clustering algorithm is as follows: Firstly, the optimal initial clustering center is determined by bat-optimized algorithm; Then, fuzzy C-means (FCM) clustering is implemented for users. When bat optimization algorithm [32] is used, each bat is represented as a cluster center matrix . Bat optimization clustering is shown in **Algorithm 1**.

Algorithm 1 Bat optimization clustering

1: **Input:** Fine-grained user-item type preference matrix, bat population size h, cluster number, maximum number of iterations T.
2: **Output:** Membership matrix of user cluster containing W and c user clusters.
3: The randomly generated initial bat population C, and the initialized bat individual C_i, velocity V_i, position x_i, pulse emissivity r_i, loudness A_i, and frequency f_i.
4: Calculating membership.
5: The individual fitness values of all bats in the population are calculated and sorted, and the bat individuals with the best fitness values are selected.
6: The individual position and speed parameters of bats are modified.
7: Generating a random number r_0 and traversing individual bats within the population. When the conditions are met $r_0 < r_i$ to generate x_{new} and the fitness F_{fit} is calculated.
8: The random number r_1 is generated and the individual bats in the population are traversed. If condition $r_1 < A_i$ is true, receive x_{new} of step 5 and modify loudness and pulse emissivity.
9: When the number of iterations is less than T or the convergence condition of cluster center is not satisfied, the iteration is carried out again in step 2. On the contrary, bat individual x_{best} under the condition of optimal fitness value is output, and it is regarded as the optimal initial cluster center and cluster division is implemented to generate user cluster membership matrix W and c user clusters.

3.2. News Resource Ontology Construction Based on Improved Graph Neural Network

At present, most online education platforms provide users with services that integrate high-quality online courses and other learning resources. However, in the face of massive learning resources, it is difficult for users to quickly find learning materials suitable for themselves, and even problems such as ”cognitive load” and ”information trek” may occur. Therefore, it is an urgent need of online education platform to intelligently analyze the rules of learning behavior and cognitive characteristics of users according to their learning process, and to provide personalized learning resource recommendation.

Information management. The management module of the NEWS ontology provides a vocabulary which allows the definition of metadata regarding the news item life-cycle management. Aspects like authorship information, rights, etc. are covered by this module. An application of such information is, for example, finding adequate journalists to cover a certain event (e.g. a soccer match, a political meeting, etc.). This is simply implemented by searching the NEWS system datawarehouse to find the authors of former news items reporting on similar events. Another example: management metadata could be used to

define the urgency of a certain news item. This is helpful in deciding the order in which news items are sent to clients or are processed by certain procedures in the agency's production workflow (automatic annotation, for instance).

In order to realize effective recommendation of news resources, it is far from enough to consider users' preferences only. In the current educational environment, learning resources show an obvious trend of complexity and diversification both in structure and form, among which the forms of expression can be divided into courseware, cases, documents, indexes, online courses, test questions, test papers, assignments, texts, etc. In this context, the ontology construction based on knowledge points is realized by improving the graph neural network. The basic principle of this process is to fully reflect the data sharing value of news resources at all levels and types. The disordered and non-structured news resources are divided into text data and media data. Among them, the text materials include courseware, cases, test papers, workbooks of auxiliary materials, etc. Media materials include video, animation, audio, text, pictures, etc. For some resources whose classification is difficult to be defined directly, the improved graph neural network is used to restructure them, and the ontology representation construction method based on metadata is adopted to solve the fuzzy problem of knowledge domain definition in graph neural network.

To this end, the news resource model is first constructed, which is as follows:

$$\gamma = \frac{sim(x_1, x_2, \cdots, x_i)}{p}. \tag{4}$$

Where γ is the learning resource model. x_i is composed of knowledge points. $sim(x_1, x_2, \cdots, x_i)$ is the degree of similarity between the knowledge points. p indicates the proportion of metadata in the resource. This model is mainly used for the management and retrieval of knowledge points in learning resources. When the knowledge points are used to describe the resource ontology, the correlation between the ontology and knowledge points is established through metadata. Using metadata attributes to define the hierarchy of different categories, a graph neural network including parent classes (data, animation, course, etc.) and sub-classes (text, graphics, video, image, audio, etc.) is constructed. At this time, the hierarchical relationship of news resources can reflect the difference in the composition of knowledge points. In order to associate different resources, the relationship between knowledge points in the resource structure of improved graph neural network is designed into four types: parent-child, reference, dependence and parallel. Among them, the father-son relationship mainly reflects the part and integrity of resources, the dependence relationship mainly reflects the overlap of resources, and the parallel relationship mainly reflects the independence of knowledge points between resources.

3.3. Title Encoder

The title encoder vectorizes news headlines. The title encoder includes a preprocessing layer, a word vector training layer, a CNN layer and an attention layer.

The preprocessing layer performs word segmentation of the news titles, and the Jieba word segmentation tool is used to perform preprocessing operations such as word segmentation and deactivation of the titles. The word after the news title participle is expressed as $[s_1^t, s_2^t, \cdots, s_M^t]$, and M is the number of title words.

The CNN layer extracts the features between words by passing the word vectors of the training layer through the CNN network. The width of the convolution kernel is consistent with the word embedding dimension, and the number of window words for each convolution operation is denoted as h, that is, the convolution kernel $\varpi \in R^{h\times d}$. The convolution result of sliding the i-word window of the news title is:

$$N_i^t = ReLU(\varpi w_{i:i+h-1}) + b. \tag{5}$$

Where $ReLU$ is a nonlinear activation function. $w_{i:i+h-1}$ is the number of words taken for each convolution operation. $b \in R$ is offset. The title words are expressed as: $[N_1^t, N_2^t, \cdots, N_P^t]$.

The attention layer uses the attention mechanism to assign weight to each word. α_i^t represents the attention weight of the i-th word in the news title, and its calculation formula is:

$$a_i^t = \frac{q_t^T \tanh(V_t \times c_i^t + v_t)}{||q_t^T||_{L2}||q_t^T \tanh(V_t \times c_i^t + v_t)||_{L2}}. \tag{6}$$

$$\alpha_i^t = \frac{exp(a_i^t)}{\sum_{j=1}^{M} exp(a_i^t)}. \tag{7}$$

Where V_t represents the weight, v_t represents the offset item, c_i^t represents the vector of the $i-th$ word in the news title, and q_t represents the query vector. The news title vector is represented as:

$$r^t = \sum_{i=1}^{M} \alpha_i^t c_i^t. \tag{8}$$

3.4. Text Encoder

Text encoder vectorizes the news text. The text encoder includes a preprocessing layer, a word vector training layer, a CNN layer and an attention layer.

The preprocessing layer divides the news text into words, the Jieba word segmentation tool is used to divide the text into words, the words of the news text are represented as $[s_1^b, s_2^b, \cdots, s_P^b]$, P is the number of words in the text. The word vector training layer is to obtain the word vector of the news text from the word segmentation of the preprocessing layer through the BERT model, and the word vector of the text is represented as $[c_1^b, c_2^b, \cdots, c_P^b]$.

The CNN layer extracts the features between words by passing the word vectors of the training layer through the CNN network. The width of the convolution kernel is consistent with the word embedding dimension, and the number of window words for each convolution operation is denoted as h, that is, the convolution kernel $\varpi \in R^{h\times d}$. The convolution result of sliding the i-word window of the news text is:

$$N_i^t = ReLU(\varpi w_{i:i+h-1}) + b. \tag{9}$$

Where $ReLU$ is a nonlinear activation function. $w_{i:i+h-1}$ is the number of words taken for each convolution operation. $b \in R$ is offset. The title words are expressed as: $[N_1^t, N_2^t, \cdots, N_P^t]$.

The attention layer uses the attention mechanism to assign weight to each word. α_i^{tb} represents the attention weight of the $i-th$ word in the news title, and its calculation formula is:

$$a_i^b = \frac{q_b^T \tanh(V_b \times c_i^b + v_b)}{||q_b^T||_{L2}||q_b^T \tanh(V_b \times c_i^b + v_b)||_{L2}}. \tag{10}$$

$$\alpha_i^b = \frac{exp(a_i^b)}{\sum_{j=1}^{M} exp(a_i^b)}. \tag{11}$$

Where V_b represents the weight, v_b represents the offset item, c_i^b represents the vector of the $i-th$ word in the news title, and q_b represents the query vector. The news title vector is represented as:

$$r^b = \sum_{i=1}^{P} \alpha_i^b c_i^b. \tag{12}$$

3.5. Decoder

Since RNN has ability to learn current word from the previous words in the sequence and it can consider the encoder's representation and its internal state at the same time, we use RNN as the decoder. We denote x_d^i as the word embedding of the $i-th$ word in the scientific papers and apply gated recurrent unit (GRU) to solve gradient exploding or vanishing problem. The attention mechanism leans a weighted interpolation c_i based on the encoder's representation, i.e.,

$$c_i = \sum_j \alpha_{ij} s_j. \tag{13}$$

where α_{ij} is the output value derived from the softmax function and the $i-th$ word must be aligned with the $j-th$ output.

3.6. User Long-term Preference Fusion

The short-term interests of users are dynamic and changeable, and we usually learn users' short-term preferences from their recent news reading history. To capture sequential information in a user's reading history, we use a GRU network to learn a user's short-term interest representation. The news representation vector obtained by our news modeling is d, and the representation vector of the user's short-term reading history sequence can be expressed as $d_1, d_2, \cdots, d_N$. N is the length of the short-term sequence, then the GRU network calculates the expression of the user's short-term interest as shown in equations (14-18).

$$r_i = sigmoid(W_r[h_{t-1}, d_t]). \tag{14}$$

$$z_i = sigmoid(W_z[h_{t-1}, d_t]). \tag{15}$$

$$\tilde{h}_t = \tanh(W_{\tilde{h}}[r_1 \circ h_{t-1}, d_t]). \tag{16}$$

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \tilde{h}_t. \tag{17}$$

$$u_{short} = h_N. \tag{18}$$

Where $\circ$ indicates the hadamard product. W_r, W_z, and $W_{\tilde{h}}$ are parameters to be learned in the GRU network. The hidden state of the last output is recorded as the user's short-term interest indicating u_{short}.

3.7. Model Training

For the prediction of ratings between users and news, we use a simple dot product to efficiently calculate the matching score between users and news, i.e. $score = u^T d$.

We use negative sampling technique to train the model. For each positive sample, that is, the news clicked by the user, we randomly select K news that has not been clicked as negative samples, so as to transform the original complex prediction problem into $K + 1$ binary classification problem, so as to simplify the training process and improve the computational efficiency. Under this condition, the posterior probability of each positive sample being clicked can be expressed as equation (19).

$$positive = \frac{exp(u^T d_i^P)}{exp(u^T d_i^P) + \sum_{j=1}^{K} exp(u^T d_{i,j}^N)}. \tag{19}$$

Where P represents the number of positive samples. d_i^P represents the $i-th$ positive sample. $d_{i,j}^N$ represents the $j-th$ negative sample of the $i-th$ positive sample, and the sum of the negative likelihood logarithm of the click-through rate of all positive samples is taken as the loss function of the model, namely,

$$Loss = \sum_{i=1}^{P} \log(\frac{exp(u^T d_i^P)}{exp(u^T d_i^P) + \sum_{j=1}^{K} exp(u^T d_{i,j}^N)}). \tag{20}$$

4. Experiment Results and Analysis

4.1. Datasets

This paper conducts experiments on the Chinese data set Baidu news, the English dataset MIND (Microsoft news data set) Adressa data set and Digg data set.

Baidu news is to crawl the browsing history of some users in September 2018 on Baidu News website, including users, news titles, news text, and browsing time stamps. We build the test set using data from September 25 to September 30, and the training set using data from September 1 to September 24.

MIND [33] is a large scale news recommendation English data set collected by Microsoft News from its anonymous user logs from October 12 to November 22, 2019, which contains users, news headlines, news text, news summaries, news categories, and news entities. This article builds a test set using data from November 15 to November 22, and a training set using data from October 12 to November 14.

Adressa data set [34] consists of newslogs collected from the Adresseavisen website over a period of 3 months, in both full and small datasets. The full version includes 3,083,438 users, 48,486 articles and 27,223,576 clicks in 10 weeks. The minor version has 561,733 users, 11,207 articles, and 2,286,835 clicks in one week.

The Digg dataset [35] consists of 3553 news items collected by the Information Science Institute of the University of Southern California on the Digg website in June 2009, including the digg_votes table and the digg_friends table. The digg_votes table contains 139,409 users and 3,018,197 votes; The digg_friends table contains 1731,658 link relationships between 71367 users.

4.2. Evaluation Index

In this paper, T is set as the test set to recommend news for user u. F_u is the recommended result. Suppose that the news set that user u is interested in in the test data set is denoted as T_u, and the performance indicators in this paper take Precision (P), Recall (R) and F1 index as references.

Precision refers to the proportion of content in the news list recommended by the news recommendation system that conforms to user preferences. The higher the accuracy rate, the more accurate the news recommendation is. The calculation formula is as follows:

$$P = \frac{\sum_u |F_u \cap T_u|}{\sum_u |F_u|}. \tag{21}$$

Recall refers to the proportion of news recommended by the recommendation system in the process of users interacting with many news, that is, the share of news recommended by the system in the news viewed by users, and its calculation formula is as follows.

$$R = \frac{\sum_u |F_u \cap T_u|}{\sum_u |T_u|}. \tag{22}$$

The measurement standard of the recommendation system is that the higher the accuracy rate and recall rate, the better. However, because these two indicators are contradictory to some extent, F1 is the evaluation index that combines the two. Therefore, this paper uses F1 values to measure the performance of the recommendation system. The higher the F1 value, the better the comprehensive performance. The calculation formula is as follows:

$$F1 = \frac{P + R}{PR}. \tag{23}$$

4.3. Influence of News Recommendation List N on P, R and F1 Values

The test set news count for the Chinese data set is 3654 click-throughs from 853 users, the test set for the English data set is 6325 click-throughs from 920056 users, test set of

Adressa is 4892 click-throughs from 10057 users and the test set of Digg is 7884 click-throughs from 50478 users.

The different length of the recommendation list will directly affect the P, R, and F1 values of the result. In order to verify the effectiveness of Attention-BTE under different length of recommendation list, in the test set, the range of recommendation list N is taken as 10,20,30,40,50 in this experiment. However, the value of N cannot be infinite, because it cannot be separated from the limit of the test set. Figures 3-6 (tables 1-4) respectively show the comparison of P, R and F1 values of Chinese, English, Adressa, Digg data sets under different recommendation list lengths N.

Table 1. Comparison of the number of different news recommendations in Chinese data set

Number of news recommendations	P	R	F1
10	0.38	0.40	0.20
15	0.40	0.40	0.21
20	0.42	0.40	0.22
25	0.51	0.56	0.28
30	0.60	0.72	0.34
35	0.56	0.60	0.29
40	0.46	0.48	0.24
45	0.43	0.44	0.22
50	0.40	0.39	0.19

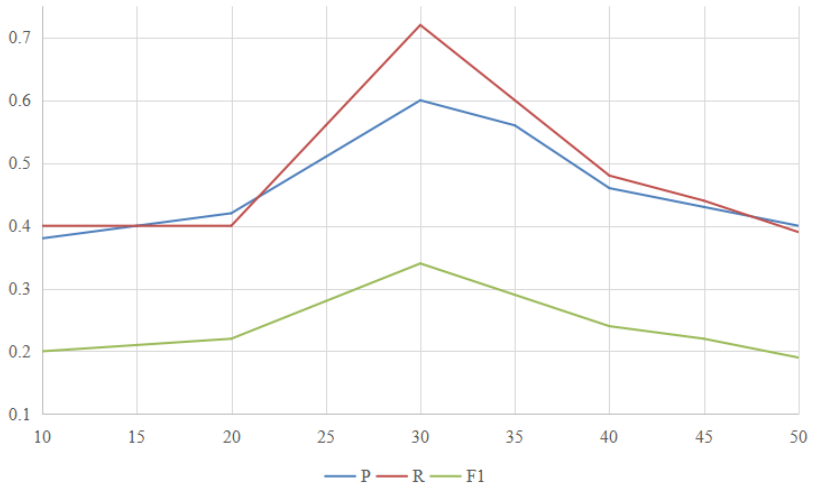


Fig. 3. Comparison of the number of different news recommendations in Chinese data sets

Table 2. Comparison of the number of different news recommendations in English data set

Number of news recommendations	P	R	F1
10	0.50	0.45	0.24
15	0.55	0.57	0.28
20	0.59	0.69	0.32
25	0.67	0.77	0.36
30	0.75	0.84	0.40
35	0.66	0.77	0.36
40	0.63	0.70	0.33
45	0.61	0.67	0.32
50	0.59	0.64	0.31

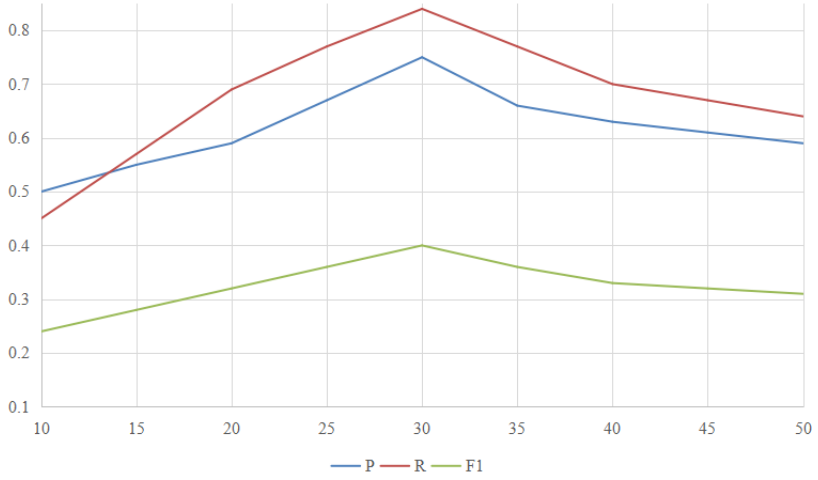


Fig. 4. Comparison of the number of different news recommendations in English data sets

Table 3. Comparison of the number of different news recommendations in Adressa data set

Number of news recommendations	P	R	F1
10	0.61	0.56	0.35
15	0.66	0.68	0.39
20	0.70	0.80	0.43
25	0.78	0.88	0.47
30	0.86	0.95	0.51
35	0.77	0.88	0.47
40	0.74	0.81	0.44
45	0.72	0.78	0.43
50	0.70	0.75	0.42

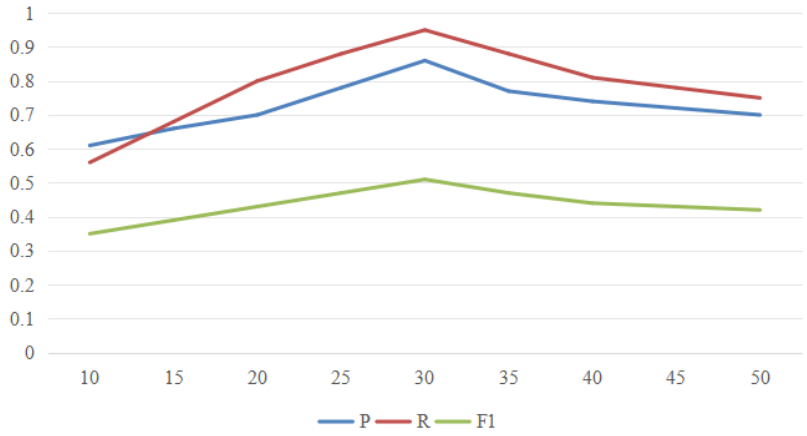


Fig. 5. Comparison of the number of different news recommendations in Adressa data sets

Table 4. Comparison of the number of different news recommendations in Digg data set

Number of news recommendations	P	R	F1
10	0.49	0.51	0.31
15	0.51	0.51	0.32
20	0.53	0.51	0.33
25	0.62	0.67	0.39
30	0.71	0.83	0.45
35	0.67	0.71	0.40
40	0.57	0.59	0.35
45	0.54	0.55	0.33
50	0.51	0.50	0.30

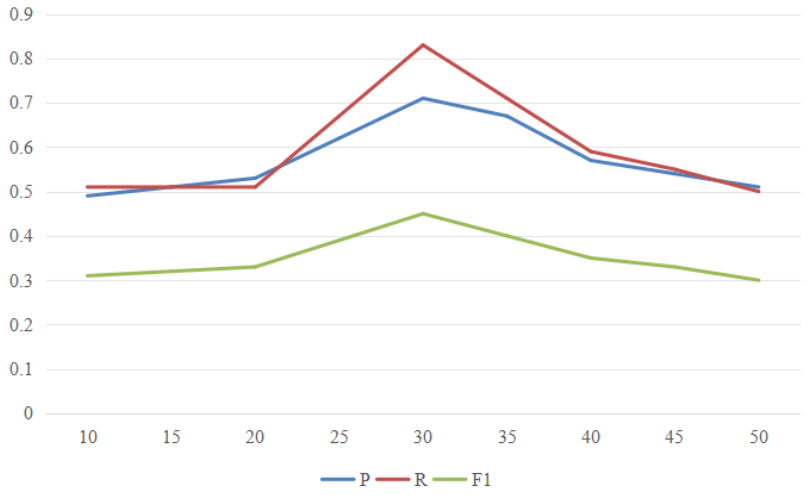


Fig. 6. Comparison of the number of different news recommendations in Digg data sets

It can be concluded from Figures 3-6 that when 30 news articles are recommended to users, P, R and F1 can achieve the maximum value, and the news recommendation effect is the best. As the list of recommendations grows, P, R, and F1 gradually increase. This is because the length of the recommendation list increases to include more items that interact with the user. However, the values of P, R and F1 show a decreasing trend when the number of recommendations is 40 and 50, which is due to the over-fitting phenomenon caused by the increase in the number of recommendations. That is, too small a number of recommendations cannot increase the number of projects that interact with users, and in the case of a small number of test sets, too large a number of recommendations is susceptible to the impact of quantitative over-fitting.

4.4. Comparison With Other State-of-the-art Methods

When the number of recommendation lists is the same, this new method is compared with news recommendation algorithms including KEHB [36], Feedrec [37], MACR [38], GACF [39]. The recommended precision, recall rate and F1 values in Chinese, English, Adressa, Digg data sets are shown in Tables 5-8 respectively. Meanwhile, pictorial diagrams of them are shown in figures 7-10. This article has the following observations.

Table 5. Comparison results of different models on Chinese data sets

Method	P	R	F1
KEHB	0.48	0.67	0.28
Feedrec	0.54	0.68	0.30
MACR	0.56	0.68	0.31
GACF	0.58	0.69	0.32
Proposed	0.61	0.72	0.33

Table 6. Comparison results of different models on English data sets

Method	P	R	F1
KEHB	0.62	0.75	0.29
Feedrec	0.62	0.76	0.29
MACR	0.69	0.80	0.32
GACF	0.72	0.82	0.33
Proposed	0.74	0.84	0.36

First, since the number of users and the number of news in English data set is larger than that in Chinese data set, the accuracy of Feedrec in Chinese data set is lower than that of KEHB, while the opposite is true in English data set. This may be due to the small number of users and news in the Chinese data set, and the serious problem of sparse user-item scoring matrix encountered in Feedrec, so the size of the data set is also the key to affecting the accuracy of news recommendation.

Table 7. Comparison results of different models on Adressa data sets

Method	P	R	F1
KEHB	0.51	0.64	0.24
Feedrec	0.51	0.65	0.23
MACR	0.58	0.79	0.31
GACF	0.69	0.81	0.32
Proposed	0.75	0.82	0.35

Table 8. Comparison results of different models on Digg data sets

Method	P	R	F1
KEHB	0.59	0.78	0.39
Feedrec	0.65	0.79	0.41
MACR	0.67	0.79	0.42
GACF	0.69	0.80	0.43
Proposed	0.72	0.83	0.44

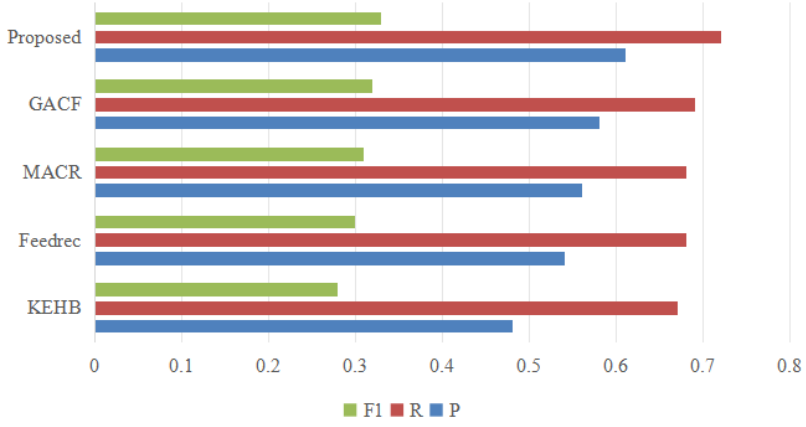


Fig. 7. Visualization display of table 5

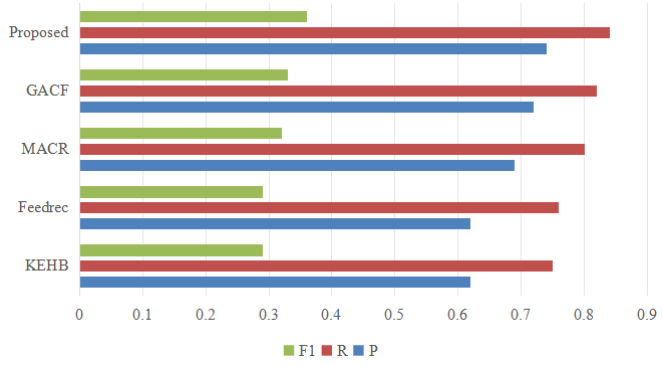


Fig. 8. Visualization display of table 6

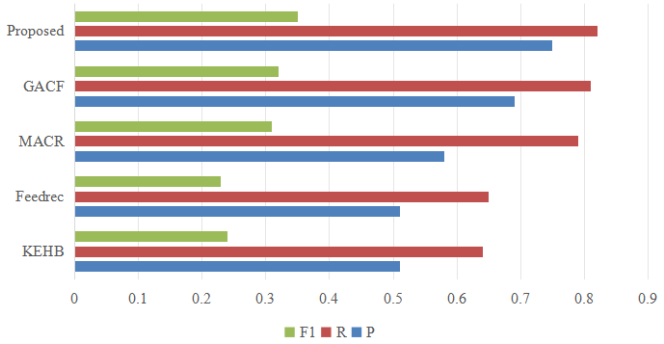


Fig. 9. Visualization display of table 7

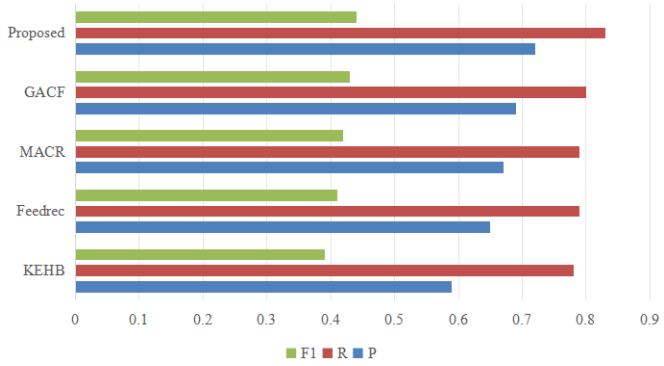


Fig. 10. Visualization display of table 8

Second, the method that combines headlines, body and categories (e.g. MACR) is superior to the method that only uses news headlines (e.g. GACF), which may be that rich content information helps to learn more accurate news presentation and proves the effectiveness of integrating multiple news information.

Third, the model in this paper is superior to KEHB, Feedrec, MACR and GACF, because the model makes full use of the theme, text and event of the news, and takes the event as a perspective of news information to form an informative news representation, and the F1 value is increased by about 6%.

4.5. Ablation Experiment

In this section, we conduct ablation experiments to show the effectiveness of proposed method. The results are shown in tables 9-12. From the results, we can see that the proposed method has the better effect.

Table 9. Ablation comparison results on Chinese data sets

Method	P	R	F1
CNN	0.49	0.66	0.29
CNN+GNN	0.57	0.68	0.31
CNN+GRU	0.58	0.69	0.32
Proposed	0.61	0.72	0.33

Table 10. Ablation comparison results on English data sets

Method	P	R	F1
CNN	0.66	0.78	0.30
CNN+GNN	0.68	0.79	0.31
CNN+GRU	0.71	0.82	0.34
Proposed	0.74	0.84	0.36

Table 11. Ablation comparison results on Adressa data sets

Method	P	R	F1
CNN	0.67	0.74	0.29
CNN+GNN	0.68	0.75	0.31
CNN+GRU	0.72	0.79	0.34
Proposed	0.75	0.82	0.35

Table 12. Ablation comparison results on Digg data sets

Method	P	R	F1
CNN	0.66	0.79	0.39
CNN+GNN	0.68	0.80	0.42
CNN+GRU	0.71	0.81	0.42
Proposed	0.72	0.83	0.44

5. Conclusions

Research on news recommendation has important practical significance and application needs. The accuracy of news recommendation lies in whether the content information of news is fully considered. This paper proposes a novel news recommendation model based on encoder graph neural network and Bat optimization in online social networks, news headlines, text and events are integrated through attention mechanism to get news vectorization. First of all, because words are the most basic unit of news headlines, text and events, and the same word has different meanings in different news, starting with the most basic words, each word is assigned different weights, which can fully consider news semantic information. Secondly, it can accurately vectorize users. Since each user is interested in news from different angles, the attention mechanism can accurately describe the user's perspective of interest in news, so as to accurately vectorize the user. Finally, the attention mechanism is used to assign weights to the headlines, text and events of the news, and the ratio of each Angle to the vectorization accuracy of the news is fully considered. The validity of this method is verified by experiments with real experimental data.

In exploring the diversity of news recommendation, the existing research usually only focuses on the diversity of recommended content, often ignoring the balance between diversity and accuracy, which is likely to lead to a large difference between the recommended news and the actual preferences of users. Thus, the user's satisfaction with the recommendation system is reduced, and even the user information is lost. Therefore, it is of great significance to design a more comprehensive diversified news recommendation model for improving the quality of online news service. In the follow-up research, the vectorization of events can be calculated using the advanced neural network method to further improve the accuracy of news recommendation.

Acknowledgments. The author is grateful for the anonymous review by the review experts.

References

1. Wu Y. The current situation and innovation of news communication under the environment of financial media[J]. Journal of Environmental and Public Health, 2022, 2022.
2. Li J. Research on readers' Reading Promotion in the digital age[C]//2022 International Conference on Artificial Intelligence in Everything (AIE). IEEE, 2022: 511-516.
3. Lian J. Research on Tourism Marketing Strategy Innovation in New Media Era[J]. Frontiers in Business, Economics and Management, 2023, 8(1): 103-106.
4. Pido N W T, Sujitno F D C. The effect of digital technology on students' reading Behavior[J]. Journal of English Teaching and Linguistic Issues (JETLI), 2022, 1(2): 81-93.

5. Jisi A, Yin S. A new feature fusion network for student behavior recognition in education[J]. Journal of Applied Science and Engineering, 2021, 24(2): 133-140.
6. Zhang P, Zheng J, Lin H, et al. Vehicle Trajectory Data Mining for Artificial Intelligence and Real-Time Traffic Information Extraction[J]. IEEE Transactions on Intelligent Transportation Systems, 2023.
7. Bhadrashetty A, Patil S. Movie Success and Rating Prediction Using Data Mining[J]. Journal of Scientific Research and Technology, 2024: 1-4.
8. Le H, Maragh R, Ekdale B, et al. Measuring political personalization of Google news search[C]//The World Wide Web Conference. 2019: 2957-2963.
9. Ko H, Lee S, Park Y, et al. A survey of recommendation systems: recommendation models, techniques, and application fields[J]. Electronics, 2022, 11(1): 141.
10. Das D, Sahoo L, Datta S. A survey on recommendation system[J]. International Journal of Computer Applications, 2017, 160(7).
11. Yin S, Wang L, Shafiq M, et al. G2Grad-CAMRL: An Object Detection and Interpretation Model Based on Gradient-weighted Class Activation Mapping and Reinforcement Learning in Remote Sensing Images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 3583-3598, 2023
12. Niu R, Sun X, Tian Y, et al. Hybrid multiple attention network for semantic segmentation in aerial images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-18.
13. Ren L, Cheng X, Wang X, et al. Multi-scale dense gate recurrent unit networks for bearing remaining useful life prediction[J]. Future generation computer systems, 2019, 94: 601-609.
14. Zhu L, Song J, Zhu X, et al. Adversarial learning-based semantic correlation representation for cross-modal retrieval[J]. IEEE MultiMedia, 2020, 27(4): 79-90.
15. Chang J, Gao C, Zheng Y, et al. Sequential recommendation with graph neural networks[C]//Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. 2021: 378-387.
16. Fkih F. Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental comparison[J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(9): 7645-7669.
17. de Koning E, Hogenboom F, Frasincar F. News recommendation with CF-IDF+[C]//Advanced Information Systems Engineering: 30th International Conference, CAiSE 2018, Tallinn, Estonia, June 11-15, 2018, Proceedings 30. Springer International Publishing, 2018: 170-184.
18. Yang Z, Cui L, Wang X, et al. MIAR: Interest-Activated News Recommendation by Fusing Multichannel Information[J]. IEEE Transactions on Computational Social Systems, 2022.
19. Liu L, Yan R J, Maruvanchery V, et al. Transfer learning on convolutional activation feature as applied to a building quality assessment robot[J]. International Journal of Advanced Robotic Systems, 2017, 14(3): 1729881417712620.
20. Huang Z, Yu C, Ni J, et al. An efficient hybrid recommendation model with deep neural networks[J]. IEEE Access, 2019, 7: 137900-137912.
21. Cheng H T, Koc L, Harmsen J, et al. Wide & deep learning for recommender systems[C]//Proceedings of the 1st workshop on deep learning for recommender systems. 2016: 7-10.
22. Tiwari S, Pangtey M S, Kumar S. Location aware personalized news recommender system based on Twitter popularity[C]//Computational Science and Its ApplicationsCICCSA 2018: 18th International Conference, Melbourne, VIC, Australia, July 2C5, 2018, Proceedings, Part IV 18. Springer International Publishing, 2018: 650-658.
23. Liang Y, Meng F, Zhang Y, et al. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(15): 13343-13352.
24. Guo N, Fu Z, Zhao Q. Multimodal News Recommendation Based on Deep Reinforcement Learning[C]//2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP). IEEE, 2022: 279-284.

25. Wu C, Wu F, Qi T, et al. Mm-rec: Visiolinguistic model empowered multimodal news recommendation[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022: 2560-2564.
26. Xun J, Zhang S, Zhao Z, et al. Why do we click: visual impression-aware news recommendation[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 3881-3890.
27. Yang C, Xiao Y, Zhang Y, et al. Heterogeneous network representation learning: A unified framework with survey and benchmark[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(10): 4854-4873.
28. Liu D, Shan L, Wang L, et al. P3oi-melsh: Privacy protection target point of interest recommendation algorithm based on multi-exploring locality sensitive hashing[J]. Frontiers in Neurorobotics, 2021, 15: 660304.
29. Zhang M, Wu S, Gao M, et al. Personalized graph neural networks with attention mechanism for session-aware recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(8): 3946-3957.
30. Ahmed M, Chakraborty P, Choudhury T. Bangla document categorization using deep RNN model with attention mechanism[C]//Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021. Springer Singapore, 2022: 137-147.
31. Peng D, Yuan W, Liu C. HARSAM: A hybrid model for recommendation supported by self-attention mechanism[J]. IEEE Access, 2019, 7: 12620-12629.
32. Yang X S, Hossein Gandomi A. Bat algorithm: a novel approach for global engineering optimization[J]. Engineering computations, 2012, 29(5): 464-483.
33. Yunanda G, Nurjanah D, Meliana S. Recommendation system from microsoft news data using TF-IDF and cosine similarity methods[J]. Building of Informatics, Technology and Science (BITS), 2022, 4(1): 277-284.
34. Z. Ou, Z. Han, P. Liu, S. Teng and M. Song, "SIIR: Symmetrical Information Interaction Modeling for News Recommendation," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2023.3299790.
35. K. Abbas et al., "Co-Evolving Popularity Prediction in Temporal Bipartite Networks: A Heuristics Based Model," in IEEE Access, vol. 11, pp. 37546-37559, 2023, doi: 10.1109/ACCESS.2023.3262587.
36. Hui B, Zhang L, Zhou X, et al. Personalized recommendation system based on knowledge embedding and historical behavior[J]. Applied Intelligence, 2022: 1-13.
37. Wu C, Wu F, Qi T, et al. Feedrec: News feed recommendation with various user feedbacks[C]//Proceedings of the ACM Web Conference 2022. 2022: 2088-2097.
38. Wei T, Feng F, Chen J, et al. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 1791-1800.
39. Long L, Yin Y, Huang F. Graph-aware collaborative filtering for top-N recommendation[C]//2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.

Jing Yu is with the Lu Xun Academy of Fine Arts, Shenyang 110004, China. She is an associate professor, has published many journal and conference papers related to her research. Research direction: Art Design, Multimedia art, interest recommendation, human-computer interaction.

Lu Zhao is with the Lu Xun Academy of Fine Arts, Shenyang 110004, China. She is an associate professor, has published many journal and conference papers related to her research. Research direction: Art Design, Multimedia art, interest recommendation, human-computer interaction.

Shoulin Yin received the B.S. and M.S. degrees in image processing from Software College, Shenyang Normal University, Shenyang, China, in 2013 and 2015, respectively. He is currently a PHD with School of Information and Communication Engineering, Harbin Engineering University. He has published more than 20 technical articles in scientific journals and conference proceedings. His research interests include image fusion, object detection, and image recognition.

Mirjana Ivanovic (Member, IEEE) has been a Full Professor with the Faculty of Sciences, University of Novi Sad, Serbia, since 2002. She has also been a member of the University Council for informatics for more than 10 years. She has authored or coauthored 13 textbooks, 13 edited proceedings, 3 monographs, and of more than 440 research articles on multi-agent systems, e-learning and web-based learning, applications of intelligent techniques (CBR, data and web mining), software engineering education, and most of which are published in international journals and proceedings of high-quality international conferences. She is/was a member of program committees of more than 200 international conferences and general chair and program committee chair of numerous international conferences. Also, she has been an invited speaker at several international conferences and a visiting lecturer in Australia, Thailand, and China. As a leader and researcher, she has participated in numerous international projects. She is currently an Editor-in-Chief of Computer Science and Information Systems Journal.

Received: January 15, 2024; Accepted: April 10, 2024.

Computer Science and Information Systems 21(3):1013–1031 https://doi.org/10.2298/CSIS230715029N

Agile Culture Clash: Unveiling Challenges in Cultivating an Agile Mindset in Organizations*

Michael Neumann[1], Thorben Kuchel[2], Philipp Diebold[2,4], and Eva-Maria Schön[3]

[1] University of Applied Sciences and Arts Hannover
Dpt. of Business Information Systems
Hannover, Germany
michael.neumann@hs-hannover.de
[2] Bagilstein GmbH
Mainz, Germany
{thorben.kuchel,philipp.diebold}@bagilstein.de
[3] University of Applied Sciences Emden/Leer
Dpt. of Economics
Emden, Germany
eva-maria.schoen@hs-emden-leer.de
[4] IU International University
Frankfurt, Germany
phillip.diebold@iu.org

Abstract. *Context:* In agile transformations, there are many challenges such as alignment between agile practices and the organizational goals and strategies or issues with shifts in how work is organized and executed. One very important challenge but less considered and treated in research are cultural challenges associated with an agile mindset. Although research shows that cultural clashes and general organizational resistance to change are part of the most significant agile adoption barriers. *Objective:* We identify challenges that arise from the interplay between agile culture and organizational culture. In doing so, we tackle this field and come up with important contributions for further research regarding a problem that practitioners face today. *Method:* This is done with a mixed-method research approach. First, we gathered qualitative data among our network of agile practitioners and derived in sum 15 challenges with agile culture. Then, we conducted quantitative data by means of a questionnaire study with 92 participants. *Results:* We identified 7 key challenges out of the 15 challenges with agile culture. These key challenges refer to the technical agility (doing agile) and the cultural agility (being agile). The results are presented in type of a conceptual model named the Agile Cultural Challenges (ACuCa). *Conclusion:* Based on our results, we started deriving future work aspects to do more detailed research on the topic of cultural challenges while transitioning or using agile methods in software development and beyond.

Keywords: Agile software development, agile methods, agile culture, agile mindset, organizational culture, challenges, barriers, conceptual model

* This paper is an extended version of our conference paper titled *Which Challenges Do Exist With Agile Culture in Practice?* [24]. In this paper, we updated the related work and added an in-depth presentation of the sample of our study. Furthermore, we provide a detailed discussion of each key challenge including a clustering of the challenges related to the dimensions of being and doing agile. Finally, we explain in more detail our conceptual model and present key take aways for practitioners and researchers.

1. Introduction

Agile methods have been used in software development for more than 20 years [1] and are widely adopted to respond to rapidly changing market conditions [4]. With an iterative approach and emphasis on collaboration, communication and interaction, agile methods allow for adequate responses to changing requirements and individual customer needs, enabling companies to remain competitive even in volatile markets. Nowadays, the term VUCA[5] (volatility, uncertainty, complexity, and ambiguity) is used to characterize this context [4]. However, values and principles are of high importance when introducing and using agile methods in practice due to their focus on social aspects such as communication, interaction, and stakeholder integration [8]. The Agile Manifesto [3] provides a ground for a value-based approach to product development. Furthermore, the guidelines for specific agile methods such as Scrum [36] or Kanban [2] define further and specific values for these approaches.

Many companies face various challenges when transitioning to agile methods [43]. A successful agile transformation depends on several factors, including the involvement of stakeholders [18], correct application of agile practices across different organizational levels [30], and cultural aspects of the organization [11]. The State of Agile Report [45] identifies contradictions between organizational culture and agile values as a core challenge in agile transition. For example, actively promoting the self-organization of agile teams requires a culture that values transparency and critical thinking regarding the agile approach in use [37]. However, challenges arising from cultural aspects when introducing and applying agile methods in practice are diverse [48]. Thus, contextual differences, such as regions, industries, or organizational levels, should be considered when studying agile methods [14, 32].

One well-known distinction in applying agile is between doing and being agile [25, 39]. Doing agile (technical agility) involves the methodological application of agile methods and practices, while being agile (cultural agility) emphasizes an agile mindset based on underlying values and principles. The interplay between these two dimensions is of high relevance for a successful agile transformation and provides a starting point for an understanding of an agile culture independent from the organizational context. Thus, we define agile culture as follows:

Definition ***Agile Culture [24]***: An Agile Culture reflects the behavior of people working in an organization using agile practices based on the underlying values and principles defined in the Agile Manifesto and the guidelines of agile methods.

Literature presents many studies regarding the application of technical agility [6, 19]. However, agile culture has been underrepresented in the recent literature (e.g., [14]). To address this gap, our mixed-method study aims to identify key challenges associated with the interplay between agile culture and organizational culture.

[5] The VUCA acronym was introduced by the U.S. military and describes markets that are characterized by highly dynamic conditions and thus, rapidly changing environments like customer needs and requirements. [4]

Hence, this paper addresses the following research questions:

- **RQ1:** *What are the key challenges with agile culture?*
  As described above, literature show various challenges with regard to agile transformations or the use of agile methods in practice. However, we identified a lack of understanding which challenges relate to an agile culture. Thus, we want to fill this gap by answering this research question by identifying key challenges with an agile culture.
- **RQ2:** *How can we describe the relationships of the key challenges with agile culture in a systematic manner?*
  We know that identifying and analyzing challenges can be complex as the challenges may have interrelations with each other. Thus, this second research questions aims to provide a systematic description of the relationships between the identified challenges.

The paper at hand is structured as follows: In Section 2, we outline the related work. Section 3 explains our mixed method research approach with qualitative and quantitative data. Section 4 presents our identified key challenges with agile culture and categorizes them to technical agility and cultural agility. Then, Section 5 presents our conceptual model that shows the relationships between the key challenges with agile culture. Section 6 discusses the implications on practitioners and researchers and presents the limitations of our study. Finally, the paper closes with Section 7, a conclusion of this work and future research directions.

2. Related Work

We searched for both, primary and secondary studies in order to provide an overview of the literature related to key challenges of cultural aspects while transitioning or using agile methods or practices. The search was performed using Google Scholar. We argue the choice to use Google Scholar by considering diverse publishers (e.g. ACM, IEEE) and the high consistency of results with digital libraries like Scopus [47]. The literature review was designed according to the guidelines for rapid reviews by Cartaxo et al. [7]. We focused our literature search on agile cultural related aspects. Table 1 gives a brief overview of the identified related work.

Various authors point to the multiple facets of culture [17, 32, 38, 40]. The overview of the related work also shows that different cultural levels (country, regions, companies, departments, teams, and individuals) may have different influences on the transition and use of agile methods and practices. In addition, agile methods can be very diverse in practice due to the tailoring of the respective method [10, 30]. Most of the articles we identified address the influence of organizational culture on the transition or use of agile methods and practices. Interestingly, different levels of detail are considered here. While some use organizational culture and its characteristics on an abstract level [21, 38], others try to make a specific reference to teams and individuals [44].

However, it is worth mentioning that most of the identified studies were published a decade ago. Agile has evolved since these studies were published. The diversity of agile practices and methods has grown steadily. Furthermore, the challenges in organizations have changed due to factors such as increasing digitization, volatile markets, and the Covid-19 pandemic [31].

Table 1. Overview of the related work in accordance with [24]

Reference	Cultural level(s)	Findings
Neumann et al. (2023) [32]	National culture & organizational culture	The authors present a causal model describing the impact of cultural aspects on agile practices. They focus on national culture using the six cultural dimensions by Hofstede and the two dimensions from the Competing Values Model. The hypothetical causal impact is described systematically per each cultural dimension on one agile practice. In total, their model presents 384 possible relationships of cultural impact on agile practices.
Gelmis et al. (2022) [14]	National culture	Hierarchical cultural structures, which are described for Turkish culture, are challenging the use and transition of agile methods. The authors recommend transforming to a more flat organizational structure, which supports the use of agile methods and should optimize the performance of agile teams.
Gupta et al. (2019) [17]	Organizational culture	The results of the study show that a hierarchical culture hinders the use of social and technical agile practices. Development cultures support both social and technical agile practices. The authors recommend considering the underlying cultures in the organization before starting an agile transformation and describe a six-step process according to Cameron and Quinn [5].
Iivari and Iivari (2011) [21]	Organizational culture	Based on an empirical study, the authors point out that a hierarchical culture is incompatible with the use of agile methods. For example, it is shown that the transition and use of agile methods in hierarchical organizational cultures leads to a steadily increasing formalization of the agile method and that it can lose key elements of agile work over time.
Siakas and Siakas (2007) [38]	National & organizational culture	In their work, the authors describe an agile culture based on well-known success factors such as user satisfaction and stakeholder involvement. They present the connections between organizational and cultural aspects and elements of agile working such as practices and roles. The authors also point out the importance of employee motivation and dynamics and draw parallels to established frameworks such as ETHICS [28] or TQM [9].
Strode et al. (2009) [43]	Organizational culture	Based on their empirical study, the authors present specific correlations between organizational culture characteristics and the successful application of agile methods. They point to the importance of positive evaluation of feedback and learning, a trustful social interaction, a collaborative cooperation and focusing on results.
Strode et al. (2022) [42]	Organizational culture	The authors discuss the challenges of transforming into an agile organization, which involves significant changes in strategy, structure, culture, operations, and technology. These transformations led to various tensions, of which 13 are highlighted. The study demonstrates that such tensions arise not only in fully agile organizations but also during the transformation process.
Tolfo et al. (2011) [44]	Organizational culture	The authors address the supporting and hindering influences of organizational culture on the transition of agile methods. A superficial examination of these influences can be a hindrance and lead to incorrect measures. Rather, it is important to consider the different levels of organizational culture in order to be able to develop an agile culture in the long term.

Although we tried to find related work dealing with key challenges of cultural aspects while using agile methods and practices, we could not identify literature dealing closely with the aim of our study. We identified some additional papers that deal with cultural facets or an agile mindset. However, we did not identify these papers as related work because they are not strongly related to our main topic, particularly the agile cultural aspect. So, to the best of our knowledge, this is the first publication that identifies key challenges with agile culture.

3. Research Design

This paper aims to identify challenges that arise from the interplay between agile culture and organizational culture.

Based on the aforementioned aim, we formulate the following research questions:

- **RQ1:** What are the key challenges with agile culture?
- **RQ2:** How can we describe the relationships of the key challenges with agile culture in a systematic manner?

RQ1 is answered with a mixed-method approach (see Fig. 3), where we collect qualitative and quantitative data and derive the key challenges. At each step of this mixed-method research approach, we gathered feedback from practitioners and discussed it extensively among the group of authors. Based on this, we develop the ACuCa model to answer our RQ2. The ACuCa model shows how the identified key challenges are interrelated. In the subsequent discussion, we show how practitioners can apply the key challenges and the ACuCa model in practice to mitigate problems with the agile culture.

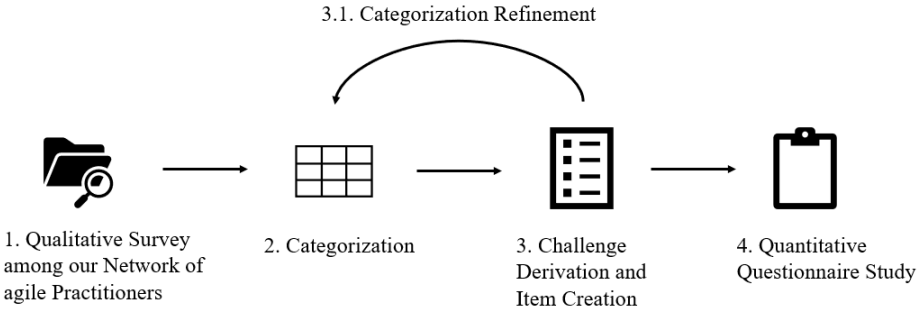


Fig. 1. Overview mixed-method research approach [24]

Step 1: In the beginning, we started with a qualitative survey among our network of agile practitioners in order to gather input for our quantitative questionnaire. Therefore, the third author of the paper asked the question on the professional network LinkedIn: “What do you see as tangible problems with agile and culture?” We discussed how to phrase this questions in such a manner that our target group understands it and we receive reasonable

data for our study. With this regard, we tried to avoid the term "agile culture" since this opens another field of discussions, that was not our focus in this research. In sum, we received 34 qualitative answers from 28 participants.

Step 2: We analyzed the data with a qualitative context analysis according to Hsieh and Shannon [20]. We used a combination of directed and conventional context analysis. For the former one, we used the research literature on Agile, and the latter one complemented the analysis since research on agile culture is rather limited. The data was categorized by a coding scheme, which ended up with 15 groups in Excel (agile values, willingness to change, fixed structures, and hierarchies, agile leadership, technical vs. cultural agility, respect, transparency of decisions, transparency of processes, trust, perseverance, feedback culture, experiments, failure culture, comfort zone, and flight levels). This preliminary result was intensively discussed among the authors and then refined where appropriate.

Step 3: In the next step, we derived items for our quantitative questionnaire. We took our 15 codes from Step 2 and used them for this purpose. First, we formulated one problem statement for each code based on the corresponding data sets. Then we reformulated these problem statements into questions. These items were discussed and refined among the authors in several iterations. An example that illustrates this process is shown in Table 2.

Table 2. Example of the derivation of a challenge (see Step 3) [24]

ID	Problem statement	Item (1. Iteration)	Item (Final)
C5	The existing structures of an organization prevent an agile way of working.	How important is it to create structures that support an agile way of working?	How important is it to create organizational structures that support an agile way of working?

Step 4: We then conducted the quantitative questionnaire study. Accordingly, we set up an online survey using Google Forms. We used appropriate guidelines for developing our questionnaire [15, 13, 35] and pretested it with five participants. In sum, the questionnaire comprises 21 items. Five of those items queried socio-demographic data which helps us to describe our sample, and 15 items queried the potential key challenges with agile culture (see C1-C15 in Appendix A.1) and one item posed additional comments. The participants of the survey rated the potential key challenges with agile culture (see C1-C15 in Appendix A.1) in terms of importance using a 7-point Likert scale (totally unimportant, unimportant, rather unimportant, neutral, rather important, important, and totally important). In addition, participants had the option to give no statement.

3.1. Data Collection and Analysis

In the following, we present more details regarding the data collection and analysis of our quantitative questionnaire (see Step 4). The quantitative questionnaire was online between 11/26/2021 and 12/17/2021. It was spread via the professional network LinkedIn and personal contacts of the authors. In sum, we received 93 fully completed datasets.

No participant terminated the questionnaire before the end. So, the dropout rate is 0%. This surprised us, but can be explained by the good design of the questionnaire, oriented towards the target group. However, we removed one set of data from the sample because one participant showed unusual response behavior. The unusual response behavior was identified by analyzing the data set in detail. We found one respondent who selected the same value of the Likert scale for all questions except one.

Furthermore, we used statistical data (mean, standard deviation, and confidence interval) for the analysis of the data. We defined challenges as key in those cases where 50% of the participants weighted them as totally important. Based on this interpretation, we were able to identify a total of seven key challenges (see Table 4).

3.2. Description of the Sample

The participants of the quantitative questionnaire (sample size N=92) had between 0 to 30 years (mean 6.13) of experience with an agile way of working. Of the 92 participants, 74 work in the private sector, 9 in the public sector, 10 work at universities or research institutes while 18 are self-employed (multiple answers were possible). A detailed description of the sample is given below in Table 3.

Table 3. Description of the sample of the quantitative questionnaire study (Step 4)

ID	Item	Options	n	N
E1	Experience with agile ways of working	Yes	85	92
		No	7	
E2	Years of experience with agile ways of working	$<$ 3 years	21	85
		3-5 years	18	
		6-9 years	20	
		$>$ 9 years	26	
E3	General opinion on agility	Overrated	2	92
		Neutral	1	
		Applied where useful	58	
		Urgently needed	31	
D1	Type of organization	Private	74	92
		Public Institutes	19	
		Self-employed	18	
D2	Industry	Consulting	20	92
		E-Commerce	12	
		R&D	7	
		IT	56	
		Other	22	

Moreover, participants work in different industries such as IT, consulting, e-commerce, research and development, automobile, logistics, eHealth, insurance, energy, and social. We asked the participants for their general opinion on agility and they answered as follows: “overrated, a buzzword to me” (2,2%), “should be applied where appropriate” (63%), “urgently needed in today’s world” (33,7%), and “neutral” (1,1%). Summarizing this information, we can observe that on the one hand, the sample has a good mix

in terms of their experience with agile ways of working as well as working in different industries. On the other hand, the sample is not opposed to agile ways of working.

4. Key Challenges with Agile Culture

In this section, we answer our first research question *RQ1: What are the key challenges with agile culture?* In doing so, we have identified seven key challenges related to agile culture, which are presented in Table 4. To evaluate the responses, we used the Likert scale options to range from 1 (totally unimportant) to 7 (totally important). The challenges in Table 4 were rated as totally important by over 50% of our survey participants. Additionally, all key challenges were found to be between important and totally important, as indicated by the small confidence interval. The statistical data for all queried challenges is provided in Appendix A.1.

Table 4. Key Challenges with Agile Culture [24]

ID	Item (EN)	Mean	Stand. deviation	Confid. (p=0.05)	totally important	N
C1	Humans in an organization do not treat each other with respect.	6.70	0.81	0.17	80.4%	92
C2	Management expects change from employees without embodying agile values themselves.	6.43	1.17	0.24	69.6%	92
C3	The organizational culture does not create a context for trusting interactions.	6.60	0.63	0.13	67.4%	92
C4	Humans in an organization are not allowed to make mistakes.	6.49	0.75	0.15	60.9%	92
C5	The existing structures of an organization prevent an agile way of working.	6.40	0.96	0.20	59.8%	92
C6	Agile cultural change does not occur at all levels (individual, team, management) of the organization.	6.23	1.12	0.23	53.3%	90
C7	Feedback is not valued in an organization.	6.35	0.76	0.16	51.1%	92

Our analysis revealed that there are challenges related to embracing agile values in the organization, particularly with respect (C1) and trust (C3, C4, and C7). Furthermore, our data highlighted issues with existing organizational structures (C5) that hinder agile ways of working, as well as issues with cultural change at various levels within an organization (C6).

4.1. Discussion of the Key Challenges with Agile Culture

The identified key challenges with agile culture (see Table 4) can be categorized into technical agility (doing agile) and cultural agility (being agile). Figure 4.1 shows the result of the clustering process. The challenges leadership (C2) and feedback culture (C7) are borderline cases as parts of the problem relate to technical agility and other parts to

cultural agility. In the case of leadership (C2, Management expects change from employees without embodying agile values themselves) parts such as existing control structures and micro management in the respective organization relates to technical agility, whereas embodying agile values or having a fixed mindset [12] is part of cultural agility. The case of feedback culture (C7, Feedback is not valued in an organization) relates to technical agility, e.g., if no tools or agile practices for providing feedback are in place, whereas valuing feedback is part of cultural agility.

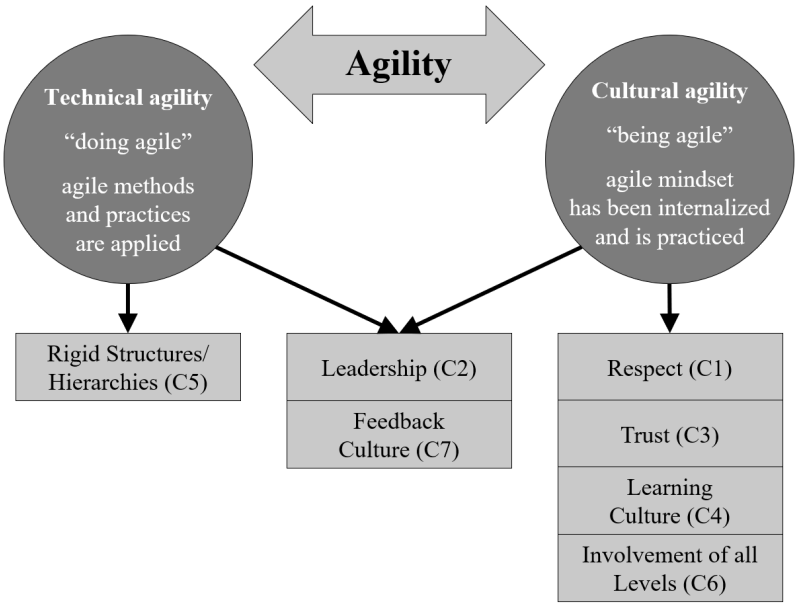


Fig. 2. Clustering of key challenges according to technical agility and cultural agility

In the following, we present a detailed explanation of each key challenge with agile culture.

Item **C1 (respect)** highlights that within an organization, there is a lack of respectful behavior towards each other, as indicated by 80% of the respondents considering it very important. The workplace often faces the common issue of lacking respect in communication and interactions, which can be addressed to create a positive work environment. Fostering a respectful atmosphere promotes openness and transparency, empowering teams to self-manage their work. Agile leaders have a responsibility to exemplify respectful behavior, and when respect becomes a core value actively practiced within an organization, it can motivate employees and extend to the respectful treatment of customers. Mutual respect cultivates open communication and collaboration, fostering constructive and productive work processes.

Item **C2 (leadership)** represents another highly relevant aspect. The underlying leadership structure should emphasize agile values such as trust, respect, or open communication to support the continuous and step-wise adaption of agile practices in organizations.

Today's complex work environment requires a different style of leadership to that of previous industrial revolutions. For example, agile leaders take responsibility and inspire performance. They also show empathy and vulnerability. Probably one of the biggest changes that modern organizations and leaders have to go through in the course of agile transformation is the delegation of responsibility to self-organized teams [23]. This change requires a shift in the mindset of leaders and an adaptation of existing control structures.

Item **C3 (trust)** points to the significance of trust in an agile work culture. It emphasizes that trust among team members and between leaders and employees is crucial for cooperation and interdisciplinary work. The item reveals that 67% (n=62) of respondents rated the absence of a framework for a trusting work environment as "very important." The lack of trust can impede the sharing of knowledge and skills, which are vital for applying agile methods and practices. Agile leaders should embody company values such as respect, courage, and openness to establish trust. A trusting work environment can eliminate time-consuming and costly control structures, allowing individuals to focus on their core tasks.

Item **C4 (learning culture)** is highly important within an organization in order to view mistakes as growth opportunities rather than solely blaming individuals. We propose that organizations should prioritize learning from mistakes instead of assigning blame. This can be achieved through open communication that encourages experimentation and public acknowledgment of mistakes. Agile leaders should lead by example by admitting and openly discussing their own mistakes. By fostering an environment where mistakes are not feared but embraced as chances for improvement, organizations can enhance innovation and motivation. It is crucial to emphasize that allowing individuals to make mistakes is insufficient; handling mistakes constructively and positively is equally vital.

Item **C5 (rigid structures/hierarchies)** points to the importance towards an agile way of working due to adapting the existing structures and hierarchies. A hierarchical culture has been shown to have a negative impact on the use of social agile practices (e.g. facilitating social interaction, collaboration, and direct communication) and a negative impact on the use of technical agile practices (coding/testing-oriented software engineering practices) [17]. For this reason, communication at eye level should be encouraged within an organization, regardless of the level (individual, team, management) at which it takes place.

Item **C6 (involvement of all levels)** addresses the significance of a cultural shift in introducing agile working within an organization. The management level expects this change to initiate from the level below them, known as the flight levels (operational level, coordination, and strategic portfolio management) [26]. However, a top-down approach cannot enforce a cultural change, and intrinsic motivation among employees at all levels is necessary for successful agile transformation. Agile leaders play a critical role by involving and engaging employees, helping them understand and identify with the change. Flight levels, which represent different perspectives on agile work within an organization, can be utilized to gradually introduce agile structures, without requiring every individual to work in an agile manner simultaneously. A cultural change on all levels, including leadership, is indispensable for long-term success in adapting to evolving organizational requirements.

Item **C7 (feedback culture)** is a pivotal aspect of agile work and influences organizational culture [43]. Feedback enables continuous reflection and adaptation, while providing a safe space for individuals to express their thoughts and ideas. However, many organizations undervalue feedback. The quality of dialogue among colleagues and superiors influences relationship quality, which, in turn, reflects the company's culture. The retrospective, conducted at the end of an iteration, plays a key role in establishing a secure environment for a feedback culture. Feedback is recognized as the key to continuous development, lifelong learning, and constructive collaboration [34]. Embracing an open feedback culture can help organizations better navigate ongoing environmental changes and transition towards agile work practices through continuous and constructive dialogue.

5. Model of Agile Cultural Challenges

This section answers our second research question *RQ2: How can we describe the relationships of the key challenges with agile culture in a systematic manner?* The clustering of the key challenges (see Fig. 4.1) enables us to derive a conceptual model that shows the interrelationships between the key challenges with agile culture. A conceptual model allows us to clarify complex ideas and relationships by providing a visual representation. The result of the clustering process (see Fig. 4.1) provided the ground for creating our conceptual model named **A**gile **Cu**ltural **Cha**llenges (ACuCa) (see Fig. 5). The ACuCa model helps to identify causal factors, dependencies and potential outcomes associated with the key challenges. In this Section, we first give a brief explanation about the creation process of the ACuCa conceptual model. Next, we introduce the model on detail and explain in particular the relationships among the clustered key challenges.

The model was created by the second author and improved based on a thorough review by the fourth author. The creation process consisted of two steps. In the first step, we clustered the key challenges. Our analysis led to the result that we have mainly different facets in the conceptual model. One facet covers the value-related key challenges, while the other facet consists of such challenges that implicitly or explicitly affect each other. Thus, we decided to create one cluster for the value-related key challenges C1, C3, C4, and C7. We also created three clusters, each consisting of one key challenge.

The second step aimed to analyze and identify the realtionships among the key challenges using existing literature. We decided to use the guidelines of the well-known agile method Scrum [36] as well as the Agile Manifesto [3]. In addition, the analysis included the extracted insights from the identified related work. For example, our analysis led to an in-depth understanding of how specific key challenges (e.g. C5) influence the core of clustered value-related key challenges.

We found that fixed structures, established thought patterns, and strict hierarchies with a command-and-control approach are major barriers to agile ways of working. The existing structures of an organization hinder an agile way of working, and a change in both culture and structural conditions is required for introducing more agility. The interaction between the two dimensions of cultural and technical agility (see Fig. 4.1) cannot be neglected, as structural conditions can have both performance-enhancing and inhibiting effects. Hierarchy, silos, and micromanagement pose a major hurdle to agile cultural development, and the biggest challenge lies in overcoming old paradigms and enabling a shift towards a structure that focuses on customer and business agility [29]. In larger

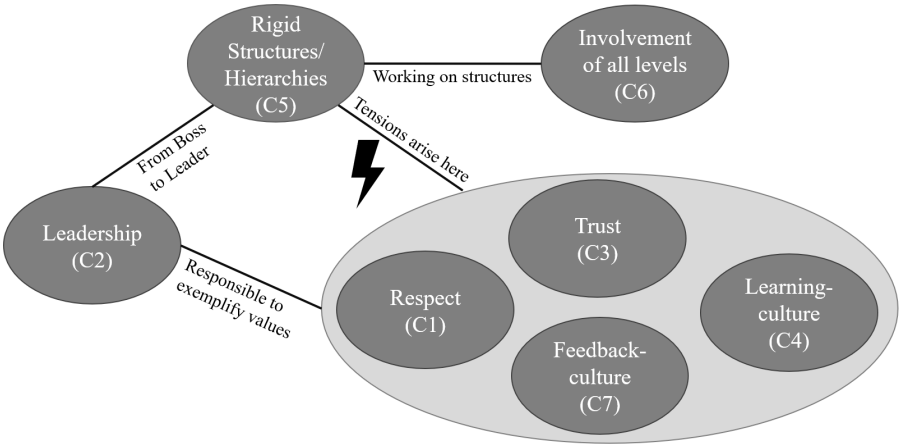


Fig. 3. The ACuCa conceptual model [24]

organizations, this can lead to the creation of silos that are more product-, value-, and customer-centric. It is important to adapt the dimensions of culture and structure in small steps that are coordinated with each other.

Organizational structures involve all levels (individual, team, organization), and organizations should iteratively adapt their structures to optimize interactions between levels while considering the organizational culture. Additionally, organizational structures contain the management level, and changing from a boss to a leader culture is a key factor for increased agility. This involves exemplifying values, behaviors, and a culture of open feedback and learning to establish a growth mindset [12] among employees. We define a boss as a manager who has a hierarchical understanding of the organization and a top-down approach to decision-making. We believe it is crucial for agile leaders to act as role models for agile teams within an organization.

To lead effectively in an agile environment, leaders must have experienced agility themselves and embody agile values. The role of a leader shifts towards becoming a coach and role model, rather than a commander. Self-organized teams require guidance through shared visions, values, and transparency, with a focus on communication and motivation. Leaders must set the frame for the team and plan adaptively while providing protection, trust, continuity, and support for employees' development through coaching and mentoring. By demonstrating and embodying agile values, leaders can facilitate the organization's adoption of agile practices step-by-step. The transformation towards an agile leadership style requires targeted training and development for future agile leaders.

6. Discussion

In the following section, we discuss the implications for practitioners and researchers and suggest ways in which the key challenges we have identified can help to promote an agile culture. In addition, we discuss the limitations of this study.

6.1. Discussion of Implications for Practitioners and Researchers

Comparing our research findings (see Section 4) to the existing literature (see Section 2), we discovered that cultural conflicts and resistance to change are major barriers to agile adoption [45]. Our key challenges with agile culture (listed in Table 4) emphasize the importance of respect (C1) and trust (C3, C4, and C7) as crucial factors in the success of an agile transformation. These challenges could be attributed to inadequate support and guidance during the transformation process [22].

To address these challenges, our research provides practical solutions for both practitioners and researchers. Practitioners can use our key challenges as a tool to drive cultural change in their organizations. The challenges raise awareness of problems that come up when trying to integrate agility into an existing organizational culture, and help to prioritize which challenges are most important to tackle. For example, C6 highlights the importance of considering the pace of cultural change across different levels (individual, team, organization), and practitioners can use this information to develop appropriate solutions.

Changing organizational culture is a complex process that requires time and appropriate models or frameworks [22, 23]. The Cynefin framework [41], Haufe Quadrant [16] or the Deming circle [9] are useful models for initiating and managing change. Researchers can support agile transformation and cultural change by developing appropriate tools, for example, to measure agility and thus support agile transformation [27, 33, 38]. Based on this tool support, researchers may develop new or adapted approaches, methods and even tools specifically focused on agile cultural change in an organization, using our identified key challenges as a basis for its development. Our key takeaways for practitioners and researchers are summarised below:

Key takeaways:

- Live the agile values, especially respect and trust
- Leadership must embody agile values
- Trust is the basis for cooperation, knowledge transfer and interdisciplinary interaction
- Innovation requires experimentation, which sometimes fails
- Focus on customers and users
- Communication at eye level instead of power and hierarchy
- Transformation should take place at all levels (Flightlevels model)
- Agile working is about living and sharing in real time, so it needs a culture that values feedback

6.2. Threats to Validity & Limitations

Despite our rigorous adherence to establishing a research protocol by means of using appropriate guidelines and using a profound research approach, here are limitations and threats in our study that must be considered and discussed, according to Wohlin et al. [46].

Construct validity: Our mixed-method research approach included pretests and feedback from the target group to ensure the validity of our data collection. To prevent interpolations in our quantitative survey, we used a 7-point Likert scale and provided the option for participants to choose ”no comment” for each item. As a lesson we learned, next time

we would rephrase ”no comment” to ”no answer” even if from our perspective no comment fits better to our concrete question. Furthermore, for the comparison and check with the existing state of the art, we used a rapid review [7]. We are aware of the fact that the replicability of the survey is limited due to the open target group of our professional network. Nevertheless, we did not only focus on agile practitioners in the network but spreaded it to all. However, we know that the practitioners in our community are more related to agile than other topics. Apart from a higher dropout rate or lower participation rate, we accepted this threat for our survey.

Internal validity: The design of the questionnaire is crucial for accurate data collection. Therefore, we conducted several pretests in different stages with individuals who met the criteria of the target group. All these pretests resulted in a 0% dropout rate. In addition, we removed the data set of the one participant who showed conspicuous response behaviour.

External validity: The surveys were conducted in German, as we aimed to reach individuals within our personal professional/business networks. While this approach allowed us to achieve a 0% dropout rate in the quantitative survey, it may limit the generalizability of our results to other groups. However, since our demographics showed that many participants work in the international IT industry, the general transferability of our findings may not be significantly impacted. To assess the generalizability of our results further we have started further studies to conduct surveys in other languages.

7. Conclusion and Future Work

This paper is an extended version of our study *Which Challenges Do Exist With Agile Culture in Practice?* [24]. This extended version presents the findings of a study aimed at addressing the key challenges of implementing agile culture in practice. We further present ACuCa as a conceptual model explaining the relationships between the identified challenges. Seven key challenges, including respectful treatment between humans (C1), agile leadership (C2), trust in interactions (C3), learning culture (C4), rigid hierarchies (C5), involvement of all organizational levels (C6), and not valued feedback (C7), emerged from the interplay of organizational culture and agile practices.

The ACuCa model reveals that a rigid hierarchical structure impedes the development of an agile culture aimed at optimizing aspects such as value-based work and learning culture. Moreover, involvement of all organizational levels, including management, teams, and individuals, is essential for optimizing both the structure and culture of the organization. To support agile software development teams' self-optimization, we encourage the management level to transform from a boss to a leader and exemplify relevant values for successful agile method usage. Based on our results, we provide practical implications and specific recommendations for cultural change in organizations, such as utilizing transformation frameworks like the Cynefin framework or the Deming circle.

However, our findings reveal the need for further research on cultural influences on agile methods and practices. As agile methods become more tailored, selecting the right approach in specific situations becomes more complex. The various cultural aspects and influences described at different levels, such as the department or specific teams, may pose complex challenges in organizations. Therefore, we call on other researchers to study the relationship and interplay of cultural influences on the use and transition of agile methods

in software development and beyond. We support the findings of other researchers who have suggested the usage of transformation frameworks and encourage more research to be conducted in this area, as there is still much to be explored.

References

1. Abrahamsson, P., Salo, O., Ronkainen, J., Warsta, J.: Agile software development methods: Review and analysis (478), 7–94 (2002)
2. Anderson, D.J.: Kanban - Successful Evolutionary Change for your Technology Business. Blue Hole Press. (2010)
3. Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Greening, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R., Mellor, S., Schwaber, K., Sutherland, J., Thomas, D.: Agile manifesto (2001), https://agilemanifesto.org/
4. Bennett, N., Lemoine, G.: What a difference a word makes: Understanding threats to performance in a vuca world. Business Horizons 57(3), 311–317 (2014), https://www.sciencedirect.com/science/article/pii/S0007681314000020
5. Cameron, K.S., Quinn, R.E.: Diagnosing and changing organizational culture: Based on the competing values framework. Addison-Wesley, Reading (2006)
6. Campanelli, A.S., Parreiras, F.S.: Agile methods tailoring – a systematic literature review. Journal of Systems and Software 110, 85–100 (2015)
7. Cartaxo, B., Pinto, G., Soares, S.: Rapid Reviews in Software Engineering, pp. 357–384. Springer International Publishing, Cham (2020)
8. Chow, T., Cao, D.B.: A survey study of critical success factors in agile software projects. Journal of Systems and Software 81(6), 961–971 (2008)
9. Deming, W.E.: Out of the Crisis: Quality, Productivity and Competitive Position. MIT Press (1986)
10. Diebold, P., Dahlem, M.: Agile practices in practice: A mapping study. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (2014)
11. Diebold, P., Küpper, S., Zehler, T.: Nachhaltige agile transition: Symbiose von technischer und kultureller agilität. In: Engstler, M., Fazal-Baqaie, M., Hanser, E., Mikusz, M., Volland, A. (eds.) Projektmanagement und Vorgehensmodelle 2015. pp. 121–126. Gesellschaft für Informatik e.V., Bonn (2015)
12. Dweck, C.: Mindset: Changing The Way You think To Fulfill Your Potential (Updated ed.). Robinson, London (2017)
13. Finstad, K.: Response interpolation and scale sensitivity: Evidence against 5-point scales. J. Usability Studies 5(3), 104–110 (may 2010)
14. Gelmis, A., Ozkan, N., Ahmad, A., Guler, M.: Impact of turkish national culture on agile software development in turkey. In: Lean and Agile Software Development. pp. 78–95. Springer International Publishing, Cham (2022)
15. Gräf, L.: Assessing internet questionnaires: The online pretest lab. Online social sciences pp. 49—-68 (2002)
16. Group, H.: Haufe quadrant (2023), https://www.haufe.de/konzept/haufe-quadrant
17. Gupta, M., George, J., Xia, W.: Relationships between it department culture and agile software development practices: An empirical investigation. International Journal of Information Management 44, 13–24 (2019), https://www.sciencedirect.com/science/article/pii/S0268401218306662
18. Hoda, R., Noble, J., Marshall, S.: The impact of inadequate customer collaboration on self-organizing agile teams. Information and Software Technology 53(5), 521–534 (2011)
19. Hoda, R., Salleh, N., Grundy, J., Tee, H.M.: Systematic literature reviews in agile software development: A tertiary study. Information and Software Technology 85, 60–70 (2017)

20. Hsieh, H.F., Shannon, S.E.: Three approaches to qualitative content analysis. Qualitative Health Research 15(9), 1277–1288 (2005)
21. Iivari, J., Iivari, N.: The relationship between organizational culture and the deployment of agile methods. Information and Software Technology 53(5), 509–520 (2011), https://www.sciencedirect.com/science/article/pii/S095058491000193X, special Section on Best Papers from XP2010
22. Kotter, J.: Leading change: Why transformation efforts fail (1995), https://hbr.org/1995/05/leading-change-why-transformation-efforts-fail-2
23. Krieg, A., Prenner, N., Guckenbiehl, P., Theobald, S., Schneider, K.: A scientific baseline for agile leadership - a workshop study. In: Fazal-Baqaie, M., Linssen, O., Volland, A., Yigitbas, E., Engstler, M., Bertram, M., Kalenborn, A. (eds.) Projektmanagement und Vorgehensmodelle 2022. pp. 25–36. Gesellschaft für Informatik e.V., Bonn (2022)
24. Kuchel, T., Neumann, M., Diebold, P., Schön, E.M.: Which challenges do exist with agile culture in practice? In: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing. p. 1018–1025. SAC '23, Association for Computing Machinery, New York, NY, USA (2023)
25. Küpper, S., Kuhrmann, M., Wiatrok, M., Andelfinger, U., Rausch, A.: Is there a blueprint for building an agile culture? In: Volland, A., Engstler, M., Fazal-Baqaie, M., Hanser, E., Linssen, O., Mikusz, M. (eds.) Projektmanagement und Vorgehensmodelle 2017 - Die Spannung zwischen dem Prozess und den Mensch im Projekt. pp. 111–128. Gesellschaft für Informatik, Bonn (2017)
26. Leopold, K.: Practical Kanban From Team Focus to Creating Value. Leanpub (2017), http://leanpub.com/practicalkanban
27. Looks, H., Fangmann, J., Thomaschewski, J., Escalona, M., Schön, E.M.: Towards a standardized questionnaire for measuring agility at team level. In: Agile Processes in Software Engineering and Extreme Programming. XP 2021. pp. 71—-85 (2021)
28. Mumford, E.: Socio-technical design: an unfulfilled promise or a future opportunity? In: Organisational and Social Perspectives on Information Technology, IFIP TC8 WG8.2 International Working Conference (2000)
29. Narayan, S.: Agile IT organization design: for digital transformation and continuous delivery. Pearson Education, Inc (2015)
30. Neumann, M.: The integrated list of agile practices - a tertiary study. In: Lean and Agile Software Development. pp. 19–37. Springer International Publishing, Cham (2022)
31. Neumann, M., Bogdanov, Y.: The impact of covid 19 on agile software development: A systematic literature review. In: Proceedings of the 55th Hawaii International Conference on System Sciences (2022)
32. Neumann, M., Schmid, K., Baumann, L.: Characterizing the impact of culture on agile methods: The moca model. In: Proceedings of the 17th International Conference on Software and System Processes (2023)
33. Patel, C., Ramachandran, M.: Agile maturity model (amm): A software process improvement framework for agile software development practices. International Journal of Software Engineering 2(I), 3—-28 (2009)
34. Ram, P., Rodriguez, P., Oivo, M., Martinez-Fernandez, S.: Success factors for effective process metrics operationalization in agile software development: A multiple case study. In: ICSSP 2019. pp. 14–23 (2019)
35. Runeson, P., Höst, M.: Guidelines for conducting and reporting case study research in software engineering. Empirical Software Engineering 14(2), 131–164 (2009)
36. Schwaber, K., Sutherland, J.: The scrum guide (2020), https://www.scrumguides.org/scrum-guide.html
37. Schön, E., Escalona Cuaresma, M., Thomaschewski, J.: Agile values and their implementation in practice. International Journal of Interactive Multimedia and Artificial Intelligence 3(5), 61–66 (2015)

38. Siakas, K., Siakas, E.: The agile professional culture: a source of agile quality. Software Process: Improvement and Practice 12(6), 597–610 (2007)
39. Sidky, A., Arthur, J., Bohner, S.: A disciplined approach to adopting agile practices: the agile adoption framework. Innovation in Systems and Software Engineering 3(3), 203—-216 (2007)
40. Šmite, D., Gonzalez-Huerta, J., Moe, N.B.: "when in rome, do as the romans do": Cultural barriers to being agile in distributed teams. In: Proceedings of the 21st International Conference on Agile Software Development 2020. pp. 145–161. Springer International Publishing, Cham (2020)
41. Snowden, D., M.E., B.: A leader's framework for decision making (2007), https://hbr.org/2007/11/a-leaders-framework-for-decision-making
42. Strode, D.E., Sharp, H., Barroca, L., Gregory, P., Taylor, K.: Tensions in organizations transforming to agility. IEEE Transactions on Engineering Management 69(6), 3572–3583 (2022)
43. Strode, D., Huff, S., Tretiakov, A.: The impact of organizational culture on agile method use. In: Proceedings of the 42nd Hawaii International Conference on System Sciences. pp. 1–9 (2009)
44. Tolfo, C., Wazlawick, R., Gomes Ferreira, M., Forcellino, F.: Agile methods and organizational culture: reflections about cultural levels. Journal of Software Maintenance and Evolution: Research and Practice 12(6), 597–610 (2011)
45. VersionOne, Collabnet: 17th annual state of agile survey (2023), stateofagile.com
46. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in Software Engineering. Springer Berlin, Heidelberg (2012)
47. Yasin, A., Fatima, R., Wen, L., Afzal, W., Azhar, M., Torkar, R.: On using grey literature and google scholar in systematic literature reviews in software engineering. IEEE Access 8, 36226–36243 (2020)
48. Šmite, D., Moe, N.B., Gonzalez-Huerta, J.: Overcoming cultural barriers to being agile in distributed teams. Information and Software Technology 138, 106612 (2021)

A. Appendices

A.1. Challenges with Agile Culture

The full data set including the 15 challenges with agile and organizational culture is shown in Table 5.

Table 5. Appendix 1: Key Challenges with Agile Culture [24]

ID	Item (EN)	Mean	Stand. deviation	Confid. (p=0.05)	Confid. intervall min.	Confid. intervall max.	totally important	N
C1	Humans in an organization do not treat each other with respect.	6.70	0.81	0.17	6.53	6.86	80.4%	92
C2	Management expects change from employees without embodying agile values themselves.	6.43	1.17	0.24	6.20	6.67	69.6%	92
C3	The organizational culture does not create a context for trusting interactions.	6.60	0.63	0.13	6.47	6.73	67.4%	92
C4	Humans in an organization are not allowed to make mistakes.	6.49	0.75	0.15	6.34	6.64	60.9%	92
C5	The existing structures of an organization prevent an agile way of working.	6.40	0.96	0.20	6.21	6.60	59.8%	92
C6	Agile cultural change does not occur at all levels (individual, team, management) of the organization.	6.23	1.12	0.23	6.00	6.47	53.3%	90
C7	Feedback is not valued in an organization.	6.35	0.76	0.16	6.19	6.50	51.1%	92
C8	Humans in an organization do not want to change the way they work.	6.16	1.03	0.21	5.94	6.37	46.7%	90
C9	In an organization, there is no transparency regarding decisions.	6.35	0.76	0.16	6.13	6.44	45.7%	92
C10	In an organization, there is no transparency regarding processes.	6.25	0.76	0.16	6.09	6.41	45.1%	91
C11	Humans in an organization do not understand that agile transformation needs perseverance.	6.13	0.94	0.19	5.94	6.33	44.4%	90
C12	Humans in an organization don't want to leave their comfort zone.	5.97	1.05	0.22	5.75	6.18	37.8%	90
C13	There is no safe environment for experiments in the organization.	5.97	0.87	0.18	5.79	6.15	30.8%	91
C14	Humans in an organization have not internalized the difference between technical and cultural agility.	5.73	1.18	0.24	5.49	5.97	30.3%	89
C15	The cultural values of humans within an organization do not align with agile values.	5.96	1.12	0.23	5.72	6.19	30.0%	90

Michael Neumann is a PhD candidate at the University of Hildesheim. Currently, he works as an administrative professor for the Department of Business and Informatics at the University of Applied Sciences & Arts Hannover. His research focuses on influencing and success factors in the field of agile software development. He has many years of practical experience in agile software development in various different contexts.

Eva-Maria Schön received her PhD in computer science from the University of Seville (Spain) in 2017. She is a Professor of Business Informatics at University of Applied Sciences Emden/Leer since 2022. In addition, she has many years of practical experience in the development of digital products and has broad experience in training and consulting for various companies.

Philipp Diebold is the founder, managing partner and agile coach at Bagilstein GmbH since 2018. Together with his colleagues, he continues the application-oriented research from his dissertation (2018) at Fraunhofer with scientifically sound and practice-oriented consulting and coaching. He is also a professor of business informatics at the iu International University in Frankfurt.

Thorben Kuchel holds a Bachelor degree in Business Informatics from HAW Hamburg. He is working as an Agile Coach at Bagilstein GmbH aiming to support organizations in their agile transformation.

Received: July 15, 2023; Accepted: May 10, 2024.

Computer Science and Information Systems 21(3):1033–1054 https://doi.org/10.2298/CSIS230426026B

Graph Rewriting Primitives for Semantic Graph Databases Sanitization*

Adrien Boiret, Cédric Eichler, Benjamin Nguyen, and Sara Taki

INSA Centre Val de Loire, Laboratoire d'Informatique Fondamentale d'Orléans
88 Bd Lahitolle, 18000 Bourges, France
{adrien.boiret,cedric.eichler, benjamin.nguyen, sara.taki}@insa-cvl.fr

Abstract. Due to the rapid proliferation of data online, an important quantity of private or sensitive informations is being stored as linked data in graph databases (*e.g.*, represented as RDF). For such databases to be shared without jeopardizing privacy, they must first undergo a process known as database sanitization. During this process, databases are transformed following graph transformations that are usually described informally or through ad-hoc processes. However, a more thourough formalization of these transformations would aid in analysing the sanitization process, ensuring its correctness, and demonstrating the resulting privacy guarantees. This paper is an effort toward bridging the gap between the rigorous graph rewriting approaches and graph sanitization. We propose a graph transformation language to serve as a basis for constructing various sanitization mechanisms. This language relies on a set of elementary transformation operators formalized using a generic algebraic graph rewriting approach. Our language takes into account semantic and supports the equivalent of WHERE and EXCEPT clauses. As a proof of concept, we use these operators to implement two mechanisms from the literature, one generic (Local Differential Privacy) and one specifically introduced for semantic graph databases (sensitive attribute masking through anatomization). We propose an open-sourced tool implementing the elementary operators and the privacy mechanisms we derive from them relying on the Attributed Graph Grammar System (AGG) and its java API, providing a concrete tool implementing formal graph rewriting mechanisms to sanitize semantic graph databases. We present experimental results on this implementation regarding both proposed schemes and discuss its efficiency and scalability.

1. Introduction

In many collaborative data centric applications that collect personal data, such as car pooling, or smart metering, the data collectors (*i.e.*, the entity managing the application) need to publish and share raw data with various parties. These parties can range from internal developpers who need test data to data analysts in charge of producing predictive or explicative models, or even simply the scientific community. Since this database is composed of personal data, the European General Data Protection Regulation (GDPR) rules apply. In this paper, we consider the case where the database is sanitized (anonymized) prior to its release, which is compatible with the GDPR. Sanitization of a database means

* This paper is an extended version of [1]

that it is no longer possible (or at least very difficult, costly, and time consuming) to reidentify individuals in the dataset.

Nowadays, data are often organized as graphs with an underlying semantic to allow efficient querying and support inference engines. Such is the case in, for example, linked data and semantic web typically relying on RDF representation. Yet, while anonymization of tabular databases and untyped homogeneous graphs are well-researched, anonymization processes in such databases have been mildly studied. Most existing graph sanitization processes describe the underlying graph transformations in an informal, add-hoc, format-specific, and/or ambiguous way. We argue that such transformations should be 1) formalized, to ease the analysis of the process and the verification of both its correctness and the guarantees it provides; 2) generic and format-agnostic to improve the applicability of sanitization processes. In these regards, rigorous graph rewriting approaches appear to be perfect candidates.

In general, and despite the importance of graphs in databases and ontology representations, the use of formal graph rewriting techniques to model database evolutions is seldom studied. To the best of our knowledge, [1] is the only proposal using rigorous graph rewriting to model graph transformations for graph sanitization.

The present paper is an extension of [1]. We generalize this approach by formalizing a language of basic operators using attributed graph rewriting rules to serve as a basis for a greater number of graph sanitization mechanisms. We show this formalization in action on two examples privacy schemes, one classical (Local Differential Privacy (LDP)) already expressed in [1] and one specific to semantic graphs (anatomization) not explored in [1].

Contributions In this paper, we formalize and implement basic operators using AGG – The Attributed Graph Grammar System [22], one of the most mature development environments supporting the definition and application of typed graph rewriting systems [19]. These operators demonstrate the feasibility of the approach and should be enriched with other operators to build a library of operators supporting various anonymization schemes.

We choose to focus on eight basic operators. The seven first operators create or delete nodes, copy, cut, or merge edges, or randomize the targets of a relation. The eighth operator help better identify sets of nodes. Each of these operators can be represented as a single graph rewriting rule. Together, they can be combined into procedures expressive enough for our announced examples, randomization providing LDP guarantees and sensitive attribute anatomization.

We show that, given our formalization and implementation it is easy to describe sets of nodes or edges as the scope of our operators. This can be intuitively seen as clauses:

- WHERE allows restriction of the operator to a subset of nodes defined by a set of relations (*e.g.*, nodes that share a type or specific attribute).
- EXCEPT allows exclusion by the operator of a subset of nodes or relations defined by a set of relations.

The next Section presents the related work. In Section 3, we present the formalism of graph databases and graph rewriting we base our approach on, and the visual conventions used in the AGG rewriting tool. We also provide a running example for the rest of the paper. Section 4 describes the semantics and syntax of the eight basic operators, among which `JoinSet` that processes the keywords WHERE and EXCEPT for restriction and

exclusion in the scope of the other operators. Section 5 presents the description of two privacy procedures through our operators. Finally, Section 6 presents the results of a preliminary experimental evaluation made by applying one of the procedures of Section 5 to a sampling of the Sentiment140 dataset containing information about tweets.

2. Related Work

Private publication of semantic graph. A massive amount of work has focused on privacy in data presented as tables. They have resulted in multiple well-established models, such as k-anonymity [21] and differential privacy [9]. In this article, we consider *semantic databases*, such as RDF, which are *graph databases*. Such databases present the advantage of managing the semantics of the data, which we view as an advantage in the context of anonymization, since it helps the database designer better understand which data is sensitive, and how it can be modified. The aforementioned general anonymization concepts have been translated and applied to graph representations, but mainly in the context of homogeneous graphs such as social or computer networks [24,25].

Only a small batch of work has tackled the publication of semantic data-graphs. Radulovic et al. [17] introduced an anonymization framework for RDF that considers the specific characteristics of RDF specification. The framework aims to protect the privacy of particular entities of interest within RDF graphs. The underlying privacy concept is k-RDF anonymity, where an entity of interest within the graph is indistinguishable from k-1 others with regard to their quasi identifiers. They introduced different add-hoc anonymization operations that can be employed to implement such a model: generalization, suppression, atomization, and perturbation.

K-RDF-Neighborhood anonymity is an approach presented by Heitmann et al. [12]. Their work joins and builds upon other works conducted for homogeneous graphs, specifically by Zhou et al. [26], as well as on heterogeneous graphs, such as RDF, specifically by Radulovic et al. [17]. The central point behind their proposal is that indistinguishability should be guaranteed regarding not only quasi-identifiers but the whole one hop neighborhood of entities of interests. They presented a graph modification algorithm that only deletes edges

In [23], intermediary nodes regroup semantically related values of sensitive attributes via *semantic anatomization*, a process made to regroup attributes with strong semantic similarities. The privacy and utility of the process are then dependant on the creation of relevant or diverse groups. Modifications are expressed as SPARQL update queries.

Delanaux et al. [6] developed a declarative framework for anonymizing RDF graphs by replacing sensitive nodes by blank nodes. They consider a set of utility SPARQL queries and a set of privacy SPARQL queries, and generates operations on a graph so that utility queries can be answered while privacy ones cannot. The generated anonymization operations are in the form of SPARQL update queries of DELETE type (deletion of triples) and DELETE/INSERT (triples update).

While SPARQL is a standard description rather than add-hoc, it is RDF-specific and may exhibit ambiguities [8].

Graph rewriting for database modification. To generalize and abstract consistent updating methods, different works have used formalisms such as tree automata or grammars for XML (see [18] for a survey) or first order logic for graph databases (*e.g.*, [4,10]).

Despite the importance of graphs in databases and ontology representations, the use of formal graph rewriting techniques to model database evolutions is seldom studied. Formal graph rewriting techniques are usually based on *category theory*, an abstract way to deal with different algebraic mathematical structures and the relationships between them. Algebraic approaches of graph rewriting allow a formal yet visual specification of rule-based systems characterizing both the effect of transformations and the contexts in which they may be applied.

Few approaches relying on graph rewriting to formalize ontology evolutions have already been proposed [5,20,16]. They usually focus on formalization but do not provide an implementation. Recently, Chabin et al. [3,2] proposed SETUP, a tool for the management of RDF/S updates. In SETUP, graph rewriting rules formalize atomic updates and guarantee the preservation of RDF/S intrinsic constraints.

3. Background and setting

We consider databases to be modeled as attributed oriented multigraphs. In such models, it is customary for nodes and edges to have properties (among a finite set) and attributes (as words on a signature). In [2], RDF/S databases are modeled as a typed graph with 4 node types and 6 edge types. These types are inherent to RDF and thus the model can not be applied natively to arbitrary graph databases (*e.g.*, neo4j).

We argue that considering a single node type and a single edge type having a single attribute (named *att* and *prop*, respectively) is in fact at least as expressive. Indeed, typing and additional properties can be encoded via special kinds of relation. We believe this model to be able to capture most –if not all– graph database representations.

In what follows, we present a running example of such databases and target privacy-preserving scenarios for its exploitation. Then, we introduce the formalism used to model their transformations.

3.1. Running example and target scenarios

As a running example, we consider a graph database that contains information on travels, both professional and personal. An example of such a database is provided in Fig. 1.

It has nodes for relevant entities, people and travels, whose attributes are identifiers. It also has nodes for literals, *i.e.*, information that would have been stored as a raw value instead of a link to another node in formats like RDF, *e.g.*, last name, first name, and address for people, date and destination for travels. We do not differentiate nodes representing entities or literals.

Its edges describe both relations between entities, *e.g.*, "this person participated in this travel," represented by an edge of attribute ``attends'', but also relations between entities and their information, *e.g.*, "this person's name is in this literal," represented by an edge of attribute ``name''. Typing falls within this second case *e.g.*, "this node is a person" or "this literal is a city," represented by an edge of attribute ``type''.

In the example of Fig. 1, id105 (named Miller) attended travel id207 to Paris for professional reasons.

We give two motivating examples consisting in the application of two privacy mechanisms in this database, whose implementation will be shown to be possible in Section 5. In

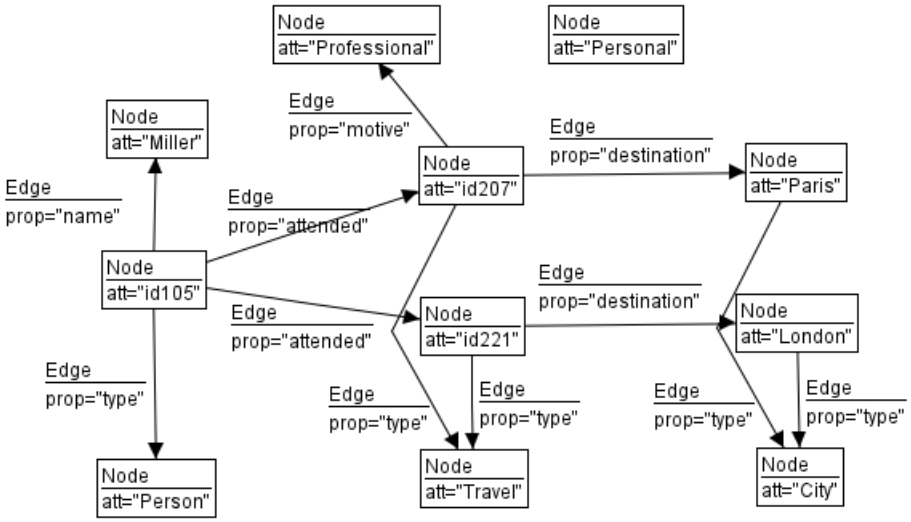


Fig. 1. Running example: instance of a database

these examples, we will note that the described procedures should only apply to specific nodes and relations.

- We want to provide plausible deniability with regard to the relation "destination" between personal travels and cities. To do so, we want to randomize this relation for personal trips exclusively, to preserve privacy with a bias towards correct answers to preserve utility. This corresponds to guaranteeing local differential privacy (LDP [14,7]) on trips with a "motive" edge leading to "personal". More precisely, we want to modify the database such that querying it to output the destination of personal trips would be locally differentially private.
- For professional trips, we want to obfuscate the relation "destination" between travels and cities, as to hide precise dates and frequency of collaboration between the database's company and its collaborators, for instance. This can be done, for instance, by grouping trips in certain cities together (*e.g.*, "Paris," "Bordeaux," "Toulouse" all grouped in the more nebulous group "France") and rerouting the "destination" edges towards those groups rather than a precise value. This would mean that we want to apply anatomization [17,23] where the "destination" attributes of travels with attribute "motive" set to "professional" is considered sensitive.

3.2. Graph Rewriting Rules

We adopt the Single Push Out (SPO) formalism ([15]) to specify rewriting rules as well as one of its extensions to specify additional application conditions and restrict their applicability, *Negative Application Conditions* (NACs) [11], and *Positive Application Conditions* (PACs).

The SPO approach is a simple yet expressive way of fully formalizing graph transformations. Furthermore, it offers an easy-to-understand yet formal graphical view of the

graph transformation. In what follows, we adopt the graphical representation provided by AGG's graphical interface.

The SPO approach is an algebraic approach based on category theory. A rule is defined by two graphs – the Left and Right Hand Side of the rule, denoted by L and R – and a partial morphism m from L to R (*i.e.*, an edge-preserving morphism m from a subgraph of L to R).

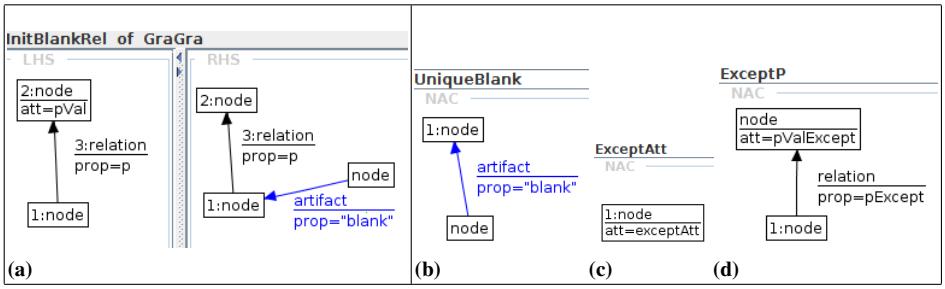


Fig. 2. Identifying nodes having some relation with some value to be replaced by a blank: (a) SPO core and its NACs (b) uniqueness of the blank replacing a node, (c) EXCEPT a specific node, (c) EXCEPT nodes with some relation having some value

Fig. 2a formalizes the SPO core of the first step of the *blank* operator. It identifies the nodes to be replaced and creates a blank for each. Its L is composed of a relation `p` from some node (1) to a node (2) with an attribute `pVal`. R has the same pattern plus a new node that is the source of an artifact edge labeled ``` ``blank'' ``` whose target is node (1). Note that the attributes of two nodes of R are not represented. In general, this can stem from three possibilities: (i) it does not matter, *e.g.*, an unattributed node in L or NAC will match any node; (ii) it can be inferred, *e.g.*, a node in R or NAC has the same attributes as the node it is matched with in R. This is the case for node (2) in the example; (iii) a node in R is created without an attribute value. This is the case for the node representing a blank.

By convention, an attribute value within quotation mark (*e.g.*, ``` ``blank'' ```) is a fixed constant, while *a value noted without quotation mark (e.g., p) is a variable that is either a wildcard matching any value or whose is given as input.*

The partial morphism from L to R is specified in the figure by tagging graph elements - nodes or edges - in its domain and range with a numerical value. An element with value i in L is part of the domain of m and its image by m is the graph element in R with the same value i. For instance, in Fig. 2a, the notation *1:* for the nodes with an unspecified attribute in L and R indicates that they are mapped through m. In the following, we refer to such nodes as `1:node`.

A graph rewriting rule $r = (\mathsf{L}, \mathsf{R}, m)$ is applicable to a graph G iff there exists a total morphism $\tilde{m} : \mathsf{L} \to G$. The result of the application of r to G w.r.t. $\tilde{m}$ is the object of the push-out of the diagram composed by L, R, G, m, and $\tilde{m}$. Informally, the application of r to G with regard to $\tilde{m}$ consists of modifying G by (1) removing the image by $\tilde{m}$ of all elements of L that are not in the domain of m (*i.e.*, removing $\tilde{m}(\mathsf{L} \backslash Dom(m))$); (2) removing

all dangling edges (*i.e.*, deleting all edges that were incident to a node suppressed in step (1)); (3) adding an isomorphic copy of all elements of R that are not mapped through m.

In Fig. 2a, the rule is applicable to any non-artifact edge with the possibility to specify the attribute of the edge (`p`) and/or of its target (`pVal`) as input of the rule. Indeed, such attributes are either wildcard that matches any value or their value may be given as input. For example, with the input (``rdf:type'', ``city''), the rule only maps `1:node` of *type city*. The application of the rule consists of adding an unattributed (blank) node and a ``blank'' artifact edge from the new node to `1:node`, representing the fact that the latter will be replaced by the former.

NACs are well-studied extensions that restrict rule application by forbidding certain patterns in the graph. EXCEPT clauses will mostlty be encoded through NACs. A NAC for a rule r is defined as a constraint graph which is a super-graph of its left-hand side. An SPO rule $r = (L, R, m)$ with NACs is applicable to a graph iff: (i) there exists a total morphism $\tilde{m} : \mathsf{L} \to G$ (this is the classical SPO application condition); (ii) for all NAC N associated with r, there exists *no* total morphism $\bar{m} : N \to G$ whose restriction to L is $\tilde{m}$.

By convention, since NACs are super-graphs of L, unnecessary parts of L are not depicted when illustrating a NAC. Graph elements common to L and NAC are identified by a numerical value, similar to elements mapped by m.

Figures 2b, 2c, and 2d represent NACs associated to the SPO core of Fig. 2a. The first specifies that `1:node` must not be the target of a ``blank'' artifact edge, ensuring that we will not create several blank nodes linking to the same existing node. The second maps `1:node` with a node whose attribute `exceptAtt` is given as input. This forbids the application of the rule to any node with such attribute, akin to an EXCEPT clause excluding a particular node. The third NACs forbids the application of the rule to any `1:node` either source of a relation `pExcept` and/or with value `pValExcept`. Indeed, like variables appearing in L those in NACs can be either wildcards or their value may be given as input. *pExcept* given as ``capital city of'' the rule exclude all `1:node` that are a *capital city*. With `pValExcept` given as ``New York' the rule will exclude all `1:node` that are linked to the a node with the attribute ``New York' (regardless of the relation). Note that here, at least one of the two should be given a input value, or the NAC would match L and the rule would never be applicable. For instance, with

A rewriting procedure –or rule sequence– as we consider it here is a succession of steps. Each step is the application of a rewriting rule as long as the rule applies or a specified number of times. We consider that *when a rule is applicable w.r.t. several morphisms, it chooses one uniformly at random*.

4. Our Language

Our language's goal is to support the specification of privacy procedures and mechanisms that will apply operators (*e.g.*, randomization) on some particular sets of nodes and edges.

In its general form, we would expect each of our instructions to be the combination of an operator, a set of sources (or subject), a set of destinations (or objects), and potentially additional operator- or procedure-specific requirements (*e.g.*, new targets for randomization, or lists of identifiers and sensitive attributes). To this end, we aim to provide a small

number of simple operators, as well as a a way to descrinbe the sets of nodes and edges on which they will apply.

This section introduces the language we build and its design choices: procedures composed of elementary operators applied to simple cases, and the special operator `JoinSet` that allow it to preserve its expressiveness.

4.1. Design principle

Simple operators Our main design choice is to keep the basis of our language as simple as possible. We propose a set of eight operators, each described by a single graph rewriting rule, with an easy-to-match pattern. They allow to create and delete nodes, copy, modify, cut or merge edges, randomize the sources or destination of a relation. Al these operators work on very basic patterns, and the main role of the eighth operator, `JoinSet`, is to reduce complex application cases to the simple cases handled by our operators.

Node and Edge identification To limit the use of an operator to specific nodes, our language needs a way to define and manipulate sets of node and edges.

As discussed in the example of 3, these sets can be defined as types, can exclude types, or can require the presence or absence of a certain property at a certain value. It is possible to specify these kinds of restrictions directly in graph rewriting rules, using L, PACs, and NACs. However, this would have the important drawback of necessitating numerous versions of the same rule and/or complicating the pattern matching phase. We therefore believe a pre- and post-processing procedure to construct temporary add-hoc sets to be necessary.

To identify *nodes*, our language considers either the label of the node (*e.g.*, the node.s labeled "id207"), or the existence of labels with a given edge towards nodes of a certain label (*e.g.*, the node.s with edges "type" heading to nodes labeled "City"). As a shorthand, we sometimes say that the nodes with a label S and an edge labeled p towards a node labeled O as "matching (S, p, O)". Such a triplet (S, p, O) can be called a **set**, and identified with a single letter $\mathcal{X}$.

We note that we can specify two special value in these sets:

- $\star$ corresponds to any label we want
- `null` corresponds to no expectation, *i.e.*, not only any label but the existence is not necessary either

For instance, $(\text{City}, null, null)$ would match any node whose label is "City", whether they have an outgoing edge or not, which includes one node in our example. $(\text{City}, *, *)$ would match any node whose label is "City" with any outgoing edge at all, which includes no node of our example. Note that `null` differs from $\star$ only when both p and O are equal to `null`. Indeed, $(\text{City}, *, null)$ would not constrain O but impose the existence of some p, and thus of some O.

To identify *edges*, our language considers their labels and identifiers of their source or target (*e.g.*, all edges labeled "destination" from a node matching $(*, \text{"type"}, \text{"Travel"})$ to a node matching $(*, \text{"type"}, \text{"City"})$).

Pre-processing using `JoinSet` To reduce a wide variety of operators' scopes and targets to the basic case, we only define operators on very simple cases, *i.e.*, by considering simple node sets, and when applicable the edges that link them. For this to not hinder the

expressiveness of our language, we need a way to create or populate sets that correspond to nodes matching intricate conditions. For this, we use the following instruction:
`JoinSet` (x, X) `Where` $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ `Except` $\{\mathcal{Y}_1, \ldots, \mathcal{Y}_m\}$

This primitive matches nodes that match every $\mathcal{X}_i$ in the `Where` section and match none of the $\mathcal{Y}_i$ in the `Except` section. When it does, it create from it an edge labeled x towards a node labeled X.

Graph rewriting primitives: We present here the few primitives, formalized and implemented directly using AGG, that create or delete nodes, copy, divide, or merge edges, and randomize specific relations[1].

- `NewNode`(X) creates a new node labeled X
- `DeleteNode`$(\mathcal{X})$ matches all nodes matching $\mathcal{X}$ and deletes them.
- `JoinSet` (x, X) `Where` $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ `Except` $\{\mathcal{Y}_1, \ldots, \mathcal{Y}_m\}$ (explained above)
- `EdgeCopy`$(\mathcal{S}, p, \mathcal{O}, p')$ matches couples of source nodes matching $\mathcal{S}$ and destination nodes matching $\mathcal{O}$, where there is an edge from the source to the destination labeled p.
  When a match is found, creates an edge labeled p' from the source to the destination.
- `EdgeReverse`$(\mathcal{S}, p, \mathcal{O}, p')$ matches couples of source nodes matching $\mathcal{S}$ and destination nodes matching $\mathcal{O}$, where there is an edge from the source to the destination labeled p.
  When a match is found, creates an edge labeled p' from the destination to the source.
- `EdgeCut`$(\mathcal{S}, p, \mathcal{O}, p_I, M, p_O)$ matches pairs of source nodes matching $\mathcal{S}$ and destination nodes matching $\mathcal{O}$, where there is an edge from the source to the destination labeled p.
  When a match is found, creates a new intermediary node of label M. Then, it creates an edge labeled p_I from the source to the intermediary, an edge labeled p_O from the intermediary to the destination, and finally it deletes the edge p from the source to the destination.
- `EdgeChord`$(\mathcal{S}, p_I, \mathcal{M}, p_O, \mathcal{O}, p)$ is the converse of `EdgeCut`, and matches triplets of source nodes matching $\mathcal{S}$, intermediary nodes matching $\mathcal{M}$ and destination nodes matching $\mathcal{O}$ where there is an edge from the source to the intermediary labeled p_I and there is an edge from the intermediary to the destination labeled p_O.
  When a match is found, creates an edge labeled p from the source to the destination.
- `RandomTarget`$(\mathcal{S}, p, \mathcal{O}, \mathcal{T})$ matches all edges labeled p between a source matching $\mathcal{S}$ and a destination matching $\mathcal{O}$, and it reroutes this edge by picking a new target uniformly among nodes matching $\mathcal{T}$.

Procedures: A procedure is a sequence of instructions, executed in order. They can be as simple as intermediary operators, or complex enough to describe graph manipulations guaranteeing certain types of privacy. For instance, a procedure `DeleteEdge`$(\mathcal{S}, p, \mathcal{O})$ meant to delete all edges labeled p between a source matching $\mathcal{S}$ and a destination matching $\mathcal{O}$ would be:

1: `EdgeCut`$(\mathcal{S}, p, \mathcal{O}, p_I, "ToBeDeleted", p_O)$
2: `DeleteNode`$("ToBeDeleted", null, null)$

[1] Rules specifications are available at `https://github.com/ceichler/granon/blob/master/anonOperator.ggx`

This procedure first cuts every occurrence of p between $\mathcal{S}$ and $\mathcal{O}$ in two with a new node labeled "ToBeDeleted" (line 1). Then, it immediately deletes all those "ToBeDeleted" nodes (line 2), thus erasing the original edges.

5. Privacy Procedures

This section introduces procedures that provide privacy guarantees: Local Differential Privacy [7] and anatomization [17]. These procedures are expressed as combinations of our primitives.

5.1. Local Differential Privacy

The first proposed procedure support randomization to achieve local differential privacy (LDP) as defined and used in [14,7]. Assuming a query outputing the value.s of a property for a particular node or a set of nodes, we wish to modify the graph to make the query satisfy LDP.

Definition 1 (Local DP (Duchi *et al.*) [7]). *Let χ be a set of possible values and Y the set of noisy values. A mechanism $\mathcal{M}$ is ε-locally differentially private (ε-LDP) if for all $x, x' \in \chi^2$ and for all $y \in Y$ we have*

$$Pr[\mathcal{M}(x) = y] \leq e^{\varepsilon} \times Pr[\mathcal{M}(x') = y]$$

LDP-mechanisms outputing a value in Y achieve optimal utility for a given ε by giving an answer randomly drawn from a staircase distribution over Y, with the most probable value being the *real* value -whose probability depends solely on $|Y|$ and ε- and all other values being equiprobable [13].

To achieve ε-LDP for a set R of relations, our random operator should therefore transform each $(s,t) \in R$ into a relation $(s,t') \in R'$ under the following specification:

$$P(t') = \begin{cases} 0 & t' \notin T \\ \frac{K}{|T|-1+K} & t' = t \\ \frac{1}{|T|-1+K} & t' \in T \wedge t' \neq t \end{cases}$$

with K an integer approximation of e^{ε}, $K = \lfloor e^{\varepsilon} \rfloor$.

To do so, we pick a new target at random, but, to obtain a staircase distribution from a uniform distribution –used to choose the morphism with regard to which the transformation rule is applied–, we skew the odds by creating $K-1$ dummies. Picking a dummy as a target should ultimately result as giving the true answer to recreate a staircase distribution. The procedure, defined in Alg. 1, and detailed thereafter, has the following arguments:

- $\mathcal{S} = (X_S, s, S)$ to match the sources of the relation to randomize
- p to match the edges of the relation to randomize
- $\mathcal{O} = (X_O, o, O)$ to match the destinations of the relation to randomize
- K the factor by which correct answers are more likely than other values

Algorithm 1 Procedure LDP(S,p,O,K)

1: **for** i from 1 to $K-1$ **do**
2: `NewNode`$(Dummy)$
3: **end for**
4: `JoinSet`(o,O)`Where`$\{(Dummy,*,*)\}$ `Except`$\{\}$
5: `EdgeCut`$(\mathcal{S},p,\mathcal{O},p_I,"Intermediary",p_O)$
6: `EdgeCopy`$((*,null,null),p_O,(*,null,null),p_N)$
7: `RandomTarget`$((*,null,null),p_N,(X_O,o,O),(*,o,O))$
8: `EdgeChord`$(\mathcal{S},p_I,(*,p_N,Dummy),p_O,\mathcal{O},p)$
9: `EdgeChord`$(\mathcal{S},p_I,(*,null,null),p_N,\mathcal{O},p)$
10: `DeleteNode`$(Dummy,*,*)$
11: `DeleteNode`$(Intermediary,*,*)$

Recurring example: We illustrate the steps of this procedure in a recurring example presented in Figures 3 to 8. We present a $\ln(4)$-LDP-providing randomization of persons attending travels, *i.e.*, the procedure LDP((type,Person),attended,(type,Travel),3).

In this example, nodes and edges in black are definitive data that started or are meant to remain in the graph. Nodes and edges in blue are temporary nodes that are only used as intermediary elements in the rewriting process (e.g. the Dummy nodes). Nodes and edges in bold/thick have recently been created (e.g. the "attended" edge in Figure 6). Nodes and edges that are dashed with reduced opacity have recently been deleted (e.g. the "Dummy" nodes in Figure 8).

Initialization: We start by creating a bias for the correct value, by creating $K-1$ dummy nodes thanks to the `NewNode` operator This rule is to be repeated as many times as required to obtain ε-LDP, then they are matched to $(*,o,O)$ with `JoinSet` (lines 1 and 2). Creating two dummies makes the truth three times likelier, and suits $\varepsilon \geq \ln(3)$. If we create 0 dummies, then we are 0-LDP as the edges' targets will be uniformly randomized.

This step is illustrated in Figure 3 with the creation of 3 dummy nodes.

Edge Cut: We want to use `RandomTarget` to reroute edges p towards a new target matching $(*,o,O)$ at random. However, picking a dummy as a new target means we want to keep the original target. This means that instead of rerouting the edges p directly, we would like to create "forks" that lead both to the original and new targets. Since hyper-edges are not objects of our graphs, we use `EdgeCut` and `EdgeCopy` to emulate this behavior (lines 3 and 4):

- First, we cut the edges p from $\mathcal{S}$ to $\mathcal{O}$ into p_I and p_O edges coming from a middle node labeled "Intermediary"
- Then, we create a copy of p_O edges labeled p_N

This is pictured in Figure 4.

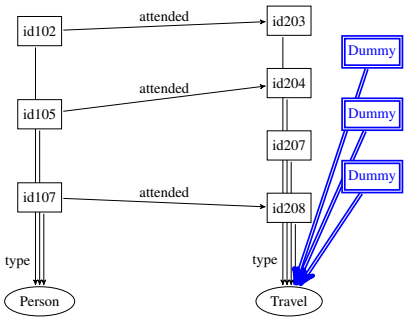


Fig. 3. Create dummy travel nodes

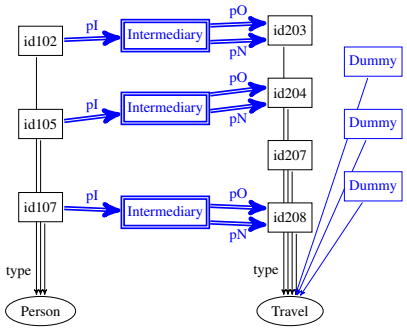


Fig. 4. Split "attended" edges into pI, pO, pN

Randomizing: We then use `RandomTarget` (line 5) toto redirect the edges p_N towards any target matching $(*, o, O) = \mathcal{O}$, with uniform probability. We have one chance to pick the original node, and $K - 1$ chances to pick a dummy. This means we are K times more likely to pick one of these nodes than any other real node matching $\mathcal{O}$. This is pictured in Figure 5, where node matching *(*,type,Travel)* are depicted in red.

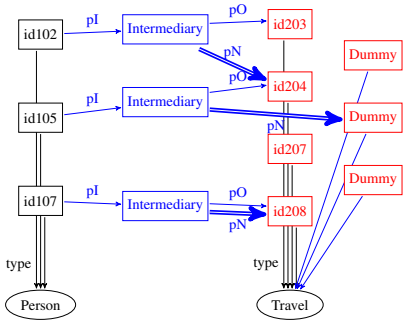


Fig. 5. Randomize the target of pN edges

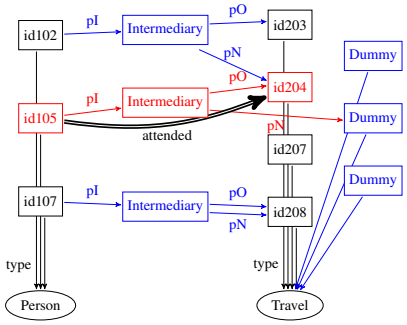


Fig. 6. Chord rerouting: pN targets a Dummy

Rerouting: We start by dealing with the case where p_N was rerouted to a dummy (line 6). In that case, the intermediary node matches $(*, p_N, Dummy)$, and we want to keep the old target of p, which means the current target of p_O. We say that if there is a chord from a node matching $\mathcal{S}$ to an intermediary node matching $(*, p_N, Dummy)$ to a destination node matching $\mathcal{O}$ with edges p_I, p_O, then we create the edge p from the source to the old destination. This is pictured in Figure 6, the matched chord being depicted in red.
We then deal with the case where p_N was rerouted to a real target (line 7). We say that if there is a chord from a node matching $\mathcal{S}$ to any intermediary node to a destination node matching $(*, o, O)$ with edges p_I, p_N, then we create the edge p from the source to the new destination. This is pictured in Figure 7, the matched chords being depicted in red.
We note that this also creates unnecessary edges from sources to dummies. This is not a problem, as dummies will be deleted shortly, as pictured in Figure 8.

Termination: Since dummies and intermediary nodes are artifacts we created rather than real nodes of the graph, we end the procedure by using `DeleteNode` to delete all nodes of label "Dummy" and "Intermediary" (lines 8 and 9). This also deletes all edges p_I, p_O, p_N linked to intermediary nodes, as well as edges p unduly targeting dummies generated during rerouting.

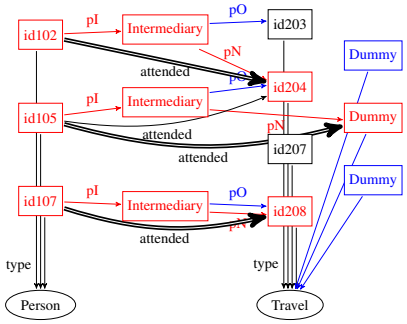


Fig. 7. Chord rerouting: every path pI, pN generates a new "attended" edge

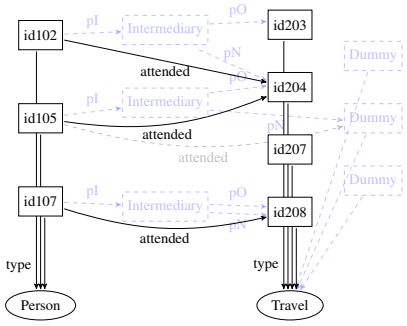


Fig. 8. Cleanup

5.2. Anonymization through Anatomization

In this section, we show that sorting sensitive attributes in groups to prevent inferences can be naturally expressed in our elementary operators.

We consider that some attributes of our graphs are sensitive, as described in [17,23]. These papers describe an anonymization process that involves the deletion of explicit identifiers (*e.g.* first and last name), and the separation of nodes with quasi-identifiers (*e.g.* date of birth, zip code) to a set of sensitive attributes (*e.g.* religion, sexual orientation). To cut links between quasi-identifiers and the sensitive attributes, the possible values of sensitive attributes are put into groups through a process called semantic anatomization [23]. Their approach is qualified as semantic as it concentrates on semantic-aware grouping of sensitive attributes. Furthermore, it retains the correlation between entity quasi-identifiers and semantically related sensitive values.

While the creation of these groups itself is a potentially involved and complex process that goes beyond the scope of graph rewriting, the redirection of links from identifiers to sensitive attributes is possible if the groups are provided.

We consider that our graph contains explicit identifiers (edge labels $e_1, \ldots, e_n$), quasi-identifiers (edge labels $q_1, \ldots, q_m$), sensitive attributes (edge labels $p_1, \ldots, p_k$), and group nodes that aggregate attribute values. A value X is part of a group if nodes of label X have an edge of label "inGroup" pointed at this group's node. The anonymization is made as follows:

Recurring example: The algorithm is detailed thereafter and illustrated through a toy example where the destinations of travels are considered sensitive $(p_i) = (destinatiation)$, the identifiers are names $(e_i) = (name)$, and the quasi-identifiers are types $(q_i) = (types)$.

More specifically, we illustrate the steps of sensitive attributes redirection in a recurring example presented in Figures 9 to 14.

Erasing explicit identifiers: Line 1 deletes the edges towards sensitive attributes. The values are preserved, but the links are cut for anonymization. In the example, this would lead to the deletion of every edge whose attribute is "name".

Flagging quasi identifiers: Lines 2 and 3 create a node QI and relate every node with at least one quasi identifiers to QI through an edge $hasQI$. In the example, all nodes with a *type* are the source of an edge $hasQI$ whose target îs QI.

Algorithm 2 Procedure Anat$((e_i)_{1\leq i\leq n}, (q_i)_{1\leq i\leq m}, (p_i)_{1\leq i\leq k}, inGroup)$

1: For all e_i do `DeleteEdge`$((*, null, null), e_i, (*, null, null))$
2: `NewNode`(QI)
3: For all q_j do `JoinSet`$(hasQI, QI)$ `Where`$\{(*, q_j, *)\}$ `Except`$\{\}$
4: For all p_i do `EdgeChord`$((*, hasQI, QI), p_i, (*, null, null), inGroup, (*, null, null), p'_i)$

5: For all p_i do `EdgeCut`$((*, hasQI, QI), p_i, (*, null, null), p^i_I, "Intermediary", p^i_O)$
6: For all p_i do
`EdgeChord`$(("Intermediary", null, null), p^i_O, (*, null, null), inGroup, (*, null, null), p''_i)$

7: For all p_i do `EdgeReverse`$(("Intermediary", null, null), p''_i, (*, null, null), p''_i)$
8: For all p_i do
`EdgeChord`$((*, null, null), p''_i, ("Intermediary", null, null), p^i_O, (*, null, null), hasOne)$

9: For all p_i do `EdgeCopy`$((*, null, null), p'_i, (*, null, null), p_i)$
10: For all p_i do `EdgeCut`$((*, hasQI, QI), p'_i, (*, null, null), p^i_I, "Intermediary", p^i_O)$
11: `DeleteNode`$("Intermediary", null, null)$
12: `DeleteNode`$(QI, null, null)$

Redirecting sensitive attributes: Line 4 matches cases where a node with quasi-identifiers has an edge p_i pointing towards the value of a sensitive attribute, which itself has an edge $inGroup$ designating a group of the semantic anatomization. These matches are depicted in red on Fig. 9.
When such a match is detected, an edge p'_i is directly traced from the initial node to the group. The label used is p'_i instead of p_i to avoid undue deletions of redirected edges in line 5.
Should we want to preserve those sensitive but ungrouped edges, we would only split edges pointing towards grouped values by replacing line 5 with:

For all p_i
do `EdgeCut`$((*, hasQI, QI), p_i, (*, inGroup, null), p_{I,i}, "Intermediary", p_{O,i})$

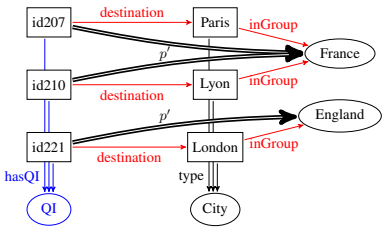


Fig. 9. Short-circuit sensitive values

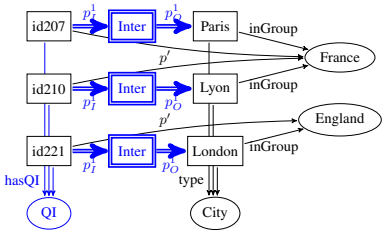


Fig. 10. Split sensitive edges into p_{I_1}, p_{O_1}

Keep track of sensitive attribute values: Lines 5 through 8 aim to keep track of the cardinalities of each sensitive value in the anatomization groups, represented by the edges p_i we delete.
Line 5 creates an intermediary node between the source and the value of its sensitive attribute, as depicted in Fig. 10. Line 6 links those intermediary nodes with the group of their sensitive attributes with an edge labeled p_i'', illustrated in Fig. 11. The matched chords are depicted in red.

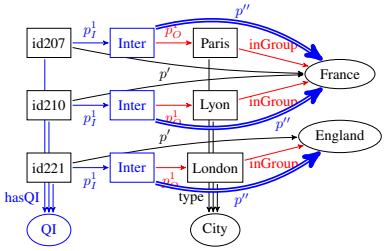


Fig. 11. Link intermediaries with groups

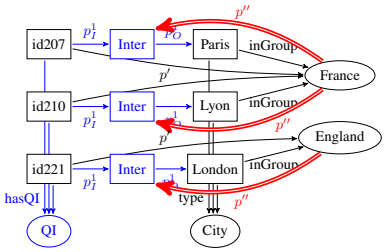


Fig. 12. Reverse from group to intermediaries
This edge is inverted in line 7, then used in line 8 to create a direct link between the group and the sensitive attribute, as shown if Fig. 12 (where the inverted p'' are depicted in red) and 13 (where the p'', p_O^i chords are depicted in red), respectively.

Bringing back p_i: Line 9 creates a copy p_i of each p_i' linking a node with the group its sensitive value is in. Then line 10 cut the p_i' edges to prepare them for deletion.

Cleanup: Finally, lines 11 and 11 erase the nodes we created in previous steps. This notably includes the last trace of edges p_i that could not have been redirected –*i.e.*, those that did not belong to an anatomization group– for which the information is gone for good. This last step is depicted in Fig. 14.

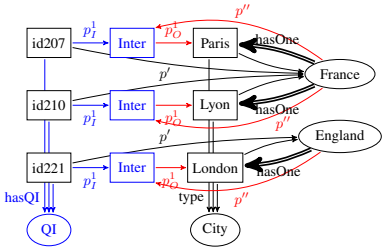


Fig. 13. Chord from groups to values

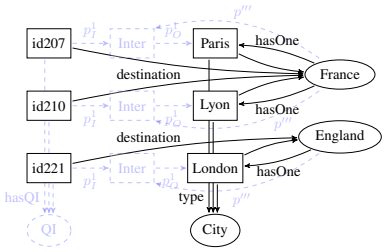


Fig. 14. Cleanup

6. Experimental evaluation

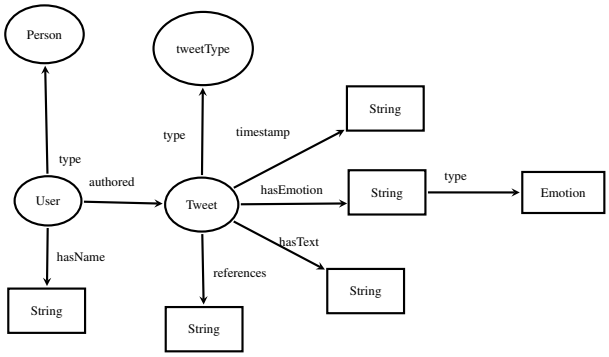


Fig. 15. Schema for Sentiment140

This section investigates the feasibility of the approach and provides a preliminary experimental evaluation by applying two privacy procedures to a real dataset. We implemented

the formal specification of the operators using AGG – The Attributed Graph Grammar System ([22]), one of the most mature development environment supporting the definition and application of typed graph rewriting systems ([19]). Using AGG's Java API, we implemented Granon, a tool in Java[2] to handle their management and on-the-fly modification (*e.g.* the definitions of multiple NACs and PACs in the *JoinSet* operator). Granon also supports the procedures defined herein. Experiments are conducted in a single-thread on an Intel Xeon Gold 5215 2.5GHz with 64 GB RAM and a 20Go JVM heap size.

Dataset(s): The experiments are conducted on a sampling of the Sentiment140 dataset composed of 1.6 million tweets[3], which we have parsed to load as a knowledge graph conforming to the schema shown in Fig. 15. The dataset is composed of *tweets* authored by a named user. Tweets each have a timestamp, a full text, and an emotion (positive, neutral, or negative). They may reference users' name. The users, tweets, and emotions are all typed.
The most important factor regarding runtime is the number of tweets and the size of the graph. Therefore, we apply our experiments on the graphs resulting from the parsing of the t first tweets of the datasets, with $t = 200$, 400, 600, 800, 1000, 1200, and 2000.

Nature of the experiments: To investigate the scalability of the proposal, we apply an instantiation of both our privacy schemes:

1) *Procedure LDP((*,"type", "tweetType"), "hasEmotion", (*, "type", "Emotion"), ln(2))* as described in Sec. 5.1. We arbitrarily consider the emotion of a tweet to be the sensitive value and use $\varepsilon = ln(2)$, the value of $epsilon$ having a negligible influence on the runtime.
2) *Procedure Anat((hasText), (references), (timestamp), inGroup)* described in 5.2. We consider texts to be identifiers, timestamps to be sensitive and references to be quasi-identifiers. Therefore, the procedure will delete all "hasText" edges and generalize the timestamp of any tweet that references someone. After considering the dataset, we construct anatomization groups representing a one minute timewindow: for instance a timestamp with the value "Mon Apr 06 22:20:19 PDT 2009" belongs to the group "Mon Apr 06 22:20 PDT 2009". Groups are constructed while parsing the experimental datasets, resulting in slightly bigger graphs for the same t.

Experimental results: Average and median execution times (in ms) over 50 runs of the procedures are reported in Fig. 16 for the various t. The input graphs' sizes (*i.e.*, their number of vertices #V and edges #E) constructed by parsing the first t tweets of the dataset are reported in Tab. 1. The size of the graph is linear in *t*, which is consistent with the schema. The average and median runtimes for the LDP procedure are overlapping as the distribution has a very low standard deviation. The distribution of runtimes for the Anatomization procedure consistently comports high outliers for every t.

Experimental interpretation: As expected, the LDP procedure is more time consuming than the Anatomization procedure. For $t = 1000$, the median for LDP and Anatomization

[2] available at `https://github.com/ceichler/granon`
[3] `https://www.kaggle.com/kazanova/sentiment140`

Table 1. Size of the graphs resulting from parsing the t first tweets of the dataset

	LDP		Anatomization	
t	#V	#E	#V	#E
100	531	743	538	843
200	1042	1465	1054	1665
300	1558	2201	1576	2501
400	2081	2930	2105	3330
500	2605	3661	2636	4161
600	3113	4386	3150	4986
700	3623	5115	3665	5815
800	4147	5889	4195	6659
900	4648	6576	4704	7476
1000	5170	7310	5232	8310
2000			10393	16531
3000			15463	24625

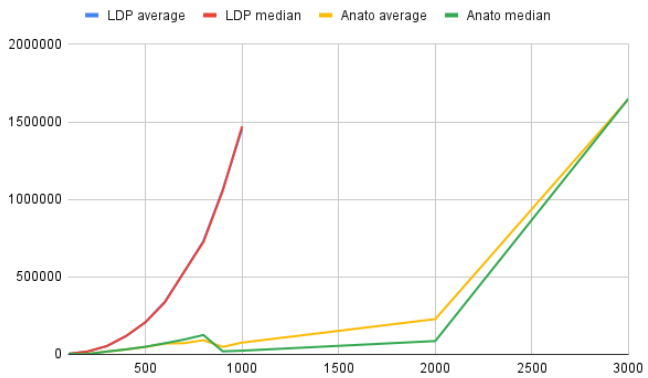


Fig. 16. Experimental results runtime (ms) with various t

are 1,47 * 10^6 ms and 2,4 * 10^4ms, respectively. Furthermore, while the asymptotic complexity of the latter is roughly quadratic for $t \geq 2000$, the former is over-quadratic. The median execution time of the LDP procedure roughly triples from $t = 400$ to $t = 500$ (1,2 * 10^5 ms to 3,3 * 10^5ms) when the number of nodes is multiplied by 1,25. The average execution time of the Anatomization procedure is multiplied by 547 regarding t = 100 and t =3000, with 28,7 times the number of nodes.

This can be explained by the randomization inherent to LDP. Indeed, randomizing the target of a relation requires restarting the matching process from scratch, without benefiting from optimizations (e.g. smart backtracking). Otherwise, the first match would influence all others. Therefore, randomizing the sensitive value of n items requires -in the randomizing step- running n search for graph homomorphisms, which is itself super-linear in the size of the graph.

A 30 minute run would allow to execute the LDP (resp. anatomization) scheme on a graph slightly bigger than #V = 5170, #E = 7310 (resp. #V = 15463, #E = 24625). While such a runtime is very much reasonable when considering the sanitization of a database prior to its release as the process is expected to be run infrequently, the asymptotic complexities discourage the application of the proposed graph-rewriting based techniques to big graphs.

7. Conclusion

This paper is a first effort toward bridging the gap between rigorous graph rewriting and graph sanitization. We propose a basic yet expressive set of atomic operators, formalized as simple Single Push-Out graph rewriting rules, providing a generic expressive rigorous definition that can be parametrized with node and set identifiers. We show how they can be used to express privacy-preserving graph databases publication mechanisms, namely local differential privacy and anatomization. These operators and procedures have been subject to an open-source implementation[4]. This work stands as a proof of concept and a first step towards a graph rewriting based approach to graph database sanitization. A lot of considerations and work remain open on the topic.

Firstly, while the current implementation would reasonably allow the sanitization of small to medium graphs, its scalability remains limited due to high asymptotic complexity. A first effort would be required to reduce multiplicative constants, for example, by expanding operators to reduce the number of transformations, e.g. implementing an operator for edge suppression or edge modification rather than relying on two operators to conduct the operation; 2) or even the asymptotic complexity of our procedures, for example by considering a subgraph for the transformation, as procedures work on local properties.

Another natural outlook would be to expand the tool to encompass other privacy protocols (e.g. [6]). This might require the use of other operators, that we hope share the simplicity and versatility of our current set.

Furthermore, our language and operators are designed to work on simple relations between two sets of nodes. In real world usecases, most queries are joins of several relations (or to speak in graph terms, path queries on more than one edge). For a sanitization mechanism to preserve good qualities on such requests while still providing privacy guarantees, it is likely that we will need operators adapted to the preservation of invariants on composite paths. It is possible, but not yet shown, that some such operators can be built as compositions of one-relation operators.

Ultimately, as a more general goal, we aim to expand our current tool to allow a user to specify semantic and privacy constraints. Such a tool would compile those requests as sequences of our operators and offer to perform the resulting graph transformations.

Acknowledgments. This work is part of the SENDUP project, supported by the French National Research Agency, under grant ANR-18-CE23-0010.

[4] available at `https://github.com/ceichler/granon`

References

1. Boiret, A., Eichler, C., Nguyen, B.: Privacy operators for semantic graph databases as graph rewriting. In: Chiusano, S., Cerquitelli, T., Wrembel, R., Nørvåg, K., Catania, B., Vargas-Solar, G., Zumpano, E. (eds.) New Trends in Database and Information Systems. pp. 366–377. Springer International Publishing, Cham (2022)
2. Chabin, J., Eichler, C., Ferrari, M.H., Hiot, N.: Graph rewriting rules for RDF database evolution: optimizing side-effect processing. Int. J. Web Inf. Syst. **17**(6), 622–644 (2021)
3. Chabin, J., Eichler, C., Halfeld-Ferrari, M., Hiot, N.: Graph rewriting rules for rdf database evolution management. In: Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications & Services. p. 134–143. ACM (2020)
4. Chabin, J., Halfeld Ferrari, M., Laurent, D.: Consistent updating of databases with marked nulls. Knowledge and Information Systems (2019)
5. De Leenheer, P., Mens, T.: Using graph transformation to support collaborative ontology evolution. In: Schürr, A., Nagl, M., Zündorf, A. (eds.) Applications of Graph Transformations with Industrial Relevance. pp. 44–58. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
6. Delanaux, R., Bonifati, A., Rousset, M.C., Thion, R.: Query-based linked data anonymization. In: International Semantic Web Conference. pp. 530–546. Springer (2018)
7. Duchi, J.C., Jordan, M.I., Wainwright, M.J.: Local privacy and statistical minimax rates. In: 54th Annual IEEE Symposium on Foundations of Computer Science. pp. 429–438. IEEE Computer Society (2013)
8. Duval, D., Echahed, R., Prost, F.: An algebraic graph transformation approach for RDF and SPARQL. In: Hoffmann, B., Minas, M. (eds.) Proceedings of the Eleventh International Workshop on Graph Computation Models, GCM@STAF 2020, Online-Workshop, 24th June 2020. EPTCS, vol. 330, pp. 55–70 (2020). https://doi.org/10.4204/EPTCS.330.4, https://doi.org/10.4204/EPTCS.330.4
9. Dwork, C.: Differential privacy. In: Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II. pp. 1–12 (2006)
10. Flouris, G., Konstantinidis, G., Antoniou, G., Christophides, V.: Formal foundations for RDF/S KB evolution. Knowl. Inf. Syst. **35**(1), 153–191 (2013)
11. Habel, A., Heckel, R., Taentzer, G.: Graph grammars with negative application conditions. Fundam. Inf. **26**(3,4), 287–313 (Dec 1996)
12. Heitmann, B., Hermsen, F., Decker, S.: k-rdf-neighbourhood anonymity: Combining structural and attribute-based anonymisation for linked data. PrivOn@ ISWC **1951** (2017)
13. Kairouz, P., Oh, S., Viswanath, P.: Extremal mechanisms for local differential privacy. Journal of Machine Learning Research **17**(17) (2016), http://jmlr.org/papers/v17/15-135.html
14. Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A.D.: What can we learn privately? SIAM J. Comput. **40**(3), 793–826 (2011)
15. Löwe, M.: Algebraic approach to single-pushout graph transformation. Theoretical Computer Science **109**(1–2), 181 –224 (1993)
16. Mahfoudh, M., Forestier, G., Thiry, L., Hassenforder, M.: Algebraic graph transformations for formalizing ontology changes and evolving ontologies. Knowledge-Based Systems **73**, 212 – 226 (2015)
17. Radulovic, F., García Castro, R., Gómez-Pérez, A.: Towards the anonymisation of rdf data (2015)
18. Schwentick, T.: Automata for XML - A survey. J. Comput. Syst. Sci. **73**(3), 289–315 (2007)
19. Segura, S., Benavides, D., Ruiz-Cortés, A., Trinidad, P.: Automated Merging of Feature Models Using Graph Transformations, pp. 489–505. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)

20. Shaban-Nejad, A., Haarslev, V.: Managing changes in distributed biomedical ontologies using hierarchical distributed graph transformation. Intern. Journal of Data Mining and Bioinformatics **11**(1), 53–83 (2015)
21. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(5), 557–570 (2002)
22. Taentzer, G.: Agg: A graph transformation environment for modeling and validation of software. In: AGTIVE (2003)
23. Thouvenot, M., Curé, O., Calvez, P.: Knowledge graph anonymization using semantic anatomization. In: 2020 IEEE International Conference on Big Data (Big Data). pp. 4065–4074. IEEE (2020)
24. Wu, X., Ying, X., Liu, K., Chen, L.: A Survey of Privacy-Preservation of Graphs and Social Networks, pp. 421–453. Springer US (2010)
25. Zheleva, E., Getoor, L.: Privacy in social networks: A survey. In: Social network data analytics, pp. 277–306. Springer (2011)
26. Zhou, B., Pei, J., Luk, W.: A brief survey on anonymization techniques for privacy preserving publishing of social network data. SIGKDD Explorations **10**(2), 12–22 (2008). https://doi.org/10.1145/1540276.1540279, http://doi.acm.org/10.1145/1540276.1540279

Adrien Boiret, Associate Professor at INSA-CVL, LIFO Lab, Systems and Data Security research team since 2022. His current focus is in data privacy for graph databases, especially as they contain semantic information. His previous work was on formal languages for semistructured data.

Cédric Eichler obtained his Ph.D. from Toulouse III in 2015. Since 2016 he is associate professor at INSA Centre Val de Loire where he developed his interest for both technical and legal aspects of privacy. He was Data Protection Officer of INSA Centre Val de Loire from 2018 to 2022 and head of the "Security of Data and Systems" research team from 2017 to 2022. His scientific contributions lie at the intersection of graphs and privacy: from private graph computations to the sanitization of graph databases with underlying semantic.

Benjamin Nguyen is a Professor at INSA Centre Val de Loire since 2014. He received his Ph.D. from Université de Paris-Sud in 2003. His research interests focus on privacy preserving data management techniques, with currently a special focus on anonymization and reidentification techniques. He was head of the Laboratoire d'Informatique Fondamentale d'Orléans (LIFO, EA 4022) from 2016 to 2022, and member of the executive board of INSA Centre Val de Loire from 2015 to 2019.

Sara Taki, Research Engineer at INSA Centre Val de Loire, LIFO Lab, Systems and Data Security research team. She earned her Ph.D. in Computer Science from INSA Centre Val de Loire in 2023. Her research focuses on data privacy within graph databases, in particular differential privacy over RDF graphs. She holds an M.Sc degree in Information Systems and Data Intelligence from Lebanese University.

Received: April 20, 2023; Accepted: May 05, 2024.

Computer Science and Information Systems 21(3):1055–1075 https://doi.org/10.2298/CSIS230530027G

A Comprehensive Review of the Data and Knowledge Graphs Approaches in Bioinformatics*

Ylenia Galluzzo

Department of Engineering, University of Palermo, Italy
ylenia.galluzzo01@unipa.it

Abstract. The scientific community is currently showing strong interest in constructing knowledge graphs from heterogeneous domains (genomic, pharmaceutical, clinical etc.). The main goal here is to support researchers in gaining an immediate overview of the biomedical and clinical data that can be utilized to construct and extend KGs. A in-depth overview of the available biomedical data and the latest applications of knowledge graphs, from the biological to the clinical context, is provided showing the most recent methods of representing biomedical knowledge with embeddings (KGEs). Furthermore, this review, differentiates biomedical databases based on their construction process (whether manually curated by experts or not), aiming to offer a detailed overview and guide researchers in selecting the appropriate database for their research considering to the specific project needs, available resources, and data complexity. In conclusion, the review highlights current challenges: integration of different knowledge graphs and the interpretability of predictions of new relations.

Keywords: Biomedical Knowledge Graph, Knowledge Graph Embeddings, Text Mining, Graph Neural Network

1. Introduction

Knowledge graphs are an area of great interest in both academia and industry, because they facilitate information extraction (facts and hypotheses) by to well-defined interconnections between relevant entities (abstract and concrete) within a given domain. In addiction, they are equally interesting for understanding how to form new relationships through the use of data semantics and linkages. Originally, knowledge graphs are represented as the knowledge base graphs in the Resource Description Framework (RDF). Information (Resource, a Property and a Property value) is represented through assertions forming SPO triples (subject, predicate, object) which express direct and complex relationships between different resources [63]. Knowledge graphs can also be described as an ontology. An ontology is a data model that represents knowledge about a specific through sets of relations among concepts within a domain and instances of objects representing the topic. The Web Ontology Language (OWL) serves as a markup language for expressing ontologies. RDF and OWL have become crucial standards within the Semantic Web. In 2012, Google popularized knowledge graphs with the introduction of its "Google Knowledge

* This is an extended version of a conference paper presented in a workshop K-GALS organized in conjunction with the conference ADBIS 2022.

Graph" [22]. This system uses Knowledge Vault, which combines probabilistic knowledge extracted from Web content with prior knowledge derived from existing knowledge repositories enabling users to receive relevant information based on the search queries. Currently, numerous open knowledge bases or ontologies have been published, including WordNet [71], DBpedia [6], Wikidata [24] etc. and industry Knowledge Graph (eg. Google, Microsoft's Bing, Facebook, eBay, IBM etc.) [77].

There are many papers summarising the current state of research on knowledge graphs. One of the most recent contributions is by Hogan et al. (2021) [45] who provide a comprehensive introduction to knowledge graphs. The authors compare existing data query models and languages, and summarize methods for creating, evaluating, and publishing knowledge graphs.

Knowledge plays an important role in reasoning-driven Natural Language Processing (NLP) tasks. Indeed, knowledge graphs have emerged as an important tool for addressing various NLP problems, such as Question Answering (KGQA) [49, 68, 83]. Semantics in information can help in extrapolating information that is more *semantically close* to the query. Structured knowledge is also a key element in conversational AI where virtual assistants (e.g. Alexa, Siri or Cortana) answer questions in an advanced way (open questions), as opposed in a more advanced manner to common chatbots programmed only to responde to strictly controlled questions (closed questions). Recently, research works have focused on collecting different techniques for constructing Knowledge Graphs (KG) and their application [134]. In particular, KGs have various application perspectives across different domains such as medical, financial, cybersecurity, news and education, social network de-anonymization, classification, geoscience.

Although several surveys on knowledge graph embeddings in general [16, 19, 112] and specifically on the biological topic [73] have been published over the past few years, this paper aims to explore and summarise the most recent advances in the application of KGs, providing a concise overview of the topic. The goal is to distinguish between biological and clinical domains, and highlight potential issues that may arise from careless construction of KGs, as well as providing information on how KGs support semantic knowledge. Current applications of the latest NLP models for creating clinical KGs are shown in this context. Furthermore, we introduce the most recent and promising future research paths (e.g., the use of multimodal approaches and Simplicial neural networks) in the fields of biomedical and precision medicine. Additionally, the paper expands on the study of usable resources for constructing a biomedical KG.

As this paper is an extended version of the conference paper [28] presented at a K-GALS workshop organized in conjunction with the ADBIS 2022 conference, new resources have been introduced for the construction of knowledge graphs (KGs). Moreover, it conducts an in-depth analysis of the methods and data currently used in biochemical and clinical applications.

2. Knowledge Graphs in Bioinformatics

The application of KGs in the field of biomedical data for decision support spanes from clinical to the biological applications. One of the earliest and most renowned rule-based systems for medical diagnosis is MYCIN [107] which has a knowledge base of 600 rules. There is a close connection between KG and biomedical NLP. On one hand, this connec-

tion allows for the enhancement of the amount and representation of data in KG. On the other hand KG enables improvements in predictions for solving NLP tasks (e.g. named entity recognition (NER) [57] and relation extraction [44]).

Relationship extraction systems are crucial for identifying connections between a wide variety of topics. For instance, they are needed to assess the relationship between non-pharmacological variables and COVID-19 pandemic as well as to support policy-making on COVID-19 in public health [125].

A knowledge graph in the biomedical field is used to connect a vast amount of interrelated information: genes with biological processes, molecular functions, and cellular components; genes with phenotype or interaction with other genes; drugs with the diseases they treat; genes responsible for diseases; generic symptoms related to diseases, etc. Using graphs as a representation of biomedical data seems to be the most natural solution for modelling objects of this type. In Fassetti et al. [25], graphs are used for the identification of features that characterize and at the same time discriminate gene expression among sets of healthy/diseased samples. This is accomplished through the identification of patterns within the graphs belonging to the sample sets with complementary health statuses. In the Table 1, most of the knowledge databases (KB) that are used for constructing and integrating knowledge in the context of biological and clinical data are listed. The data coverage and complexity are specified for each KB, along with their last update (release) dates.

Biomedical databases, in general, play a crucial role in scientific research contributing to drugs development, disease diagnosis and treatment, and understanding biological processes. Although some of these databases are not structured as knowledge graphs (KG), integrating them into a knowledge graph can maximize their potential. By connecting information from different databases and linking data of diverse nature (e.g., clinical data with genetic information), enables researchers to uncover hidden relationships and connections, potentially leading to the discovery of new associations and insights in biomedical research. Therefore, Table 2 lists the most well-known and utilized databases in biomedical research. In addition, the Table 2 distinguishes databases based on their creation methodology: manually curated (by experts), automatic extraction systems or mixed methodology (automatic and manually curated). This type of distinction is particularly important in the biomedical context. A manually curated biomedical database is often considered superior to one created through an automatic methods in certain contexts. Databases curated manually by experts are regarded as the best in terms of precision, reliability, contextualization, and continuous information updates. However, it is also important to note that this process is slow, expensive and requires the collaboration of industry-specific experts. The choice of information generation methods for populating these databases depends heavily on the resources available and the complexity of the data to be evaluated.

miRNA (microRNA) databases are crucial as they offer fundamental information about microRNA sequences, their functions, their interactions with target genes, and their involvement in biological and pathological processes. Currently, there are not many KG that utilize this type of data in conjunction with generic biomedical databases. The Table 3 displays the currently available human miRNA databases, for the same reasons as for general biomedical databases. These miRNA databases may not be structured as knowledge graphs but can still be used to discover new associations and insights regarding the roles

of miRNA in various biological and pathological contexts. Integrating miRNA databases into a knowledge graph provides a broader context and a more integrated approach to understanding the roles of miRNA in gene regulation and biological networks. This approach, along with the use of this data, can lead to the identification of biomarkers, and a better understanding of diseases. Using genetic, molecular and other specific KG information containing details about Human miRNA can be valuable in the context of precision medicine accelerating research and development of customized therapies.

In literature, although using different methods and algorithms, graphs are mostly used to solve common problems: making inferences about biomedicine, creating alternative ways to represent graphs on the same knowledge domain and extending information extraction.

A large part of research is currently devoted to the identification of similar entities within a KG. Embeddings generated using neural networks are used to calculate knowledge-based similarities between, for example, drugs, proteins and diseases [127].

Many ways are used to extend knowledge in the biomedical domain to discover latent information or missing information in KGs.

Completion of the knowledge graph (KGC) aims to complete the structure of the knowledge graph by predicting the missing entities or relationships in the knowledge graph and extracting unknown facts. KGC technologies may involve the use of traditional methods, such as rule-based reasoning and the probability graph model (Markov logic network). Recently, KGC techniques use methods of learning through embeddings representation: methods based on semantic correspondence models, based on learning of representation and other methods based on neural network models.

The use of models based on a generative approach to learn the embeddings of entities and relationships allows to generate hypotheses regarding the relationships associated with a connection score between graph embeds through multiple techniques: tensor factorisation (DistMult model [9]) and latent distance similarity (TransE model [124]). This type of techniques are used in polypharmacy, to evaluate the side effects that are caused by the interaction of drug combinations [70, 76].

2.1. Example of construction of knowledge graph

Constructing knowledge graphs from heterogeneous biomedical databases (see Table 1, Table 2, Table 3) involves several complex steps, such as data integration, ontology alignment, and semantic integration. To effectively navigate these challenges, it is essential to first define the objectives of the knowledge graph construction project. For instance in the context of cancer research, the objective may be to integrate diverse biomedical datasets to facilitate knowledge discovery, data-driven insights, and personalized medicine.

The following steps can guide the construction of such KG with the focus on cancer.

1. **Identify and selection relevant biomedical databases.** For example, to construct a KG specialized on cancer, we need to consider database containing genomic data (e.g., TCGA), drug data (e.g., DrugBank), molecular pathway databases, diseases-gene associations data (e.g, COSMIC, DISEASES). It's important to note that the chosen databases are mostly created through manual curation by experts (see Table 2, Table 3). In specific contexts, a manually curated biomedical database is frequently regarded as superior to one generated through automated methods, as mentioned previously.

Table 1. A list of 13 knowledge data sources that are useful in the biomedical context, is provided

Knowledge database	Description	Coverage	Last Release
STRING [101]	database of known and predicted protein-protein interactions	67.592.464 proteins from 14.094 organisms.	August 2021
iDISK [87]	Dietary Supplements (DS) Knowledge base	4.208 DS concept, 495 drugs, 776 diseases, 985 symptoms, 605 therapeutic classes, 17 system organ classes, and 137.568 DS products.	February 2020
Hetionet [43]	biomedical knowledge assembled from 29 different databases (genes, compounds, diseases, etc.)	47.031 nodes of 11 types and 2.250.197 edges of 24 types.	February 2017
DRKG [50]	biological knowledge graph relating genes, chemical compounds, biological processes, drug side effects, diseases, and symptoms.	100.000 entities of more than 12 types. 6.000.000 relationships of more than 100 types.	in 2020
KEGG [79]	reference knowledge base for integration and interpretation large-scale molecular data sets (genomic, chemical and health information)	563 pathway maps, 47.296.502 genes, 9.010 organisms, 2.640 human diseases, 12.136 drugs etc.	May 2023
PharmGKB [130]	knowledge on actionable gene-drug associations and genotype-phenotype relationships	759 drugs, 1761 genes, 213 pathways, 227 diseases, 200 clinical guidelines, and 993 drug labels.	in 2023
Gene Ontology (GO) [2]	describes knowledge of the biological domain: molecular function, biological process, cellular component	7.554.638 annotations 1.519.515 gene products 5.291 species.	May 2023
UniProtKB [3]	collection of annotated functional information on proteins	Swiss-Prot: 569.516 seq, 205.866.895 amino acids; TrEMBL: 249.308.459 seq, 86.853.323.495 amino acids.	May 2023
Reactome [31]	knowledge graph that focuses on biological pathways and their relationships	95.164 proteins, 102.459 complexes, 90.807 reactions, 22.050 pathways; 11.278 human proteins.	March 2023
OncoKB [10]	precision oncology knowledge base, consolidating biological and clinical data on genomic alterations in cancer	Memorial Sloan Kettering (MSK), provides accurate information about the biological and clinical implications of over 5.000 cancer gene alterations.	May 2023
OGB-Biokg [47]	biomedical knowledge graph constructed by Stanford University (associations between proteins, e.g., physical interactions, co-expression, homology or genomic neighborhood etc.)	5 types of entities: diseases (10,687 nodes), proteins (17,499), drugs (10,533 nodes), side effects (9,969 nodes), protein functions (45,085 nodes).	April 2023
Bioteque [26]	a resource of biological knowledge graph embeddings	12 types of biological entities (e.g. genes, diseases, drugs) and 67 types of relationships.	July 2022
NCBI [91]	knowledge graph representing gene-related information, including gene sequences, gene structures, functional annotations, and genetic variations.	242.554.936 GenBank sequences	April 2023

Table 2. A list of 24 biomedical databases, valuable for biomedical research yet not explicitly structured as KGs, is provided

Database	**Main scope**	**Manually curated?**
SCOPe [12]	protein structural relationships	Mostly manually
Protein Data Bank (PDB) [1]	archive of 3D structure data for large biological molecules	Yes
CATH [96]	hierarchical classification of protein domains	Mixed with other methods
InterPro [81]	functional analysis of protein sequences by classifying them into families and predicting domain presence and important sites	Mixed with other methods
The Human Protein Reference Database [58]	for each protein in the human proteome integrate information pertaining: domain architecture, post-translational modifications, interaction networks and disease association.	Yes
Bgee [7]	gene expression patterns across multiple animal species	Yes
HGNC [105]	relation between gene symbol and their corresponding entries in other database	Mixture with other methods
DrugBank [116]	molecular information about drugs, their mechanisms, their interactions and their targets	Yes
Supertarget [41]	integrates drug-related information associated with medical indications, adverse drug effects, drug metabolism, pathways and (GO) terms for target proteins	Yes
SIDER [62]	collects information on drug classification and side effects and links to further information, e.g. drug-target relations	No
OFFSIDES [104]	database of drug side-effects	No
TWOSIDES [104]	database drug-drug-effect	No
STITCH [102]	database of known and predicted interactions between chemicals and proteins	Yes
SIGNOR [102]	causal relationships between human proteins, chemicals of biological relevance, stimuli and phenotypes	Yes
SMPDB [51]	database containing pathways found in model organisms such as humans, mice, E. coli etc.	No
ChEMBL [30]	chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs	Yes
ChEBI [40]	dictionary of molecular entities focused on ‘small’ chemical compounds	Yes
PubChem [59]	chemical database	Yes
TISSUE [90]	tissue expression, proteomics and transcriptomics screens	Mixture with other methods
Brenda Tissue Ontology [13]	collection of enzyme functional data	Not specified
Disease Ontology [93]	ontology for human disease	Not specified
Cell Ontology [55]	repository for biomedical ontologies	Not specified
DISEASES [82]	integrates disease-gene associations, cancer mutation data, and genome-wide association studies	Mixed with other methods
COSMIC [27]	Catalogue Of Somatic Mutations In Cancer, associate genes with the related cancer type	Yes

Table 3. A list of 8 databases containing Human miRNA information, valuable for biomedical research yet not explicitly structured as KGs, is provided

Database	**Main Goal**	**Manually curated?**
mirCancer [120]	contains associations between miRNAs and human cancers.	Mixed with other methods
miRanda [52]	miRNA-target interactions.	Not specified
miRBase [60]	miRNAs sequences and annotations, associated with names, keywords, genomic locations, and references.	No
miRNASNP [34]	contains miRNA-related mutations.	Mixed with other methods
miRTarBase [48]	miRNA-target interactions, including those implicated in cancer. Experimentally validated miRNA-target interactions.	Mixed with other methods
DIANA-TarBas [108]	miRNA-target interactions, including those related to cancer. It offers information on miRNA regulation of target genes and associated functional effects.	Yes
OncomiRDB [109]	miRNA expression profiles in various cancer types, along with their putative target genes and functional implications.	Yes
TCGA [4]	Cancer Genomics Databases. It provides miRNA dysregulation patterns in cancer and their potential roles as biomarkers or therapeutic targets.	Yes

2. **Develop an integration strategy.** Use Extract, Transform, Load (ETL) pipelines or data ingestion tools to extract data from these heterogeneous sources. Each source may have its own data format (e.g Json, XML, RDF etc.), and schema.
3. **Ontology Alignment.** Within cancer research, critical ontologies include the Human Phenotype Ontology (HPO) [36], which catalogues phenotypic abnormalities linked to genetic diseases, and the Disease Ontology (DO) [93] providing a standardized vocabulary for disease classification. Complementary ontologies encompass the Gene Ontology (GO) [2], delineating molecular functions, biological processes, and cellular components, as well as the Cell Ontology [55], characterizing cell types and anatomical structures. At this step, it is necessary to establish correspondences or mappings between entities of different ontologies. Furthermore, it is important to establish common standards and formats to facilitate comparison and alignment. Software platforms like e.g BioPortal can help researchers to systematically assess the efficacy and accuracy of ontology alignment methodologies in biomedical domains.
4. **Mapping data into graph-based data model.** The goal of this step is to develop and define a standardized schema for our data, pinpointing crucial data entities and their relationships. RDF (Resource Description Framework) triples, can link together diverse elements such as genes, proteins, diseases, drugs, and pathways. This approach fosters a comprehensive comprehension of intricate biological systems by forging significant links between different components.
5. **Semantic Integration.** Employ semantic reasoning techniques (e.g, Rule-based Reasoning, Semantic Similarity etc.) to deduce relationships and uncover novel insights from the heterogeneous data.
6. **Quality check.** Define robust quality metrics to quantitatively evaluate the comprehensiveness, precision, and coherence of the constructed knowledge graph. Conduct thorough validation procedures to verify data integrity and adherence to ontology mappings, employing systematic validation protocols. In this step, for example is necessary to solicit feedback on the accuracy, relevance, and completeness of the mappings, incorporating expert insights to refine the validation process, ensuring its scientific validity and applicability in the field.

2.2. Methods of Knowledge Graph Embedding

A very common application that has grown recently is the creation of entity embeddings or assertions on KGs by training deep learning networks such as autoencoders from inputs constructed by KG nodes [69, 133]. The purpose of representing graphs in a high-dimensional space to a low-dimensional space is to capture the essence of a graph while preserving its intrinsic (global and / or local) structure in the form of a dense vector representation, both of arcs [9] and of single nodes [35]. This approach has been used to analyze knowledge graphs across different domains enabling, starting from compressed and *meaningful* information, the application of classification techniques and the development of predictors, which can aid researchers in identifying associations between diseases and bio-molecules [95], discovering new treatments for existing drugs [43] and addressing other related problems.

In the following section, we will explore how the most recent techniques of *representative learning* are applied in the biochemical and clinical context.

Application on Biochemical Data. One of the studies that facilitated the creation of a comprehensive KG in the biomedical context is PharmKG [130]. This study aggregated multi-omics data, including disease-related words, gene expression and chemical structure information, while preserving biological and semantic features through the latest KGE approaches. A significant aspect of the study focused on drug-related topics, particularly addressing drug reuse and adverse reactions, crucial for preventing harm to patients. Additionally, the study investigated potential drug-drug interactions (DDIs), drug-protein (DPIs), drug-disease, drug-target interactions (DTIs) by modelling the problem of predicting links between graph nodes with KGs. In the study by Zhu et al. [132], the authors provided a detailed overview of existing drug knowledge bases and their applications. This work used datasets containing key properties of drugs (DrugBank [116] and SuperTarget [41]) as well as datasets containing the main information about chemical compounds (PubChem [59] and ChEMBL [30]). In the study by Lin et. al [65], a Bio2RDF created from DrugBank was employed to evaluate the relationships between a potential drug and its neighbours. This evaluation is based on GNN models applied to the biological KG. Innovative work in drug research includes the development of the BioDKG-DDI model [86], that aims to identify DDI interaction relationships to support experimental work in drug development laboratories. BioDKG-DDI uses an innovative self-attention mechanism on DNNs (deep neural networks)to attenuate multi-features embeddings including molecular structures, drug structures, and drug similarity matrix. Furthermore, recent advancements in DNNs are employed to predict the search for similar drugs, e.g. to identifying molecules with antibiotic properties through graph-based retro-synthesis [66]. Drug-protein interactions are another area of interest for researchers. In BridgeDPI [118], convolutional CNN and feed-forward network layers are used to encode SMILES representation of drug and protein sequences. GNNs are then employed, as in other similar works in the literature, to build *bridge* nodes between interactions and predict new connections between nodes.

Many papers in scientific literature in recent years have focused on the study of graph models applied to KG for the discovery of Drug-Target Interactions (DTIs) [5, 56, 84, 100, 110, 113, 114]. Drug-Target Interactions (DTIs) are the interactions between drugs and molecular targets in the human body. The interest in this type of interaction is justified by the fact that it is now essential to understand how drugs act in the body, how they bind to molecular targets and how they affect biological processes. These types of interactions can determine the effectiveness of a drug in treating a disease and may highlight its unwanted side effects. In pharmaceutical research it is crucial to have a comprehensive picture of DTIs interactions in a KG in order to develop safer, more effective and targeted drugs for the treatment of complex diseases. A new DT2Vec+ [5] approach to the computational reprocessing of drugs to predict new drug-target interactions (DTIs) was proposed in 2023, which showed promising results. DT2Vec+ was created by integrating and mapping drug-target-disease triplet association graphs. The heterogeneous graph in DT2Vec+ with “drug”, “target” and “disease” entities has been mapped to low-size vectors using node embedding principles to create specific characteristics for each entity. The authors also tested the new method on DTI tasks to propose drugs targeted at specific cancer biomarkers. Another approach for DTI predictions is KG-DTI [113]. The KG is constructed using 29,607 positive drug-target pairs. To extract the built-in features, KG-DTI, uses the DistMult embedding strategy instead. KG-DTI is then applied to recommend drugs for AD (Alzheimer’s disease) by targeting apolipoprotein E.

The results presented by the authors of KG-DTI show that seven of the top ten drugs recommended for AD are supported and validated by clinical practice and literature. KG2ECapsule [100] employs entity representations obtained by recursively propagating takeovers from the receptive fields of attention-based entities, similar to DTI-GAT (Drug-Target Interaction prediction with Graph networks attention) [110]. In particular, DTI-GAT uses the attention mechanism on graphs to facilitate the topological interpretation of DTI, assigning a different attention weight to each node in KG. The accuracy rate of DTI-GAT reaches 93.75, on enzyme dataset (BRENDA [92]), surpassing that of other prediction methods. An innovative approach for DPI is employed by the authors of TransDTI [56] (Transformer-Based Language Models for Estimating DTIs). They use transformer-based language models to classify interactions between drug-target pairs as active, inactive, and intermediate. The results presented by the TransDTI authors suggest that transformer-based linguistics effectively predict new drug-target interactions from sequence data. In 2022, GCHN-DTI [114], introduced a heterogeneous network created from various data sources including drug-target interactions, drug-drug interactions, similarities between drugs, target-target interactions and similarities between targets. In GCHN-DTI utilizes a graph convolution approach for the DTI task. The method employs an attention mechanism between convolutional graph layers to combine the embedding of nodes of each layer. GCHN-DTI demonstrates superiority over several state-of-the-art methods. One knowledge graph embedding approach that integrates and works well on DDI (Drug-Drug Interaction), DTI (Drug-Target Interaction) and PPI (Protein-Protein Interaction) is ConvE-Bio [84]. While ConvE-Bio serves as a powerful tool for predicting biomedical relationships it currently faces limitations related to processing large graphs. As presented in the Table 4 several noteworthy works focus on solving different tasks. For "diseases diagnosis" a recent tool based on the study gene association information and cofunctional gene modules is MLA-GNN(multi-level attention graph neural network) [121]. MLA-GNN achieves state-of-art performance on transcriptomic data [4] and proteomic data (COVID-19). The authors also employ an innovative mechanism to try to identify the genes most involved in model analysis and prediction. Another type of task studied by researchers in this context is the application of knowledge graph-based "disease-gene" prediction. The GenePredict-KG model [29] is developed for this purpose by integrating several datasets. Despite achieving results that surpass state-of-the-art performance, the method suffers from several limitations related to class imbalance.

In order to provide a comprehensive overview of current research, it is important to mention the application of Neural networks known as Hyperbolic Graph Neural Networks (HGNN) to the DISEASE dataset, based on the SIR disease spreading model [11], demonstrating excellent results in link prediction. However, recent advancements have shown that these new sophisticated neural networks have been outperformed by Simplicial neural networks (SNNs) for link prediction [15], achieving better results in terms of ROC AUC on the same dataset.

In the following Table 4 we will succinctly present, the latest applications of knowledge graph embedding in the context of biochemical data and the tasks they aim to address.

Application on Clinical Data. Studies of KG-based recommendation systems built from electronic medical records (EMRs) aim to enhance medical decision-making for improved

Table 4. Latest KGs constructions and Graph Neural Network applications in the biochemical field

Model	Task	Dataset	Year
BioDKG-DDI [86]	DDI	DrugBank [116], SIDER [62], KEGG [79], PubChem [59] and OFFSIDES [104]	2022
BridgeDPI [118]	DPI	BindingDB [32],C.ELEGANS and HUMAN datasets [67], DUD-E [75]	2022
RetroGNN [66]	Drug Discovery	Zinc15 database [99]	2020
HGNN [11]	Link Prediction, Node Classification	Disease, PubMed	2019
ConvE-Bio [84]	DDI, DTI, PPI	DrugBank [116], Human Protein [88]	2023
DT2Vec+ [5]	DTI	CTD [20], DrugBank [116], ChEMBL [30]	2023
KG-DTI [113]	DTI, DTP	DrugBank [116]	2021
KG2ECapsule [100]	DDI	DrugBank [116], OGB-Biokg [47], KEGG [79]	2023
DTI-GAT [110]	DTI	SuperTarget [41], DrugBank [116], KEGG [79], BRENDA [13]	2021
TransDTI [56]	DTI	ChEMBL [30], Kiba [103]	2022
GCHN-DTI [114]	DTI	DrugBank [116]	2022
KG-COVID-19 [85]	ML tasks, queries	PharmGKB [130], Therapeutic Target Database (TTD) [14], ChEMBL [30], GO [2], STRING [101], IntAct Molecular Interaction Database [80]	2021
MLA-GNN [121]	Disease diagnosis	TCGA [4], COVID-19 [111]	2022
KGMultiple Ontologies [78]	Gene-Disease Association	Uniprot [18], OMIM [38], Orphanet [115]	2020
GenePredict-KG [29]	Gene-Disease Association	STRING [101], SIDER [62], DrugBank [116], Human Phenotype Ontology (HPO) [36] Genotype-Tissue Expression (GTEx) [128] Gene Ontology Annotation (GOA) [17] Mammalian Phenotype (MP) [97] Mouse Genome Informatics (MGI) [23], PubChem [59], OMIM [38]	2023
GNBR [98]	Drug repurposing	Orphanet [115], OMIM [38], UMLS [8] DrugCentral [106]	2019
Compact Walks [46]	Pathways discovery	Hetionet [43], ROBOKOP [74]	2022

patient care. Creating KGs from medical record texts containing a patient's treatment history (medical diagnoses, therapies, etc.) presents a cost-effective approach compared to building KGs based on deeper biological aspects (relationships between genes, diseases, chemical composition of drugs, etc.) that require more attention. Research in this area is still in its early stages and recent advancements in information extraction models (such as LSTM, BERT, and NER models) enable the extraction of meaningful information from unstructured data, enriching biomedical knowledge bases with non-trivial connections [39] [64] [33]. Recently, Zhang et al. [129] demonstrated the effectiveness of attention mechanisms and convolutional graphs techniques in creating embedded KGs features enhance the classification and generation of radiological reports in order to improve diagnosis and support physicians in their work.

The Table 5 presents several noteworthy research studies from last three years, which incorporate clinical data of various types (e.g, images, ontologies, etc.). It is noteworthy that currently only MKGs [117] constructs KGs that encompass both biomedical and clinical data. As a result, it has the potential to address numerous tasks such as drug-drug interactions (DDIs), drug-protein interactions (DPIs), classification of nodes, and more.

Table 5. Latest KGs constructions and applications in the clinical field

Model	Task	Dataset	Year
ClinicalBERT [89]	link prediction	PubMed abstract, MIMIC-III database [54]	2021
SMR [33]	link prediction	MIMIC-III [54], DrugBank [116], ICD-9 ontology [94]	2020
RR-KG [129]	generation of radiological reports	U-RR dataset [21]	2020
MaKG [123]	generation of radiological reports	IU XRay [21]and MIMIC CXR [53]	2022
MKGs [117]	several	real world data (EMH, EHR etc), UMLS [8], ROBOKOP [74], DrugBank [116], UniProt [18], InterProt [81], SIDER [62], GO [2], KEGG [79], Therapeutic target database [14] etc.	2023

3. Open research problems

3.1. Construction and integration of knowledge graphs

Biomedical knowledge graphs are typically curated manually by expert researchers. One such example is COSMIC [27], constructed by a group of domain experts who associated genes with related cancer types based on literature. However, the field of biological knowledge is constantly evolving, necessitating scalable intelligent systems capable of integrating real-time updates. Addressing this challenge involves not only updating knowledge bases but also ensuring the reliability of knowledge representations in KGs and their relationships. One widely studied technique for enhancing KG reliability involves aligning entities from different KGs based on their similarity. Recently, Xiang et al. [119] introduced a method that incorporates ontology hierarchies and class disjunctions to improve entity alignment accuracy and avoid mismapping.

Research has also shown that the quality of available knowledge graphs directly impacts the accuracy of knowledge graph embedding (KGE) predictions. Low data quality can propagate into embedding models, leading to decreased prediction accuracy [73]. Missing knowledge and integration errors in KGs can further exacerbate this issue, perpetuating incorrect and misleading domain knowledge. This is particularly problematic in the biomedical domain, where inaccuracies can have significant consequences.

3.2. Performance

Complex biological systems are often represented as graphs, but the exploration, training and prediction techniques applied to these graphs require significant of resources and time leading to limited scalability. While knowledge KGEs address some aspects of this problem by operating with linear time and space complexity, the challenge of dynamically encoding new entities into the graph remains unresolved. KGEs rely heavily on prior knowledge of embeddings for each type of information in the knowledge base, allowing them to maintain both local and global information. However, this dependence on prior knowledge presents scalability issues that propagate into the prediction process.

3.3. Explainable predictions

Lack of interpretability is a recurring problem with deep learning models [37]. Which becomes particularly concerning given the increasing use of neural networks in decision-making within biomedical applications. Efforts have been made in this regard to address this issue. For example, CrossE [126] explores the process of explaining graph search paths using embeddings to interpret link prediction. In the context of KGE, learning meaningful embeddings through specific optimisation techniques often leads to predictions that are difficult to interpret. In data analysis, GNN models frequently employed in the biomedical domain, generate relevant information for each data node, thereby enhancing interpretability to some extent. Recent efforts have aimed to impose constraints during training to make KGE models partially interpretable (e.g. type constraints and basic relation axioms) [61, 72]. Extracting information from NLP presents challenges for constructing reliable KG in the health domain. Complex models used for understanding natural language still have many issues [42]. Importantly, biases inherent in extracting information from EMHs should not be underestimated. Inevitably these biases in the data will propagate to some extent in the results of the predictions. Therefore is crucial for research to prioritize the development of more reliable and explainable models for the healthcare sector.

4. Discussion and conclusion

This survey aims to present the latest models and strategies to use knowledge graphs in the biomedical context. Their use has become increasingly widespread in recent years, with current research focusing on enhancing the outcomes derived from their application in biomedicine. As outlined in this survey, many knowledge graphs are typically constructed from data sources, which are either manually curated by experienced researchers or generated through sophisticated NLP techniques (NER, relation extraction). We subsequently pointed out the potential errors that this approach may introduce in the biomedical context, during KGs construction. In this regard, with the aim of ensuring the future research in the biomedical domain is increasingly reliable and accurate, this review delves into the detailed construction methods of biological, chemical and clinical databases (see Table 2 and Table 3). The differences between the types of entities used in the biomedical knowledge bases and their "size" are noted in Table 1.

The process of extending knowledge in KGEs can indeed be addressed by the low-dimensional representation of the characteristics of each entity and/or relation within the graph. This compressed and representative representation of the knowledge graph can help identify potential inconsistencies during the integration process the graphs and partially resolve some problems associated with errors in knowledge graph construction caused by misaligned entities. KGEs are currently highly active area of research, due to their ability to provide a generalisable context on the KG and probabilistically deduce new relations missing in the existing graph structure. This characteristic has accelerated the discovery of new drugs in many studies by evaluating the interaction between properties of molecules present in the KG. The importance of KG feature representation, as discussed, underscore its effectiveness in constructing increasingly comprehensive KGs.

A recent advancement in research involves the construction of a multimodal knowledge network, where additional information is incorporated into the KG to enhance rea-

soning. This approach utilizes a combination of various interaction features among KG entities to improve predictions (e.g. on drug repositioning) [122]. The multimodal approach has also been recently applied in precision medicine, where detailed knowledge and a specific focus are essential for creating KGs that represent and generate *reliable* knowledge [131].

In conclusion, this discussion highlights the current open challenges in the use of KGs in the biomedical field, emphasizing the need to improve the interpretability and quality of biomedical KG data in order to increase confidence within the community regarding predictions and thereby support advancements in specialised medicine.

References

1. Protein data bank: the single global archive for 3d macromolecular structure data. Nucleic acids research 47(D1), D520–D528 (2019)
2. The gene ontology resource: enriching a gold mine. Nucleic acids research 49(D1), D325–D334 (2021)
3. Uniprot: the universal protein knowledgebase in 2021. Nucleic acids research 49(D1), D480–D489 (2021)
4. 53, D.C.C.B.R..J.M.A..K.A..P.T..P.D..W.Y., 68, T.S.S.L.D.A.: The cancer genome atlas pan-cancer analysis project. Nature genetics 45(10), 1113–1120 (2013)
5. Amiri Souri, E., Chenoweth, A., Karagiannis, S., Tsoka, S.: Drug repurposing and prediction of multiple interaction types via graph embedding. BMC bioinformatics 24(1), 1–17 (2023)
6. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings. pp. 722–735. Springer (2007)
7. Bastian, F.B., Roux, J., Niknejad, A., Comte, A., Fonseca Costa, S.S., De Farias, T.M., Moretti, S., Parmentier, G., De Laval, V.R., Rosikiewicz, M., et al.: The bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. Nucleic Acids Research 49(D1), D831–D847 (2021)
8. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research 32(suppl_1), D267–D270 (2004)
9. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. Advances in neural information processing systems 26 (2013)
10. Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al.: Oncokb: a precision oncology knowledge base. JCO precision oncology 1, 1–16 (2017)
11. Chami, I., Ying, Z., Ré, C., Leskovec, J.: Hyperbolic graph convolutional neural networks. Advances in neural information processing systems 32 (2019)
12. Chandonia, J.M., Guan, L., Lin, S., Yu, C., Fox, N.K., Brenner, S.E.: Scope: improvements to the structural classification of proteins–extended database to facilitate variant interpretation and machine learning. Nucleic acids research 50(D1), D553–D559 (2022)
13. Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., Neumann-Schaal, M., Jahn, D., Schomburg, D.: Brenda, the elixir core data resource in 2021: new developments and updates. Nucleic acids research 49(D1), D498–D508 (2021)
14. Chen, X., Ji, Z.L., Chen, Y.Z.: Ttd: therapeutic target database. Nucleic acids research 30(1), 412–415 (2002)
15. Chen, Y., Gel, Y.R., Poor, H.V.: Bscnets: block simplicial complex neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 6333–6341 (2022)

16. Choudhary, S., Luthra, T., Mittal, A., Singh, R.: A survey of knowledge graph embedding and their applications. arXiv preprint arXiv:2107.07842 (2021)
17. Consortium, G.O.: Gene ontology consortium: going forward. Nucleic acids research 43(D1), D1049–D1056 (2015)
18. Consortium, U.: Uniprot: a worldwide hub of protein knowledge. Nucleic acids research 47(D1), D506–D515 (2019)
19. Dai, Y., Wang, S., Xiong, N.N., Guo, W.: A survey on knowledge graph embedding: Approaches, applications and benchmarks. Electronics 9(5), 750 (2020)
20. Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., McMorran, R., Wiegers, J., Wiegers, T.C., Mattingly, C.J.: The comparative toxicogenomics database: update 2019. Nucleic acids research 47(D1), D948–D954 (2019)
21. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association 23(2), 304–310 (2016)
22. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 601–610 (2014)
23. Eppig, J.T.: Mouse genome informatics (mgi) resource: genetic, genomic, and biological knowledgebase for the laboratory mouse. ILAR journal 58(1), 17–41 (2017)
24. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing wikidata to the linked data web. In: The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I 13. pp. 50–65. Springer (2014)
25. Fassetti, F., Rombo, S.E., Serrao, C.: Discovering discriminative graph patterns from gene expression data. In: Proceedings of the 31st Annual ACM Symposium on Applied Computing. pp. 23–30 (2016)
26. Fernández-Torras, A., Duran-Frigola, M., Bertoni, M., Locatelli, M., Aloy, P.: Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the bioteque. Nature Communications 13(1), 5304 (2022)
27. Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., et al.: Cosmic: somatic cancer genetics at high-resolution. Nucleic acids research 45(D1), D777–D783 (2017)
28. Galluzzo, Y.: A review: Biological insights on knowledge graphs. In: New Trends in Database and Information Systems: ADBIS 2022 Short Papers, Doctoral Consortium and Workshops: DOING, K-GALS, MADEISD, MegaData, SWODCH, Turin, Italy, September 5–8, 2022, Proceedings. pp. 388–399. Springer (2022)
29. Gao, Z., Pan, Y., Ding, P., Xu, R.: A knowledge graph-based disease-gene prediction system using multi-relational graph convolution networks. In: AMIA Annual Symposium Proceedings. vol. 2022, p. 468. American Medical Informatics Association (2022)
30. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., et al.: Chembl: a large-scale bioactivity database for drug discovery. Nucleic acids research 40(D1), D1100–D1107 (2012)
31. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., et al.: The reactome pathway knowledgebase 2022. Nucleic acids research 50(D1), D687–D692 (2022)
32. Gilson, M.K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., Chong, J.: Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic acids research 44(D1), D1045–D1053 (2016)
33. Gong, F., Wang, M., Wang, H., Wang, S., Liu, M.: Smr: medical knowledge graph embedding for safe medicine recommendation. Big Data Research 23, 100174 (2021)

34. Gong, J., Tong, Y., Zhang, H.M., Wang, K., Hu, T., Shan, G., Sun, J., Guo, A.Y.: Genome-wide identification of snps in microrna genes and the snp effects on microrna target binding and biogenesis. Human mutation 33(1), 254–263 (2012)
35. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 855–864 (2016)
36. Groza, T., Köhler, S., Moldenhauer, D., Vasilevsky, N., Baynam, G., Zemojtel, T., Schriml, L.M., Kibbe, W.A., Schofield, P.N., Beck, T., et al.: The human phenotype ontology: semantic unification of common and rare disease. The American Journal of Human Genetics 97(1), 111–124 (2015)
37. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51(5), 1–42 (2018)
38. Hamosh, A., Scott, A.F., Amberger, J., Valle, D., McKusick, V.A.: Online mendelian inheritance in man (omim). Human mutation 15(1), 57–61 (2000)
39. Harnoune, A., Rhanoui, M., Mikram, M., Yousfi, S., Elkaimbillah, Z., El Asri, B.: Bert based clinical knowledge extraction for biomedical knowledge graph construction and analysis. Computer Methods and Programs in Biomedicine Update 1, 100042 (2021)
40. Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., Steinbeck, C.: Chebi in 2016: Improved services and an expanding collection of metabolites. Nucleic acids research 44(D1), D1214–D1219 (2016)
41. Hecker, N., Ahmed, J., Von Eichborn, J., Dunkel, M., Macha, K., Eckert, A., Gilson, M.K., Bourne, P.E., Preissner, R.: Supertarget goes quantitative: update on drug–target interactions. Nucleic acids research 40(D1), D1113–D1117 (2012)
42. Helwe, C., Clavel, C., Suchanek, F.M.: Reasoning with transformer-based models: Deep learning, but shallow reasoning. In: 3rd conference on automated knowledge base construction (2021)
43. Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., Baranzini, S.E.: Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife 6, e26726 (2017)
44. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D.S.: Knowledge-based weak supervision for information extraction of overlapping relations. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 541–550 (2011)
45. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., d Melo, G.: Knowledge graphs [j]. synthesis lectures on data semantics and knowledge (2021)
46. Hou, P.Y., Korn, D.R., Melo-Filho, C.C., Wright, D.R., Tropsha, A., Chirkova, R.: Compact walks: Taming knowledge-graph embeddings with domain-and task-specific pathways. In: Proceedings of the 2022 International Conference on Management of Data. pp. 458–469 (2022)
47. Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., Leskovec, J.: Open graph benchmark: Datasets for machine learning on graphs. arXiv preprint arXiv:2005.00687 (2020)
48. Huang, H.Y., Lin, Y.C.D., Cui, S., Huang, Y., Tang, Y., Xu, J., Bao, J., Li, Y., Wen, J., Zuo, H., et al.: mirtarbase update 2022: an informative resource for experimentally validated mirna–target interactions. Nucleic acids research 50(D1), D222–D230 (2022)
49. Huang, X., Zhang, J., Li, D., Li, P.: Knowledge graph embedding based question answering. In: Proceedings of the twelfth ACM international conference on web search and data mining. pp. 105–113 (2019)
50. Ioannidis, V.N., Song, X., Manchanda, S., Li, M., Pan, X., Zheng, D., Ning, X., Zeng, X., Karypis, G.: Drkg-drug repurposing knowledge graph for covid-19. arXiv preprint arXiv:2010.09600 (2020)

51. Jewison, T., Su, Y., Disfany, F.M., Liang, Y., Knox, C., Maciejewski, A., Poelzer, J., Huynh, J., Zhou, Y., Arndt, D., et al.: Smpdb 2.0: big improvements to the small molecule pathway database. Nucleic acids research 42(D1), D478–D484 (2014)
52. John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., Marks, D.S.: Human microrna targets. PLoS biology 2(11), e363 (2004)
53. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
54. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data 3(1), 1–9 (2016)
55. Jupp, S., Burdett, T., Leroy, C., Parkinson, H.E.: A new ontology lookup service at embl-ebi. SWAT4LS 2, 118–119 (2015)
56. Kalakoti, Y., Yadav, S., Sundar, D.: Transdti: Transformer-based language models for estimating dtis and building a drug recommendation workflow. ACS omega 7(3), 2706–2717 (2022)
57. Karampatakis, S., Dimitriadis, A., Revenko, A., Blaschke, C.: Training ner models: knowledge graphs in the loop. In: The Semantic Web: ESWC 2020 Satellite Events: ESWC 2020 Satellite Events, Heraklion, Crete, Greece, May 31–June 4, 2020, Revised Selected Papers 17. pp. 135–139. Springer (2020)
58. Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al.: Human protein reference database—2009 update. Nucleic acids research 37(suppl_1), D767–D772 (2009)
59. Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., et al.: Pubchem substance and compound databases. Nucleic acids research 44(D1), D1202–D1213 (2016)
60. Kozomara, A., Griffiths-Jones, S.: mirbase: integrating microrna annotation and deep-sequencing data. Nucleic acids research 39(suppl_1), D152–D157 (2010)
61. Krompaß, D., Baier, S., Tresp, V.: Type-constrained representation learning in knowledge graphs. In: The Semantic Web-ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I 14. pp. 640–655. Springer (2015)
62. Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The sider database of drugs and side effects. Nucleic acids research 44(D1), D1075–D1079 (2016)
63. Lassila, O., Swick, R.R.: Resource description framework (rdf) model and syntax specification, w3c recommendation 22 february 1999 (1999)
64. Li, L., Wang, P., Yan, J., Wang, Y., Li, S., Jiang, J., Sun, Z., Tang, B., Chang, T.H., Wang, S., et al.: Real-world data medical knowledge graph: construction and applications. Artificial intelligence in medicine 103, 101817 (2020)
65. Lin, X., Quan, Z., Wang, Z.J., Ma, T., Zeng, X.: Kgnn: Knowledge graph neural network for drug-drug interaction prediction. In: IJCAI. vol. 380, pp. 2739–2745 (2020)
66. Liu, C.H., Korablyov, M., Jastrzebski, S., Włodarczyk-Pruszyński, P., Bengio, Y., Segler, M.H.: Retrognn: Approximating retrosynthesis by graph neural networks for de novo drug design. arXiv preprint arXiv:2011.13042 (2020)
67. Liu, H., Sun, J., Guan, J., Zheng, J., Zhou, S.: Improving compound–protein interaction prediction by building up highly credible negative samples. Bioinformatics 31(12), i221–i229 (2015)
68. Lukovnikov, D., Fischer, A., Lehmann, J., Auer, S.: Neural network-based question answering over knowledge graphs on word and character level. In: Proceedings of the 26th international conference on World Wide Web. pp. 1211–1220 (2017)
69. Ma, T., Xiao, C., Zhou, J., Wang, F.: Drug similarity integration through attentive multi-view graph auto-encoders. arXiv preprint arXiv:1804.10850 (2018)

70. Malone, B., García-Durán, A., Niepert, M.: Knowledge graph completion to predict polypharmacy side effects. In: Data Integration in the Life Sciences: 13th International Conference, DILS 2018, Hannover, Germany, November 20-21, 2018, Proceedings 13. pp. 144–149. Springer (2019)
71. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM 38(11), 39–41 (1995)
72. Minervini, P., Costabello, L., Muñoz, E., Nováček, V., Vandenbussche, P.Y.: Regularizing knowledge graph embeddings via equivalence and inversion axioms. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10. pp. 668–683. Springer (2017)
73. Mohamed, S.K., Nounu, A., Nováček, V.: Biological applications of knowledge graph embedding models. Briefings in bioinformatics 22(2), 1679–1693 (2021)
74. Morton, K., Wang, P., Bizon, C., Cox, S., Balhoff, J., Kebede, Y., Fecho, K., Tropsha, A.: Robokop: an abstraction layer and user interface for knowledge graphs to support question answering. Bioinformatics 35(24), 5382–5384 (2019)
75. Mysinger, M.M., Carchia, M., Irwin, J.J., Shoichet, B.K.: Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. Journal of medicinal chemistry 55(14), 6582–6594 (2012)
76. Nováček, V., Mohamed, S.K.: Predicting polypharmacy side-effects using knowledge graph embeddings. AMIA Summits on Translational Science Proceedings 2020, 449 (2020)
77. Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., Taylor, J.: Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done. Queue 17(2), 48–75 (2019)
78. Nunes, S., Sousa, R.T., Pesquita, C.: Predicting gene-disease associations with knowledge graph embeddings over multiple ontologies. arXiv preprint arXiv:2105.04944 (2021)
79. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M.: Kegg: Kyoto encyclopedia of genes and genomes. Nucleic acids research 27(1), 29–34 (1999)
80. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., Del-Toro, N., et al.: The mintact project—intact as a common curation platform for 11 molecular interaction databases. Nucleic acids research 42(D1), D358–D363 (2014)
81. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, G.A., Bileschi, M.L., Bork, P., Bridge, A., Colwell, L., et al.: Interpro in 2022. Nucleic Acids Research 51(D1), D418–D427 (2023)
82. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J.X., Jensen, L.J.: Diseases: Text mining and data integration of disease–gene associations. Methods 74, 83–89 (2015)
83. Purkayastha, S., Dana, S., Garg, D., Khandelwal, D., Bhargav, G.: Knowledge graph question answering via sparql silhouette generation. arXiv preprint arXiv:2109.09475 (2021)
84. Qu, X., Cai, Y.: Conve-bio: Knowledge graph embedding for biomedical relation prediction. In: 2023 International Conference on Intelligent Supercomputing and BioPharma (ISBP). pp. 10–13. IEEE (2023)
85. Reese, J.T., Unni, D., Callahan, T.J., Cappelletti, L., Ravanmehr, V., Carbon, S., Shefchek, K.A., Good, B.M., Balhoff, J.P., Fontana, T., et al.: Kg-covid-19: a framework to produce customized knowledge graphs for covid-19 response. Patterns 2(1), 100155 (2021)
86. Ren, Z.H., Yu, C.Q., Li, L.P., You, Z.H., Guan, Y.J., Wang, X.F., Pan, J.: Biodkg–ddi: predicting drug–drug interactions based on drug knowledge graph fusing biochemical information. Briefings in Functional Genomics 21(3), 216–229 (2022)
87. Rizvi, R.F., Vasilakes, J., Adam, T.J., Melton, G.B., Bishop, J.R., Bian, J., Tao, C., Zhang, R.: idisk: the integrated dietary supplements knowledge base. Journal of the American Medical Informatics Association 27(4), 539–548 (2020)
88. Rogers, D., Hahn, M.: Extended-connectivity fingerprints. Journal of chemical information and modeling 50(5), 742–754 (2010)

89. Roy, A., Pan, S.: Incorporating medical knowledge in bert for clinical relation extraction. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 5357–5366 (2021)
90. Santos, A., Tsafou, K., Stolte, C., Pletscher-Frankild, S., O'Donoghue, S.I., Jensen, L.J.: Comprehensive comparison of large-scale tissue expression datasets. PeerJ 3, e1054 (2015)
91. Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., et al.: Ncbi taxonomy: a comprehensive update on curation, resources and tools. Database 2020 (2020)
92. Schomburg, I., Jeske, L., Ulbrich, M., Placzek, S., Chang, A., Schomburg, D.: The brenda enzyme information system–from a database to an expert system. Journal of biotechnology 261, 194–206 (2017)
93. Schriml, L.M., Munro, J.B., Schor, M., Olley, D., McCracken, C., Felix, V., Baron, J.A., Jackson, R., Bello, S.M., Bearer, C., et al.: The human disease ontology 2022 update. Nucleic acids research 50(D1), D1255–D1261 (2022)
94. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.A.: Disease ontology: a backbone for disease semantic integration. Nucleic acids research 40(D1), D940–D946 (2012)
95. Shen, Z., Zhang, Y.H., Han, K., Nandi, A.K., Honig, B., Huang, D.S.: mirna-disease association prediction with collaborative matrix factorization. Complexity 2017 (2017)
96. Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P., Scholes, H.M., Pang, C.S., Woodridge, L., Rauer, C., Sen, N., et al.: Cath: increased structural coverage of functional space. Nucleic acids research 49(D1), D266–D273 (2021)
97. Smith, C.L., Eppig, J.T.: The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdisciplinary Reviews: Systems Biology and Medicine 1(3), 390–399 (2009)
98. Sosa, D.N., Derry, A., Guo, M., Wei, E., Brinton, C., Altman, R.B.: A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020. pp. 463–474. World Scientific (2019)
99. Sterling, T., Irwin, J.J.: Zinc 15–ligand discovery for everyone. Journal of chemical information and modeling 55(11), 2324–2337 (2015)
100. Su, X., You, Z.H., Huang, D.s., Wang, L., Wong, L., Ji, B., Zhao, B.: Biomedical knowledge graph embedding with capsule network for multi-label drug-drug interaction prediction. IEEE Transactions on Knowledge and Data Engineering (2022)
101. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., et al.: The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic acids research 49(D1), D605–D612 (2021)
102. Szklarczyk, D., Santos, A., Von Mering, C., Jensen, L.J., Bork, P., Kuhn, M.: Stitch 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic acids research 44(D1), D380–D384 (2016)
103. Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., Aittokallio, T.: Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. Journal of Chemical Information and Modeling 54(3), 735–743 (2014)
104. Tatonetti, N.P., Ye, P.P., Daneshjou, R., Altman, R.B.: Data-driven prediction of drug effects and interactions. Science translational medicine 4(125), 125ra31–125ra31 (2012)
105. Tweedie, S., Braschi, B., Gray, K., Jones, T.E., Seal, R.L., Yates, B., Bruford, E.A.: Genenames. org: the hgnc and vgnc resources in 2021. Nucleic acids research 49(D1), D939–D946 (2021)
106. Ursu, O., Holmes, J., Bologa, C.G., Yang, J.J., Mathias, S.L., Stathias, V., Nguyen, D.T., Schürer, S., Oprea, T.: Drugcentral 2018: an update. Nucleic acids research 47(D1), D963–D970 (2019)

107. Van Melle, W.: Mycin: a knowledge-based consultation program for infectious disease diagnosis. International journal of man-machine studies 10(3), 313–322 (1978)
108. Vlachos, I.S., Paraskevopoulou, M.D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I.L., Maniou, S., Karathanou, K., Kalfakakou, D., et al.: Diana-tarbase v7. 0: indexing more than half a million experimentally supported mirna: mrna interactions. Nucleic acids research 43(D1), D153–D159 (2015)
109. Wang, D., Gu, J., Wang, T., Ding, Z.: Oncomirdb: a database for the experimentally verified oncogenic and tumor-suppressive micrornas. Bioinformatics 30(15), 2237–2238 (2014)
110. Wang, H., Zhou, G., Liu, S., Jiang, J.Y., Wang, W.: Drug-target interaction prediction with graph attention networks. arXiv preprint arXiv:2107.06099 (2021)
111. Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., et al.: Cord-19: The covid-19 open research dataset. ArXiv (2020)
112. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering 29(12), 2724–2743 (2017)
113. Wang, S., Du, Z., Ding, M., Rodriguez-Paton, A., Song, T.: Kg-dti: a knowledge graph based deep learning method for drug-target interaction predictions and alzheimer's disease drug repositions. Applied Intelligence 52(1), 846–857 (2022)
114. Wang, W., Liang, S., Yu, M., Liu, D., Zhang, H., Wang, X., Zhou, Y.: Gchn-dti: Predicting drug-target interactions by graph convolution on heterogeneous networks. Methods 206, 101–107 (2022)
115. Weinreich, S.S., Mangon, R., Sikkens, J., Teeuw, M.e., Cornel, M.: Orphanet: a european database for rare diseases. Nederlands tijdschrift voor geneeskunde 152(9), 518–519 (2008)
116. Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M.: Drugbank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research 36(suppl_1), D901–D906 (2008)
117. Wu, X., Duan, J., Pan, Y., Li, M.: Medical knowledge graph: Data sources, construction, reasoning, and applications. Big Data Mining and Analytics 6(2), 201–217 (2023)
118. Wu, Y., Gao, M., Zeng, M., Zhang, J., Li, M.: Bridgedpi: a novel graph neural network for predicting drug–protein interactions. Bioinformatics 38(9), 2571–2578 (2022)
119. Xiang, Y., Zhang, Z., Chen, J., Chen, X., Lin, Z., Zheng, Y.: Ontoea: Ontology-guided entity alignment via joint knowledge graph embedding. arXiv preprint arXiv:2105.07688 (2021)
120. Xie, B., Ding, Q., Han, H., Wu, D.: mircancer: a microrna–cancer association database constructed by text mining on literature. Bioinformatics 29(5), 638–644 (2013)
121. Xing, X., Yang, F., Li, H., Zhang, J., Zhao, Y., Gao, M., Huang, J., Yao, J.: Multi-level attention graph neural network based on co-expression gene modules for disease diagnosis and prognosis. Bioinformatics 38(8), 2178–2186 (2022)
122. Xiong, Z., Huang, F., Wang, Z., Liu, S., Zhang, W.: A multimodal framework for improving in silico drug repositioning with the prior knowledge from knowledge graphs. IEEE/ACM Transactions on Computational Biology and Bioinformatics 19(5), 2623–2631 (2021)
123. Yan, S.: Memory-aligned knowledge graph for clinically accurate radiology image report generation. In: Proceedings of the 21st Workshop on Biomedical Language Processing. pp. 116–122 (2022)
124. Yang, B., Yih, W.t., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575 (2014)
125. Yang, Y., Cao, Z., Zhao, P., Zeng, D.D., Zhang, Q., Luo, Y.: Constructing public health evidence knowledge graph for decision-making support from covid-19 literature of modelling study. Journal of Safety Science and Resilience 2(3), 146–156 (2021)
126. Zhang, W., Paudel, B., Zhang, W., Bernstein, A., Chen, H.: Interaction embeddings for prediction and explanation in knowledge graphs. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. pp. 96–104 (2019)

127. Zhang, X.M., Liang, L., Liu, L., Tang, M.J.: Graph neural networks and their current applications in bioinformatics. Frontiers in genetics 12, 690049 (2021)
128. Zhang, X., Che, C.: Drug repurposing for parkinson's disease by integrating knowledge graph completion model and knowledge fusion of medical literature. Future Internet 13(1), 14 (2021)
129. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12910–12917 (2020)
130. Zheng, S., Rao, J., Song, Y., Zhang, J., Xiao, X., Fang, E.F., Yang, Y., Niu, Z.: Pharmkg: a dedicated knowledge graph benchmark for bomedical data mining. Briefings in bioinformatics 22(4), bbaa344 (2021)
131. Zhu, C., Yang, Z., Xia, X., Li, N., Zhong, F., Liu, L.: Multimodal reasoning based on knowledge graph embedding for specific diseases. Bioinformatics 38(8), 2235–2245 (2022)
132. Zhu, Y., Elemento, O., Pathak, J., Wang, F.: Drug knowledge bases and their applications in biomedical informatics research. Briefings in bioinformatics 20(4), 1308–1321 (2019)
133. Zitnik, M., Agrawal, M., Leskovec, J.: Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 34(13), i457–i466 (2018)
134. Zou, X.: A survey on application of knowledge graph. In: Journal of Physics: Conference Series. vol. 1487, p. 012016. IOP Publishing (2020)

Ylenia Galluzzo is PhD Student in Information Communication Technology (ICT) at the University of Palermo since November 2022. My main research interests are on Knowledge Graphs, Big Data Analytics, ML, Deep Learning and Bioinformatics. I deal with methodologies for modelling and analysing data through complex networks, in the field of Big Data. Previously, I worked in the industrial sector in an R&D team where I developed skills in the field of: Natural Language Processing (NLP), Anomaly Detection and methods Supervised and Unsupervised for Topic Modeling.

Received: June 01, 2023; Accepted: May 02, 2024.

Computer Science and Information Systems 21(3):1077–1095 https://doi.org/10.2298/CSIS230709028S

Operational Semantics in a Domain-Specific Robot Control Language: a Pedagogical Use Case*

William Steingartner and Valerie Novitzká

Technical University of Košice, Faculty of Electrical Engineering and Informatics
Košice, Slovakia
{william.steingartner, valerie.novitzka}@tuke.sk

Abstract. In this paper, we focus on some aspects of structural operational semantics for a selected domain-specific language for robot control. After defining a syntax for two independent languages for control of a robot's movements, we define a structural operational semantics for them. The integrated part of structural operational semantics is abstract implementation, which consists of defining abstract machine and transformation of a programming language to abstract machine instructions. The achieved results as well as the mentioned research are a part of the research in the field of semantic methods, where we focus on the formalization of semantic methods for software engineering. This area is also very important for the training of students and young IT experts because the semantic methods can help to understand program behavior and detect errors in program design. To make the teaching of formal semantics in the field of domain-specific languages more attractive, we have also prepared an application that serves to visualize the individual steps of the program on an abstract machine – simulation of translated code with visualization of a robot's movement.

Keywords: Abstract machine, containerization, domain-specific language, formal semantics, micro-service, online teaching, teaching software, university didactic.

1. Introduction

The main aim of this paper is to define an abstract machine and abstract implementation as a part of structural operational semantics for a simple domain-specific language. After that, we design an application that serves as an emulation tool of the defined abstract machine. For its design and implementation, a modular approach was taken. Our approach is primarily dedicated to the students visiting our graduate course Semantics of Programming Languages.

Domain-specific languages (DSLs) constitute a new class of programming languages and they can be characterized as a new programming paradigm. These languages are designed as special-purpose programming languages with a higher level of abstraction to support a particular set of tasks, meant to be used in the contexts of specific domains [2]. DSLs are defined in [8] as programming languages or executable specification languages that offer expressive power focused on, and usually are restricted to particular problem domains. At present, there are several DSLs for various domains: domain-specific functional

* This paper is an extended version of the conference paper "Abstract machine for operational semantics of domain-specific language" presented in a workshop 4th Workshop on Modern Approaches in Data Engineering and Information System Design organized in conjunction with the conference ADBIS 2022.

languages for dynamic geometry [28], for finance (for instance Marlowe [17] which allows users to create blockchain applications tailored for financial contracts, or Findel [3], designed for financial derivatives and boasts simplicity and expressiveness for complex derivative expressions); legislative, and juridical applications (e.g. ERGO [29] which focuses on legal contract execution logic, aiming for consistency between legal prose and executable logic, accessible to both lawyers and developers); for asset management (e.g. Psamathe [21]); for multidomain (e.g. Jabuti [9] catering the smart contracts in application integration, offering natural language-like constructors, ContextMapper [15] grounded in Domain-Driven Design principles aiding in context mapping and service decomposition, Adico [11] serving as the interface between human-readable institutional specifications and machine-readable Solidity contracts (referring to Ethereum, see [41]), Archetype [40] focused on developing smart contracts on the Tezos blockchain, with a specific focus on the formal verification of the contract, ink! [22] as a programming language dedicated to smart contracts (referring to Polkadot, see [42]), and iContractML [12] facilitating modeling and deploying smart contracts across multiple blockchain platforms); for smart contract formal analysis and verification (VeriSolid, Scilla), for modeling platform independent model specifications of a specific information system [18], for the development of software agents [6] or specifically for modeling application-specific functionalities of business applications [26] and many others. For more interested user, we recommend more comprehensive survey [1] in blockchain model. Because many of mentioned problem domains require deep knowledge of their principles and needs, they are not appropriate for teaching students because many details are not needed for explaining how to define the semantics of DSLs. Therefore we decided to use a simple DSL language Robot that is suitable for our purposes. We note that the first idea about this robot language was presented in [25].

A formal definition of each programming language consists of a formal syntax and formal semantics [16]. The formal syntax can be concrete and abstract. Generally, the concrete syntax is the set of rules that defines the combinations of symbols that are considered to be well-formed in a computer language. Abstract syntax defines the structure of syntactic entities without some details, and it is needed for defining the formal semantics of a language. Under semantics, we understand the meaning of such structures.

Formal semantics is an integral part of programming language's formal definition. It offers students studying computer science and IT experts to understand the meaning of the programs and the opportunity to better understand how the code execution works on the machine using abstraction, thus removing some unnecessary details which can be ignored.

Based on the above, this suggests that strong support for formal methods and principles in the field of software engineering and software development at the level of university education will help to acquire critical thinking and support abstraction in solving problems, whether from the point of view of designing more effective algorithms, verifying the properties of languages programs written in this language and thus prevent many (logical) errors in various phases of the design and development of software systems. Since all formal methods are rooted in formal semantics, we are also of the opinion that interactive support in the teaching of semantic methods will make it easier to learn the principles of these methods, as well as easier to navigate in the choice of a suitable semantic method, or to correctly interpret the results achieved at all levels of derivation. There-

fore, visualization software is one of the interactive aids that we present in an effort to modernize and make teaching more attractive. Orientation to domain-specific languages will make it easier for students to understand the principles that are standardly presented in imperative and functional languages. Students will more easily understand how these paradigms are related and how individual methods are modified when used across different paradigms. We would like to add that, from a practical point of view, support is still available when using the SLANG [31] tool. It is a software for the rapid prototyping of a programming language from its formal specification; which also allows belief in its semantic specification (currently support for denotational and natural operational semantics).

The students of our course become familiar with several semantic methods, such as natural semantics, denotational semantics, algebraic semantics, and axiomatic semantics. Natural semantics, or the semantics of big steps is the simplest kind of semantic method and is well understandable by students. Denotational semantics define the meaning of programs by functions and it does not keep track of execution details. Algebraic semantics specifies programs as a hierarchy of abstract data types and defines the semantics as heterogeneous algebras. Axiomatic semantics is suitable for verification purposes by defining preconditions and postconditions before and after executing the statements, respectively. One of the most popular semantics methods is structural operational semantics [20], also called the semantics of small steps, which defines how individual steps of programs are executed. This method requires to define also abstract implementation based on an abstract machine and we concentrate on this method in our paper.

Knowledge of structural operational semantics and the appropriate abstract machine is an important foundation in the teaching of programming languages. It offers students the opportunity to better understand how code execution works on the machine using abstraction, thus removing some unnecessary details which can be ignored. However, the related abstract machine theory can be unfriendly or difficult for students to learn.

To overcome this problem, we follow the idea of making formal semantic methods visual in particular steps of calculations. The current situation requires that students also acquire skills in working with this type of educational software, as stated by the author in [32]. In educating young IT experts, there is a need for tools that can illustrate the semantic methods to students and make it clearer how for example an abstract machine operates interactively for not only self-study for students but also for teachers to present abstract machine in their lectures. Many times, the introduction to programming is explained in languages similar to the original idea of Karel the Robot (or simply *Robot*). Therefore, we were also inspired by this approach and expanded our research in the field of domain-specific languages, where we work with a language for robot control. The domain-specific language we work with is the language Robot which is an entity moving in two-dimensional space by executing given commands.

The abstract implementation of a robotic language represents a natural continuation in the field of semantic methods, leveraging structural operational semantics for its definition. It is also a great motivation for us, as other semantic approaches are already formulated and derived: denotational and operational semantics are defined in [13], we defined natural semantics in [36] and we processed its visualization in [38]. It is also natural when extending the spectrum of semantic methods to verify that the new method is fully equivalent to the existing ones and that it provides the same results. Then we have in mind a

deterministic approach in the application of semantic methods. We formulated the proof of equivalence between the denotational semantics of the Robot language and the natural semantics defined by us in [37].

In this article, we present a practical approach to the visualization of the semantic method for a given DSL language, while the results achieved here are based on our previously mentioned works. In this way, we have, on the one hand, extended the field of semantic methods for a given language and thus given the user the opportunity to choose a method according to the pursued goal (only the result, detailed calculation procedure, verification of language properties, etc.), and on the other hand, we have also provided a visualization module that will enable the interactive use of the mentioned method, tracking the progress of the calculation or checking the results.

The aim of this paper is to define an abstract machine for the structural operational semantics of the robot language and for its abstract implementation. After that, we designed an application that serves as an emulation tool of abstract machine defined. For its design and implementation, a modular approach was taken. The application is divided into four micro-services: translation logic, translation graphical user interface, simulation logic, and simulation graphical user interface, with a database in the separate (fifth) container.

This paper is a part of our project to prepare young engineers and IT experts to deal with semantic methods. We have prepared the applications for teaching and learning denotational semantics [35], operational semantics [34] and abstract machine [33]. Our needs are so specific that we had problems finding similar applications. We followed the methods published in [30] with some modifications. The aim of our project is to provide an integrated system for learning and teaching semantic methods. This paper serves for illustrating how domain-specific languages differ in using operational semantics.

The purpose of the logic components is to parse the given input source code and to provide an output form (bytecode) for visual processing and calculation. The processed input for the translation logic is afterwards translated to abstract machine code and for the simulation logic, the data is used for simulating the execution on the abstract machine, returning simulation data. The web is used as a graphical user interface for both components using Spring with Thymeleaf template engine for displaying variable data on the HTML page which uses JavaScript scripts for additional calls. Web pages have both English and Slovak localization with the link from the compiler to the simulator and back.

The paper is structured as follows: in Section 2, we present a syntax of the language for robot and preliminaries for defining the semantics. Section 3 contains a definition of the abstract machine(s) for the robot language(s). Section 4 focuses on the design of an emulation of an abstract machine, and Section 5 presents its specification and functionality. Finally, Section 6 concludes our paper.

2. Simple Domain-Specific Language for Robot

A domain-specific language (DSL) is a programming language or executable specification language that, through appropriate notation and abstraction, offers expressive power focused on, and usually limited to, a specific domain of the problem [8]. Domain-specific languages are languages tailored to a specific application domain. They offer substantial gains in expressiveness and ease of use compared with general-purpose programming languages in their domain of application [19]. With the creation of a DSL, a user can focus

on the problem and its solution in a specific area for further processing. One of the possible benefits of using DSL is lower implementation time, reduced maintenance costs, as well as better portability, reliability, optimizability and testability [7]. The language of the robot is an external DSL which means it is a fully independent language with its syntax and semantics [14]. Language workbench is a great tool for creating such a language – it is easy to define not the only parser, but also a custom editing environment [10].

The operational semantics of the language of a robot is a base for abstract machine definition. The robot is an entity in two-dimensional space, a grid, with specified x (horizontal) and y (vertical) coordinates which are typically changed during the execution of a program. The robot can move horizontally and vertically, diagonal movement is not possible in this simplified version. Language has two syntactic domains [13]:

- $n \in \mathbf{Num}$ – numerals specifying number of steps which robot has to move;
- $C \in \mathbf{Comm}$ – commands.

The syntactic domain is specified by an abstract syntax giving the structure of the basic elements and the composite ones. The syntactic domain $\mathbf{Num}$ has from the point of view of abstract syntax no internal structure (but syntactically numerals can be represented with a regular expression $[0, \ldots, 9]^+$). We need to define abstract syntax only for syntactic domain $\mathbf{Comm}$ for commands. The structure of well-structured commands we define using BNF formalism by production rule as follows:

$$C ::= \textbf{left} \mid \textbf{right} \mid \textbf{up} \mid \textbf{down} \mid \textbf{left}\ n \mid \textbf{right}\ n \mid \textbf{up}\ n \mid \textbf{down}\ n \mid \textbf{skip} \mid \textbf{reset} \mid C; C.$$

The commands **left** and **right** express the movement of a robot one position to the left or to the right, respectively. The commands **up** and **down** express the movement of a robot tothe up or to the left, respectively. The commands **left** n, **right** n, **up** n and **down** n express the particular movements by n position in the specified direction. The **reset** statement invokes an immediate return to the initial, or default, position set by the user at the beginning (akin to *teleportation*). Next, the **skip** statement is void of action, serving primarily as a placeholder in proofs of semantic equivalence. The composition (sequence) of commands is expressed by the last pattern.

For the robot's position, we define the semantic domain

$$\mathbf{Point} = \mathbb{Z} \times \mathbb{Z}$$

with the elements p that are considered states. Each position has a form

$$p = (x, y)$$

where x and y are the coordinates. The change in the position of the robot is considered a change of state. For simplification, we consider that our grid is not limited in any direction.

Example 1: Assume a simple program of robot:

right 2;
up 2;
right 3

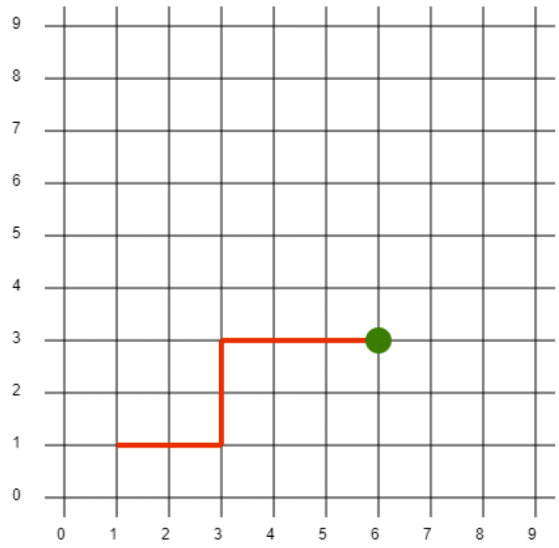


Fig. 1. Movement of a robot

A robot moves two positions right, then two positions up and finally three positions right. The robot movement according to this program is illustrated as a path along the orthogonal grid (Fig. 1). The position

$$p^* = (\mathbf{1}, \mathbf{1})$$

is the starting position of a robot, i.e. the coordinates are $x = \mathbf{1}$ and $y = \mathbf{1}$.

As we showed in [37], for natural semantics of n-step statements, we can use direct evaluation, rewriting or inductive definition. Using the first approach, the semantics of the program is evaluated by the following derivation tree:

$$\dfrac{\dfrac{}{\langle \mathbf{right}\ 2, p^* \rangle \to p_1} \qquad \dfrac{\dfrac{}{\langle \mathbf{up}\ 2, p_1 \rangle \to p_2} \quad \dfrac{}{\langle \mathbf{right}\ 3, p_2 \rangle \to p}}{\langle \mathbf{up}\ 2; \mathbf{right}\ 3, p_1 \rangle \to p}}{\langle \mathbf{right}\ 2; \mathbf{up}\ 2; \mathbf{right}\ 3, p^* \rangle \to p}$$

and the internal states are: $p_1 = (\mathbf{3}, \mathbf{1})$ and $p_2 = (\mathbf{3}, \mathbf{3})$. After performing all steps, the robot ends up at the position $p = (\mathbf{6}, \mathbf{3})$.

□

The semantics of the commands is given by evaluation of the semantic function

$$\mathscr{C} : \mathbf{Comm} \to (\mathbf{Point} \to \mathbf{Point}).$$

A structural operational semantics is known as a small-step semantics. That means each step is defined by a transition relation between configurations. A configuration is in general a pair

$$\langle C, p \rangle,$$

and it describes a simple step of execution of a command C from a state p. The transition has the form

$$\langle C, p\rangle \Rightarrow \alpha,$$

where α is a configuration and it can be either:

- a new state p', if the command C is performed in one step. A typical case is a movement in some direction in one step or the teleportation to the initial position. That means, the execution has terminated in a state p', or
- an intermediate configuration $\langle C', p'\rangle$, if the execution of command C is not completed and remaining computation continues.

Assume a state $p = (x, y)$. The transitions for the commands executed in one step are:

$$\begin{aligned}
&\langle \mathbf{left}, (x, y)\rangle &&\Rightarrow (x \ominus \mathbf{1}, y),\\
&\langle \mathbf{right}, (x, y)\rangle &&\Rightarrow (x \oplus \mathbf{1}, y),\\
&\langle \mathbf{up}, (x, y)\rangle &&\Rightarrow (x, y \oplus \mathbf{1}),\\
&\langle \mathbf{left}, (x, y)\rangle &&\Rightarrow (x, y \ominus \mathbf{1}).
\end{aligned}$$

Each command executing more movements can be considered as a sequence of commands:

$$\begin{aligned}
&\mathbf{left}\ n = \mathbf{left}; \mathbf{left}\ m,\\
&\mathbf{right}\ n = \mathbf{right}; \mathbf{right}\ m,\\
&\mathbf{up}\ n = \mathbf{up}; \mathbf{up}\ m,\\
&\mathbf{down}\ n = \mathbf{down}; \mathbf{down}\ m,
\end{aligned}$$

where $[\![m]\!] = [\![n]\!] \ominus \mathbf{1}$.

The commands that are not executed in one step have transitions expressing the first step of execution:

$$\begin{aligned}
&\langle \mathbf{left}\ n, (x, y)\rangle &&\Rightarrow \langle \mathbf{left}\ m, (x \ominus \mathbf{1}, y)\rangle,\\
&\langle \mathbf{right}\ n, (x, y)\rangle &&\Rightarrow \langle \mathbf{right}\ m, (x \oplus \mathbf{1}, y)\rangle,\\
&\langle \mathbf{up}\ n, (x, y)\rangle &&\Rightarrow \langle \mathbf{up}\ m, (x, y \oplus \mathbf{1})\rangle,\\
&\langle \mathbf{down}\ n, (x, y)\rangle &&\Rightarrow \langle \mathbf{down}\ m, (x, y \ominus \mathbf{1})\rangle,
\end{aligned}$$

where $[\![n]\!] = [\![m]\!] \oplus \mathbf{1}$, for $[\![m]\!], [\![n]\!] \in \mathbb{N}_0$ and $[\![\cdot]\!]$ is a semantic function that sends numerals to naturals (or zero) [13]:

$$[\![\cdot]\!] : \mathbf{Num} \to \mathbb{N}_0.$$

The last two commands of a language have the following transitions:

$$\begin{aligned}
&\langle \mathbf{skip}, p\rangle &&\Rightarrow p,\\
&\langle \mathbf{reset}, p\rangle &&\Rightarrow p^*,
\end{aligned}$$

where p^* is the starting position and its coordinates are stated in beginning of a program.

To make this language more general (closer to real languages), we change its abstract syntax by defining new commands for turning into the given direction according to the

angle. Because the robot moves on the grid, the angles are multiples of 90 degrees and they define movement in the current heading. We define the abstract syntax of this language as follows:

$$C ::= \mathbf{forward} \mid \mathbf{forward}\ n \mid \mathbf{turn\ left} \mid \mathbf{turn\ right} \mid \\ \mathbf{skip} \mid \mathbf{reset} \mid C; C.$$

These commands enable the robot to rotate around its axis. The command $\mathbf{forward}$ causes the robot moves one step in the direction of the current heading. The command $\mathbf{forward}\ n$ defines that robot moves n steps in the direction of the current heading. The rotation of movement defines the commands $\mathbf{turn\ left}$, which changes the direction of movement to the left, and $\mathbf{turn\ right}$, which changes the direction of movement to the right according to the direction of the current heading. The commands $\mathbf{skip}$ and $\mathbf{reset}$ work as in the previous version of a language. The last item in production rule $C; C$ is the concatenation of commands. We define a new semantic domain for the angles

$$\mathbf{Angle} = \mathbb{Z}$$

with the elements $a \in \mathbf{Angle}$. We consider the following convention: if the robot is facing north, then the current angle is $\mathbf{0}$, for east it is $\mathbf{90}$, for south $\mathbf{180}$ and for west $\mathbf{270}$. The $\mathbf{forward}$ command calculates the next position by trigonometric functions sine and cosine. For the potential input values of the angle, these functions respectively yield values of $-\mathbf{1}$, $\mathbf{0}$, or $\mathbf{1}$.

Because an angle is an element that provides the direction of the robot, we need to modify the notion of state. A new semantic domain $\mathbf{State}$ is a product:

$$\mathbf{State} = \mathbf{Point} \times \mathbf{Angle}$$

with the elements $(p, a) \in \mathbf{State}$, $p \in \mathbf{Point}$ and $a \in \mathbf{Angle}$. Then a state is a tuple and has the following form

$$\langle C, p, a \rangle.$$

Now we have two groups of commands:

- the first group contains commands $\mathbf{left}$, $\mathbf{right}$, $\mathbf{up}$, $\mathbf{down}$, $\mathbf{left}\ n$, $\mathbf{right}\ n$, $\mathbf{up}\ n$ and $\mathbf{down}\ n$;
- the second group contains the commands enabling rotation, $\mathbf{forward}$, $\mathbf{forward}\ n$, $\mathbf{turn\ left}$ and $\mathbf{turn\ right}$;
- for te second group of statements, we need to extend (redefine) also an initial state by adding an angle, $s^* = (p^*, \mathbf{0})$;
- the commands $\mathbf{skip}$, $\mathbf{reset}$ and $C; C$ belong to both groups of commands.

Commands from different groups cannot be combined in one program, because of different definitions of configurations. But they can be executed on the same abstract machine – for this reason we create two ANTLR grammars during implementation, so it is possible to apply one implementation of abstract machine emulation to several input language specifications.

The semantic function for this language is defined as

$$\mathscr{C}' : \mathbf{Comm} \to \mathbf{State} \to \mathbf{State},$$

and is defined by the following transitions:

$$\begin{aligned}
&\langle \mathbf{forward}, (x, y), a\rangle &&\Rightarrow ((x \oplus \sin a, y \oplus \cos a), a),\\
&\langle \mathbf{forward}\ n, (x, y), a\rangle &&\Rightarrow ((x \oplus [\![n]\!] \otimes \sin a, y \oplus [\![n]\!] \otimes \cos a), a),\\
&\langle \mathbf{turn\ left}, p, a\rangle &&\Rightarrow (p, (a \oplus \mathbf{270}) \bmod \mathbf{360}),\\
&\langle \mathbf{turn\ right}, p, a\rangle &&\Rightarrow (p, (a \oplus \mathbf{90}) \bmod \mathbf{360}).
\end{aligned}$$

We also need to redefine the commands **reset** and **skip**:

$$\begin{aligned}
&\langle \mathbf{reset}, (p, a)\rangle \Rightarrow (p^*, \mathbf{0}),\\
&\langle \mathbf{skip}, (p, a)\rangle \ \Rightarrow (p, a).
\end{aligned}$$

Both defined languages can be extended by other commands, for instance, $\mathbf{init}(x, y)$, which can initialize the starting position of the robot. Another useful extension can be conditional command and loop command that can serve for shortening the code and opening a lot of new options. Another extension can be an addition of flags on the environment which the robot can place and lift from. Furthermore, battery or energy management can be added, so each robot's movements deplete its energy which needs to be recharged later on [13]. The mentioned extensions are the aims for future work.

3. Abstract Machine for DSL – Theory and Development of Method

Abstract implementation of a program is used to verify the partial correctness of the implementation of programming languages and consists of

- the definition of an abstract machine;
- the translation of the program into a sequence of the instructions of an abstract machine;
- the execution of code of abstract machine and comparison with a structural operational semantics of a program.

Definition of the abstract machine [20] consists of

- the syntax of instructions;
- the semantics of instructions.

We define the new syntactic domains $\mathbf{Code}$ for a sequence of instructions and $\mathbf{Ins}$ for the instructions. An element $c \in \mathbf{Code}$ is a sequence of instructions. We define the abstract syntax of instructions and code by the following production rules:

$$\begin{aligned}
ins ::= {} & \mathbf{LEFT} \mid \mathbf{LEFT}\text{-}n \mid \mathbf{RIGHT} \mid \mathbf{RIGHT}\text{-}n \mid\\
& \mathbf{UP} \mid \mathbf{UP}\text{-}n \mid \mathbf{DOWN} \mid \mathbf{DOWN}\text{-}n \mid\\
& \mathbf{FORWARD} \mid \mathbf{FORWARD}\text{-}n \mid \mathbf{TLEFT} \mid \mathbf{TRIGHT} \mid\\
& \mathbf{SKIP} \mid \mathbf{RESET},
\end{aligned}$$

$$c ::= \quad \varepsilon \mid ins : c.$$

Our definition of abstract machine for structural operational semantics is defined for the domain-specific language of the robot, namely for both versions of the language. The state of the abstract machine determines the position of the robot in the environment

$$p \in \mathbf{Point} = \mathbb{Z} \times \mathbb{Z}$$

and possibly an angle of its rotation

$$a \in \mathbf{Angle} = \mathbb{Z}.$$

Hence, for the first version of the language containing the directional commands, the state is an element of the semantics domain $\mathbf{Point}$. For the second version of the language, comprising the rotations, a state is an element of the semantic domain

$$s \in \mathbf{State} = \mathbf{Point} \times \mathbf{Angle}.$$

Evaluation stack is not used in this abstract machine(s) but will be needed for possible extensions of robot language.

Semantic of instructions is defining using configurations, which can be written as triples

$$\langle c, v, s \rangle,$$

where c is the code, a sequence of instructions, v is an evaluation stack of values and s is the state of the abstract machine.

Some of the instructions contain numeric parameters which represent the number of steps in a given direction (it simply means that command is performed $\mathbf{n}$-times, where $\mathbf{n} = [\![n]\!]$). For situations, when numeral n is converted to boundary values ($\mathbf{0}$) or not supported values (negative numbers), an abstract machine (and in a simulation program, as well) does not do anything. Hence, the compilation of such commands provides an error.

From the semantic equivalence (proved for natural semantics in [36]) it follows that (for any parametric instruction, denoted $\mathbf{INST}$-n):

$$\begin{aligned} &\langle \mathbf{INST}\text{-}n, v, p \rangle \Rightarrow\!\!\!\!= \langle \mathbf{INST}; \mathbf{INST}\text{-}m, v, p \rangle \\ &\langle \mathbf{INST}\text{-}1, v, p \rangle \Rightarrow\!\!\!\!= \langle \mathbf{INST}, v, p \rangle \end{aligned}$$

for $[\![n]\!] = [\![m]\!] \oplus \mathbf{1}$, $[\![n]\!], [\![m]\!] \in \mathbb{N}_0$. We note that all instructions with the value $\mathbf{0}$ of the parameter don't do anything and they are (semantically) equivalent to the instruction $\mathbf{SKIP}$.

The semantics of abstract machine instructions (for the language with directional commands) expresses how a configuration is changed after execution of them. We define the semantics of instructions as follows:

$$\begin{aligned} &\langle \mathbf{LEFT} : c, \varepsilon, (x, y) \rangle \Rightarrow\!\!\!\!= \langle c, \varepsilon, (x \ominus \mathbf{1}, y) \rangle, \\ &\langle \mathbf{LEFT}\text{-}n : c, \varepsilon, (x, y) \rangle \Rightarrow\!\!\!\!= \langle \mathbf{LEFT}\text{-}m : c, \varepsilon, (x \ominus \mathbf{1}, y) \rangle, \\ &\langle \mathbf{RIGHT} : c, \varepsilon, (x, y) \rangle \Rightarrow\!\!\!\!= \langle c, \varepsilon, (x \oplus \mathbf{1}, y) \rangle, \\ &\langle \mathbf{RIGHT}\text{-}n : c, \varepsilon, (x, y) \rangle \Rightarrow\!\!\!\!= \langle \mathbf{RIGHT}\text{-}m : c, \varepsilon, (x \oplus \mathbf{1}, y) \rangle, \\ &\langle \mathbf{UP} : c, \varepsilon, (x, y) \rangle \Rightarrow\!\!\!\!= \langle c, \varepsilon, (x, y \oplus \mathbf{1}) \rangle, \\ &\langle \mathbf{UP}\text{-}n : c, \varepsilon, (x, y) \rangle \Rightarrow\!\!\!\!= \langle \mathbf{UP}\text{-}m : c, \varepsilon, (x, y \oplus \mathbf{1}) \rangle, \\ &\langle \mathbf{DOWN} : c, \varepsilon, (x, y) \rangle \Rightarrow\!\!\!\!= \langle c, \varepsilon, (x, y \oplus \mathbf{1}) \rangle, \\ &\langle \mathbf{DOWN}\text{-}n : c, \varepsilon, (x, y) \rangle \Rightarrow\!\!\!\!= \langle \mathbf{DOWN}\text{-}m : c, \varepsilon, (x, y \ominus \mathbf{1}) \rangle, \\ &\langle \mathbf{RESET} : c, \varepsilon, (x, y) \rangle \Rightarrow\!\!\!\!= \langle c, \varepsilon, p^* \rangle, \\ &\langle \mathbf{SKIP} : c, \varepsilon, (x, y) \rangle \Rightarrow\!\!\!\!= \langle c, \varepsilon, (x, y) \rangle, \end{aligned}$$

for $[\![n]\!] = [\![m]\!] \oplus \mathbf{1}$, $[\![n]\!], [\![m]\!] \in \mathbb{N}_0$ and p^* stands for an initial (starting) position defined by user specification. Instructions defining the movement of $\mathbf{n}$ steps are performed according to the principles of structural operational semantics the first movement and the remaining code contains the same instruction for $\mathbf{n} \ominus \mathbf{1}$ steps.

Similarly, the semantics of abstract machine instructions (for the language with rotations) is defined as follows:

$$\begin{aligned}
&\langle \mathbf{FORWARD} : c, \varepsilon, ((x, y), a)\rangle \Rightarrow\!\!\!> \langle c, \varepsilon, ((x \oplus \sin a, y \oplus \cos a), a)\rangle,\\
&\langle \mathbf{TLEFT} : c, \varepsilon, ((x, y), a)\rangle \Rightarrow\!\!\!> \langle c, \varepsilon, (p, (a \oplus \mathbf{270}) \bmod \mathbf{360})\rangle,\\
&\langle \mathbf{TRIGHT} : c, \varepsilon, ((x, y), a)\rangle \Rightarrow\!\!\!> \langle c, \varepsilon, ((a \oplus \mathbf{90}) \bmod \mathbf{360})\rangle,\\
&\langle \mathbf{RESET} : c, \varepsilon, ((x, y), a)\rangle \Rightarrow\!\!\!> \langle c, \varepsilon, (p^*, \mathbf{0})\rangle,\\
&\langle \mathbf{SKIP} : c, \varepsilon, (p, a)\rangle \Rightarrow\!\!\!> \langle c, \varepsilon, (p, a)\rangle,
\end{aligned}$$

The transition for moving n steps is defined similarly. We note, that in this specification, an initial configuration (state) is a tuple consisting of the position p^* and the angle $\mathbf{0}$. The initial value for the angle can be changed according to the user specification.

Generating of abstract machine code from the input language is done by the translation function

$$\mathscr{TC} : \mathbf{Comm} \to \mathbf{Code}$$

which sends an input source code written in the robot language into a sequence of instructions of abstract machine. Although we defined two forms of robot language, only one translation function for our purposes is defined:

$$\begin{aligned}
&\mathscr{TC}[\![\mathbf{left}]\!] = \mathbf{LEFT} && \mathscr{TC}[\![\mathbf{left}\ n]\!] = \mathbf{LEFT}\text{-}n\\
&\mathscr{TC}[\![\mathbf{right}]\!] = \mathbf{RIGHT} && \mathscr{TC}[\![\mathbf{right}\ n]\!] = \mathbf{RIGHT}\text{-}n\\
&\mathscr{TC}[\![\mathbf{up}]\!] = \mathbf{UP} && \mathscr{TC}[\![\mathbf{up}\ n]\!] = \mathbf{UP}\text{-}n\\
&\mathscr{TC}[\![\mathbf{down}]\!] = \mathbf{DOWN} && \mathscr{TC}[\![\mathbf{down}\ n]\!] = \mathbf{DOWN}\text{-}n\\
&\mathscr{TC}[\![\mathbf{turn\ left}]\!] = \mathbf{TLEFT} && \mathscr{TC}[\![\mathbf{turn\ right}]\!] = \mathbf{TRIGHT}\\
&\mathscr{TC}[\![\mathbf{forward}]\!] = \mathbf{FORWARD} && \mathscr{TC}[\![\mathbf{forward}\ n]\!] = \mathbf{FORWARD}\text{-}n\\
&\mathscr{TC}[\![\mathbf{reset}]\!] = \mathbf{RESET} && \mathscr{TC}[\![\mathbf{skip}]\!] = \mathbf{SKIP}\\
&\mathscr{TC}[\![C_1; C_2]\!] = \mathscr{TC}[\![C_1]\!] : \mathscr{TC}[\![C_2]\!]
\end{aligned}$$

The colon symbol ":" serves as an instruction separator delineating distinct instructions (syntactic elements) in the abstract machine instruction sequence listing.

For example, let

$$\begin{aligned}
&\mathbf{forward}\ 2; \mathbf{turn\ right}; \mathbf{forward}; \mathbf{turn\ right};\\
&\mathbf{forward}\ 1; \mathbf{turn\ left}; \mathbf{forward}\ 4;
\end{aligned}$$

be a correct program in robot language. A code of abstract machine for this program is the following:

$$\begin{aligned}
&\mathbf{FORWARD}\text{-}2 : \mathbf{TRIGHT} : \mathbf{FORWARD} : \mathbf{TRIGHT} :\\
&\mathbf{FORWARD}\text{-}1 : \mathbf{TLEFT} : \mathbf{FORWARD}\text{-}4
\end{aligned}$$

As the next step, we define the meaning of the abstract machine code using a partially defined execution function:

$$\mathscr{M} : \mathbf{Code} \to \mathbf{Point} \rightharpoonup \mathbf{Point},$$

defined as follows:

$$\mathscr{M}[\![\, c \,]\!]p = \begin{cases} p', & \text{if } \langle c, \varepsilon, p\rangle \Rightarrow^* \langle \varepsilon, \varepsilon, p'\rangle, \\ \bot, & \text{otherwise.} \end{cases}$$

A symbol $\Rightarrow^*$ represents a finite number of steps in transition relation.

By composing the translation and execution functions, we get the semantic function for the abstract machine:

$$\mathscr{C} : \mathbf{Comm} \to \mathbf{Point} \rightharpoonup \mathbf{Point},$$
$$\mathscr{C}[\![\, c \,]\!] = (\mathscr{M} \circ \mathscr{T}\mathscr{C})\,[\![\, c \,]\!].$$

This definition applies to the language version with directional commands. We note that the definition of the execution ($\mathscr{M}'$) and semantic function ($\mathscr{C}'$) for the second types of language variants (with rotations) are analogous, taking into account the semantic area:

$$\mathscr{M}'[\![\, c \,]\!]s = \begin{cases} s', & \text{if } \langle c, \varepsilon, s\rangle \Rightarrow^* \langle \varepsilon, \varepsilon, s'\rangle, \\ \bot, & \text{otherwise.} \end{cases}$$

$$\mathscr{C}' : \mathbf{Comm} \to \mathbf{State} \rightharpoonup \mathbf{State},$$
$$\mathscr{C}'[\![\, c \,]\!] = (\mathscr{M}' \circ \mathscr{T}\mathscr{C})\,[\![\, c \,]\!].$$

4. Application Design

We followed the idea of containerizing the application. We based on the methodological design, where we separated the individual layers (front-end, back-end) and implemented each functional unit as a separate Docker module. Docker is an open-source containerization platform which enables developers to package applications into containers and offers isolation of applications into containers which run safely [5] and independently thanks to Docker Engine directly on the host computer's operating system [27]. Docker containers also offer scalability and portability [24], so the system is easy to extend or change, and reusability of containers in other systems thereby achieving reproducible research [4]. The application consists of five components:

- *data structures* – contain data structures that are used in multiple components to avoid code repetition,
- *execution logic* – performs code simulation on an abstract machine, the output of which is the states and intermediate states of the abstract machine that occurred during execution,
- *GUI execution* – visualization of the simulation with which the user will interact,
- *translator logic* – translates the code from the language of the robot into the language of the abstract machine,
- *translator GUI* – the implementation of the page through which the user will enter input and perform the translation.

The last four components are containers and have their own *Dockerfile* defined, which downloads the openjdk version 11 image, defines a new user and working directory, copies

the JAR archive and runs it in the given container on the specified port. The database has its own container and a *Dockerfile* with a *postgres* version 13.3 image, a user and database set up and running it. For easier work with containers, a *docker-compose* file was also defined, which is common to all components with described image name, source code path, ports, container name, dependencies and environment variables. The interface between the front end and back end parts is obvious.

The components of applications are running on Spring Boot which is being used in a big amount of projects in areas like cloud computing, big data, reactive programming and client applications development [39]. Spring Boot data can be displayed on web pages using a template engine like Thymeleaf which is used for our GUI components. Visualization, HTTP requests and other logic are implemented using JavaScript, jQuery and other libraries.

For reading and transforming the user input, a technology of ANTLR has been used [23]. Three ANTLR grammar files are designed for our application for compiler and simulator logic components:

- simple directional commands (without rotations) for translator – translates input source code to abstract machine language where only **up**, **down**, **left**, **right** are present, then commands with parameter n as number of steps and **reset** commands are valid;
- commands with rotations for translator – translates input source code to abstract machine language where only **forward**, **forward** n, **turn left**, **turn right** and **reset** commands are valid;
- parser for simulator – parses the abstract machine code into data structures (bytecode) on which simulation can be performed.

Finally, we give a brief overview of the technologies that were used in the development of the presented application. As a development environment, the IntelliJ IDEA environment of the Jet-Brains company with a valid university license, the Git version management tool and the university Gitlab platform, as well as the Docker Desktop container management tool, which was available for free for school non-commercial work at the time of software development, were used. Maven version 3.6.3 was used for compilation. The languages Java version 11.0.11, JavaScript, HTML together with the Thymeleaf templating tool and the Spring framework were used for the implementation, for which the IntelliJ IDEA environment provides the generation of basic source and configuration files. For parsing and translation, ANTLR 4 was used as an IntelliJ IDEA plugin, installable from the plugin store in the environment settings, the grammar files associated with them, the generated Java files, and the auxiliary "interp" and "tokens" files. Other files are css style files, icons from the Flaticon site by Freepik, Kiranshastry, Vectors Market and Chanut, and configuration files such as Dockerfile and docker-compose to define containers. The jQuery library was used to manipulate the HTML page and provide a simpler programming interface. The jsPDF JavaScript library developed by MrRio was used to create PDF documents. Creating archives in ZIP format was handled by the JSZip open source library.

5. Functionality of Application

An application can be accessed through a standard web browser (the application is optimized for the Mozilla Firefox browser). This allows users to work with the application

from anywhere. An application allows to insert the user's input source code or to load a code into the application from external storage. The idea is that user input is sent to the back-end layer and compiled. If no error occurs, the source is correct and compiled bytecode is sent back for performing the visualization.

The user interface helps the user to enter the input source by providing also hints for writing the code in the robot language. The hint disappears when the user starts to write the code. Two main functionalities are available – simulation (visual computation) of abstract machine and compilation (translation) of input source to internal bytecode (a data structure according to the specification).

For the implementation of the compilation module, the ANTLR tool was used. The compiler reads an input source and identifies in which variant of language the program is written:

- if code contains rotation commands, appropriate grammar is applied;
- if code does not contain rotation commands, only directional-stepping commands, then grammar for this language, is applied;
- otherwise, no code is generated and the program provides error information (since the languages cannot be mixed).

If the commands are not mixed, but during the compilation, the lexical or syntactic error occurs, the compilation does not provide a bytecode, only information about the compilation error. Then, abstract machine code is annotated by adding HTML tags and sent back to the front end where all errors are highlighted.

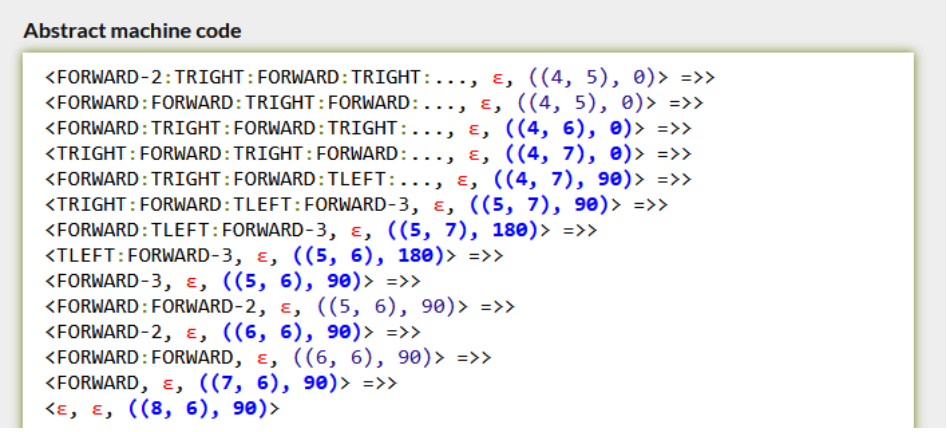


Fig. 2. A computational sequence of an abstract machine code (cropped screenshot)

The successful compilation, if no error appears, provides a bytecode for abstract machine computation steps and visualization. The result of the compilation is then stored in an internal database and sent back to the GUI component.

The main screen of the application contains two visual areas. An abstract machine code is displayed on the right-hand side of an application window (Fig. 2). The application

allows also to save the source code. The user interface supports being displayed in English and Slovak languages, a color theme can be changed from light to dark, as well.

By clicking on the execution button, the user is redirected to the simulator page with the id of the latest translation request. On the simulator web page, the given id is sent to the back end and then to the simulator logic. There the id is searched in the database for a given request. After finding the request, the abstract machine code is parsed according to the third grammar to a list of commands from which the initial configuration is created and execution of these commands is started. The result of an execution is a list of states or rather configurations of the abstract machine and this list is added to the execution request which is sent back to the user and front-end.

On the left-hand side, there is the visualization canvas showing the robot's position and rotation in an orthogonal grid (Fig. 3).



Fig. 3. Visualization canvas

Below (under the canvas area) is the simulation controller with control buttons (play, pause, step back, step forward and reset buttons) and a simulation speed slider. On the right-hand side, there are transitions of configurations from the initial configuration to the current configuration. On the top, in the middle, a currently performed command from the translated abstract machine code is highlighted. The user can download a record of the simulation in four different formats:

- standard CSV format with position, rotation, value and code stacks for each configuration in time,
- XML format with structured data similar to CSV data,

- PDF file with a title, current time, a snapshot of the canvas and transitions of configurations which are generated using jsPDF library by MrRio[1],
- a TEXsource format and picture with a snapshot of the canvas which looks similar to the PDF and is stored in a ZIP archive using JSZip[2].

Visualization is performed using JavaScript following states or configurations from simulator logic to move the robot in its environment and draw its path.

6. Conclusion

In this paper, we presented our approach to define structural operational semantics and its abstract implementation for selected domain-specific language describing the controlling of robot. We developed and defined our semantic approach for both versions of the language for natural semantics in [36]. For the existing approach in structural operational semantics (defined in [13]), we defined our approach of abstract implementation. The abstract implementation consists of a definition of an abstract machine and from translation function that transforms each command of robot language into a sequence of abstract machine instructions. We defined one abstract machine consisting of instructions suitable for both versions of the robot language, together with their semantics expressing changing configurations during instruction execution.

For this semantic approach, we also developed a visualizing software that provides a compilation of input source code written in robot language and emulates the calculation of abstract machine for both versions of the language, and it accepts direct input in abstract machine code and provides also the calculations and emulations.

The design of the micro-service system is created by using containerization, so each micro-service can be reused in other systems later (if we assume the future work is oriented to a complex software visualizing environment).

The created application offers many possible extensions, also thanks to the modular approach applied by containerization, and improvements. The list of commands can be extended in future based on a new specification. Such developed application is ready to be integrated into the teaching process for the courses oriented to formal semantics, and (possibly) formal languages. Its added value is a significant degree of interactivity, clarity and illustrativeness and, above all, the possibility to use the application in the process of present and distance teaching as well as during the independent preparation of students.

Because we see the potential for expanding and consolidating formal methods for software engineering, mainly because all formal methods are based on formal semantics, we will focus our research on semantic methods with the possibility of integrating the results into practice and into teaching young IT specialists and experts.

Acknowledgments. This work was supported by the project 030TUKE-4/2023 – "Application of new principles in the education of IT specialists in the field of formal languages and compilers", granted by Cultural and Education Grant Agency of the Slovak Ministry of Education *and* in the frame of the initiative project "Semantics-Based Rapid Prototyping of Domain-Specific Languages" under the bilateral program "Aktion Österreich-Slowakei, Wissenschafts- und Erziehungskooperation" granted by Slovak Academic Information Agency.

[1] https://mrrio.github.io/jsPDF/

[2] https://github.com/Stuk/jszip

The authors express their gratitude to Prof. Marjan Mernik for the original idea of how to formulate a domain-specific language for a given application domain. The authors also would like to thank Dániel Horpácsi and Judit Horpácsiné Kőszegi for their approach to DSL semantics for the robot that motivated us in our research.

References

1. Alam, M.T., Chowdhury, S., Halder, R., Maiti, A.: Blockchain domain-specific languages: Survey, classification, and comparison. In: 2021 IEEE International Conference on Blockchain (Blockchain). pp. 499–504 (2021)
2. Alam T., e.a.: Blockchain domain-specific languages: survey, classification, and comparison. 2021 IEEE Int.Conf. on Blockchain p. 499–504 (2022)
3. Arusoaie, A.: Certifying findel derivatives for blockchain. Journal of Logical and Algebraic Methods in Programming 121, 100665 (2021), https://www.sciencedirect.com/science/article/pii/S2352220821000286
4. Chamberlain, R., Schommer, J.: Using docker to support reproducible research. DOI: https://doi. org/10.6084/m9. figshare 1101910, 44 (2014)
5. Combe, T., Martin, A., Di Pietro, R.: To docker or not to docker: A security perspective. IEEE Cloud Computing 3(5), 54–62 (2016)
6. Demirkol, S., Challenger, M., Getir, S., Kosar, T., Kardas, G., Mernik, M.: A DSL for the development of software agents working within a semantic web environment. Computer Science and Information Systems 10(4), 1525–1556 (2013), https://dk.um.si/IzpisGradiva.php?lang=eng&id=66708
7. Deursen, A., Klint, P.: Domain-Specific Language Design Requires Feature Descriptions. Journal of Computing and Information Technology (2002)
8. Deursen, A., Klint, P., Visser, J.: Domain-specific languages: An annotated bibliography. SIGPLAN Notices 35, 26–36 (01 2000)
9. Dornelles, E., Parahyba, F., Frantz, R., Roos-Frantz, F., Reina-Quintero, A., Molina-Jiménez, C., Bocanegra, J., Sawicki, S.: Advances in a dsl to specify smart contracts for application integration processes. In: Anais do XXV Congresso Ibero-Americano em Engenharia de Software. pp. 46–60. SBC, Porto Alegre, RS, Brasil (2022), https://sol.sbc.org.br/index.php/cibse/article/view/20962
10. Fowler, M.: Domain-specific languages. Pearson Education (2010)
11. Frantz, C., Nowostawski, M.: From institutions to code: Towards automated generation of smart contracts. pp. 210–215 (09 2016)
12. Hamdaqa, M., Met, L.A.P., Qasse, I.: iContractML 2.0: A domain-specific language for modeling and deploying smart contracts onto multiple blockchain platforms. Information and Software Technology 144, 106762 (2022)
13. Horpácsi, D., Kőszegi, J.: Formal semantics. https://regi.tankonyvtar.hu/en/tartalom/tamop412A/2011-0052_05_formal_semantics/index.html (2014), accessed: Dec 14th, 2020
14. Johanson, A.N., Hasselbring, W.: Hierarchical combination of internal and external domain-specific languages for scientific computing. In: Proceedings of the 2014 European conference on software architecture workshops. pp. 1–8 (2014)
15. Kapferer., S., Zimmermann., O.: Domain-specific language and tools for strategic domain-driven design, context mapping and bounded context modeling. In: Proceedings of the 8th International Conference on Model-Driven Engineering and Software Development - MODELSWARD. pp. 299–306. INSTICC, SciTePress (2020)
16. Kollár, J., Porubän, J., Chodarev, S.: Modelovanie a generovanie softvérových architektúr [Modeling and generation of software architectures]. elfa s.r.o (2012), (*in Slovak*)

17. Lamela Seijas, P., Nemish, A., Smith, D., Thompson, S.: Marlowe: Implementing and analysing financial contracts on blockchain. In: Bernhard, M., Bracciali, A., Camp, L.J., Matsuo, S., Maurushat, A., Rønne, P.B., Sala, M. (eds.) Financial Cryptography and Data Security. pp. 496–511. Springer International Publishing, Cham (2020)
18. Luković, I., Pereira, M.J.V., Oliveira, N., Cruz, D.d., Henriques, P.R.: A DSL for PIM Specifications: Design and Attribute Grammar based Implementation. Computer Science and Information Systems 8(2), 379–403 (2011)
19. Mernik, M., Heering, J., Sloane, A.: When and how to develop domain-specific languages. ACM Comput. Surv. 37, 316– (12 2005)
20. Nielson, Riis, H., Nielson, F.: Semantics with Applications: An Appetizer. Springer Science & Business Media (2007)
21. Oei, R., Coblenz, M.J., Aldrich, J.: Psamathe: A DSL with flows for safe blockchain assets. CoRR abs/2010.04800 (2020), https://arxiv.org/abs/2010.04800
22. Parity Technologies: Ink! (2020), https://github.com/paritytech/ink, [Online; accessed February 2024]
23. Parr, T.: The Definitive ANTLR 4 Reference. Pragmatic Programmers, LLC, The, Raleigh (2013)
24. Patil, S.: Study of container technology with docker. In: International Journal of Advanced Research in Science, Communication and Technology pp. 504–509 (2021)
25. Pereira, M., Mernik, M., Cruz, D., Rangel Henriques, P.: Program comprehension for domain-specific languages. Comput. Sci. Inf. Syst. 5, 1–17 (12 2008)
26. Popovic, A., Luković, I., Dimitrieski, V., Djukic, V.: A dsl for modeling application-specific functionalities of business applications. Computer Languages, Systems & Structures 43 (05 2015)
27. Rad, B.B., Bhatti, H.J., Ahmadi, M.: An introduction to docker and analysis of its performance. International Journal of Computer Science and Network Security (IJCSNS) 17(3), 228 (2017)
28. Radaković, D., Herceg, D.: Towards a completely extensible dynamic geometry software with metadata. Computer Languages, Systems & Structures 52, 1–20 (2018)
29. Roche, N., Hernández, W., Chen, E., Siméon, J., Selman, D.: Ergo - a programming language for smart legal contracts. CoRR abs/2112.07064 (2021), https://arxiv.org/abs/2112.07064
30. Schreiner, W.: Theorem and algorithm checking for courses on logic and formal methods. In: Quaresma, P., Neuper, W. (eds.) Proceedings 7th International Workshop on Theorem proving components for Educational software, ThEdu@FLoC 2018, Oxford, United Kingdom, 18 july 2018. EPTCS, vol. 290, pp. 56–75 (2018), https://doi.org/10.4204/EPTCS.290.5
31. Schreiner, W., Steingartner, W.: The SLANG Semantics-Based Language-Generator — Tutorial and Reference Manual (Version 1.0.*). Technical report 23-13, Research Institute for Symbolic Computation (RISC), Johannes Kepler University Linz, Austria (September 2023), doi:10.35011/risc.23-13
32. Seidametova, Z.: Some methods for improving data structure teaching efficiency. Educational Dimension 58, 164–175 (Jun 2022), https://journal.kdpu.edu.ua/ped/article/view/4509
33. Steingartner, W.: Compiler module of abstract machine code for formal semantics course. In: 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI). pp. 000193–000200 (2021)
34. Steingartner, W.: On some innovations in teaching the formal semantics using software tools. Open Computer Science 11(1), 2–11 (2021), https://doi.org/10.1515/comp-2020-0130
35. Steingartner, W., Gajdoš, E.: The visualization of a graph semantics of imperative language. Polytechnica: Journal of Technology Education 2(5), 7–14 (2021), https://doi.org/10.36978/cte.5.2.1

36. Steingartner, W., Novitzká, V.: Natural semantics for domain-specific language. In: New Trends in Database and Information Systems. pp. 181–192. Springer International Publishing, Cham (2021)
37. Steingartner, W., Novitzká, V., Schreiner, W.: Proof of Equivalence of Semantic Methods for a Selected Domain-Specific Language. Journal of Applied Mathematics and Computational Mechanics 23(2) (2024)
38. Steingartner, W., Zsiga, R., Radaković, D.: Natural semantics visualization for domain-specific language. In: 2022 IEEE 16th International Scientific Conference on Informatics (Informatics). pp. 293–298 (2022)
39. Walls, C.: Spring Boot in action. Manning Publications (2016)
40. Wohrer, M., Zdun, U.: From domain-specific language to code: Smart contracts and the application of design patterns. IEEE Software 37(5), 37–42 (2020)
41. Wood, G.: Ethereum: A secure decentralized generalized transaction ledger. Ethereum project yellow paper 151(2014), 1–32 (2014)
42. Wood, G.: Polkadot: Vision for a heterogeneous multi-chain framework (2016), 21(2327): 4662

William Steingartner works as Associate Professor of Computer Science at the Department of Computers and Informatics of the Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia. He defended his PhD thesis "The Rôle of Toposes in Computer Science" in 2008. His main fields of research are semantics of programming languages, category theory, compilers, data structures and recursion theory. He also works with cybersecurity and software engineering.

Valerie Novitzká works as a Full Professor of Informatics at the Department of Computers and Informatics of the Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia. Her fields of research include semantics of programming languages, non-classical logical systems and their applications in computing science. She also works with type theory and behavioural modeling of large program systems based on categories.

Received: December 01, 2023; Accepted: April 28, 2024.

Computer Science and Information Systems 21(3):1097–1145 https://doi.org/10.2298/CSIS230217031C

Natural Language Processing: An Overview of Models, Transformers and Applied Practices

Santiago Canchila[1], Carlos Meneses-Eraso[1], Javier Casanoves-Boix[2], Pascual Cortés-Pellicer[3], and Fernando Castelló-Sirvent[3]

[1] Universidad Sergio Arboleda
Calle 74 No. 14 - 14 Bogotá, Colombia
santiago.canchila01@usa.edu.co
carlos.meneses@usa.edu.co
[2] Universidad Alfonso X El Sabio
Avda. de la Universidad, 1. Villanueva de la Cañada, 28691, Madrid, Spain
jcasaboi@uax.es
[3] Department of Business Management Universitat Politècnica de València (UPV)
Camí de Vera, s/n - 46022 València, Spain
pascorpe@omp.upv.es
fernando.castello@upv.es

Abstract. The study of utilizing human language in computer systems referred to as NLP, is becoming increasingly significant in various aspects of life, including research, daily activities, commerce, and entrepreneurship endeavors. A multitude of tech companies are dedicating resources towards the development and improvement of NLP methods, models, and products. To add to that, open-source contributions to the field are on the rise. However, with so much progress being made, it may be challenging to understand the current state of NLP and what models are considered to be the most efficient. To help those grappling with the fast-paced and constantly evolving NLP landscape, we have put together a comprehensive overview of the latest NLP research and advancements.

Keywords: NLP, Deep Learning, Transformers.

1. Introduction

The area of Natural Language Processing (NLP) appears to have a great impact on today's research, day-to-day life, commercial, and entrepreneurial endeavors. Some of the biggest tech companies have a footing in the development of newer NLP techniques, models, architectures, and products. Some open-sourcing efforts have also started to materialize. With all those new creations it may be hard to understand what the state-of-the-art is and what are the optimal models for NLP. We conduct a survey of NLP research and novelties to help those groups that may be entangled in the accelerating and messy area of current NLP.

Many applications in the field of NLP have been studied, including question answering (QA), text classification, summarization, text generation, zero-shot or few-shot classification, translation, text similarity, sentiment analysis, named entity recognition (NER), relation extraction, information retrieval, natural language inference, dialogue, factual information detection, text simplification, text clustering among others.

The area of NLP has had a massive growth fig 5 over the past years. Different approaches, techniques, and models have been proposed by researchers, trying to improve upon an, or various, aspects of previous proposals. These aspects can be summarized with different metrics, some of the most relevant ones being task-specific performance, energy efficiency, memory used, and training complexity.

Different models have been proposed to solve NLP tasks, and trade-offs have to be made when choosing the best model to use. Previous approaches such as Bag of Words or TF-IDF still prove to be useful in the NLP landscape, and should be a viable consideration when solving NLP tasks, especially when computational resources and data may be scarce. With this into consideration, it is important to note that Transformer models have achieved state-of-the-art performance on different tasks. According to Ref [46] BERT, pre-trained on a financial news corpus, allows to detect the dominant themes in the news, as well as to evaluate the tone of the articles. For their part, Ref [12] argues that, in contrast to ordinary text classification, short text presents a unique problem, consisting of less vocabulary and scarcity of features, which generates a higher need for semantic feature representation. Therefore, in order to address this problem they propose a model based on BERT.

A way to identify Spanish language hate discourse on social media using pre-trained monolingual and multilingual models, including BERT, XLM [81] and BETO [4] was proposed [8].

Ref [91] studies the possibility that removing news tagged as harmful may also eliminate satiric news. Using DistilBERT [116] they show a way to distinguish between satiric and fake news. Social media generates a large amount of data that could be used to identify the sentiment of its users using XLNet [160], DistilBERT [116] or RoBERTa [90].

Client relationships can also be improved using NLP, detecting possible weak points in client-company interactions. RoBERTa is used to detect five weak points to improve this interaction [115].

Taking those applications into consideration, it is still difficult to contemplate the whole NLP landscape. With a growing research scene, more models with various benefits, and different levels of access, it can become a difficulty to those groups wanting to use what the NLP area has to offer. With this in mind, an overview of different models and techniques will be shown, emphasizing the state-of-the-art transformer architecture [143]. A small analysis of the impact of transformers, especially Bidirectional Encoder Representations from Transformers (BERT) [43], on the research community will be presented. An analysis of the abstracts of many NLP related papers was made using various approaches to better understand their trade-offs. Some of the most relevant models, architectures, and approaches can be a point of discussion, showing that there is a trend for bigger, decoder-only, generative models, and what that implies. Finally, some guidance on what the future of NLP may look like will be provided.

2. Historic Overview

The question of whether or not computers can process natural language has been an inherent topic in computer science. In the 1950s, Alan Turing proposed a way to identify if

[4] https://github.com/dccuchile/beto

machines can think, with the imitation game [140]. This test is essentially an NLP problem, as the computer would need to process the question given by the person, understand it, and answer. This paves the way for NLP research, as it would not only advance an important field but would, if successful, also solve one of the open questions in computer science.

The earliest wave of computational NLP research came in the 1950s, introducing the concept and starting to involve the rudimentary computational power of the time. Because of the limitations of computers at that time, both in memory and processing power, it was not feasible to test or implement the proposed techniques, but a theoretical origin of NLP was starting to develop. Some groundwork was created around natural language, establishing common problems in a computational representation of language. Natural language is complicated and has rules and exceptions that are the result of thousands of years of change, so it was obvious at the beginning that, for example, hard coding a chatbot with answers was not feasible, because of the sheer amount of possibilities. Some interpretation and variability were needed in order to obtain good results. Rule-based systems started to be implemented to solve those limitations. An example from this rule-based approach to NLP was one of the first applications of NLP in a practical field, being of help in a medical setting, used by physicians in training was proposed in Ref [124]. Some areas of research started to develop, including text-to-speech [4], syntactic analysis [31], part of speech tagging [18], machine translation [5] and human interaction [152].

NLP was described as a range of computational methods driven by theories to analyze and represent natural language text at various linguistic levels, aiming to imitate human language processing for various tasks and applications [23]. NLP is a branch of machine learning that deals with text and speech [2], that began in the 1940s as an intersection of artificial intelligence and linguistics [96]. Since its inception, contributions in the theory of universal grammar, have been definitive for the development of NLP to adequately address the study of language [30,10]. Other investigations, such as the contributions of Ref [58], in which they propose the use of an associative categorical dual model for the representation of the meaning of words and its possible applications in artificial intelligence and NLP.

Ref [86], shows that NLP had two distinct focuses: Natural Language Understanding, NLU (reader/listener) for the purpose of producing a meaningful representation, and Natural Language Generation (writer/speaker) that refers to the production of language from a representation, that, according to Ref [76], the first has five sections: Phonology, Morphology, Syntax, Semantics and Pragmatics, and the second refers to identifying the objectives, planning how the objectives can be achieved by evaluating the situation and the available communication sources and making the plans as a text. research that analyzes the concepts of NLP such as Information Retrieval (IR) and Information Extraction (IE), which is the process of finding in a large data repository, the material of an unstructured or semi-structured nature, while Information Extraction (IE), consisting of obtaining sections of interest in the text to transfer them to a structured format [56]

It is important to note that previous approaches to NLP are still used, especially in the preprocessing stage, and they also give good performance on certain tasks and certain situations.

[5] https://www.ibm.com/ibm/history/exhibits/701/701_translator.html

3. Previous Approaches

Traditional approaches have various advantages over the current state-of-the-art approaches, as they are less computationally expensive, they may be a good option especially when data is scarce, and they can still obtain competitive performances. Some of those approaches will be presented next.

3.1. Rule-based and symbolic

This approach may include regex, rule-based systems, morphological analysis, tokenization, syntax analysis, parsing, and text transformation and normalization. NLP systems that are interpretable can promote trust by allowing end-users, practitioners, and researchers to grasp the model's prediction mechanisms, thereby ensuring ethical NLP practices. In the past, conventional NLP systems, like rule-based approaches, have typically operated in this manner [69,155]. Some examples of rule-based collections, topics, systems, and their possible uses are:

- One of the first books on the subject is presented, documenting many of the existing expert systems, as well as some of the available software for building them. From this, they abstract some principles and general approaches for the development of expert systems. And, some time later, the principles and history of the method are explained, exposing that rule-based systems automate knowledge for problem-solving, provide a means to capture and refine human experience [118].
- The ISCREEN system is described, a prototype for text message detection. ISCREEN includes a high-level interface for users to define rules, a text message filtering component, and a conflict detection component that examines the rules for inconsistencies. An explanation component uses text generation to respond to user queries about the system's past or potential actions, based on Grice's conversational maxims [108,55].
- Research on natural language processing involves at least four different but closely related areas of study, which are: (1) research on the psychological processes involved in understanding human language, (2) the construction of computational systems for analyzing natural language input (and/or generating natural language output), (3) the development of theories about the structure of natural language, and (4) the determination of the mathematical properties of grammatical formalisms [119].
- Some recent approaches to fully rule-based systems can still be useful. A cloud-based rule-based computing system for evaluating requests for academic scholarships for postgraduate students is presented. Among the findings, it is demonstrated that the system allowed students to submit their documents online and view the score for each item entered. The final cumulative points were automatically calculated according to the standards. The system reduced the burden on students during their applications. As a result, the system could motivate students to engage in academic scholarship applications by providing transparent and user-friendly public rules [165].

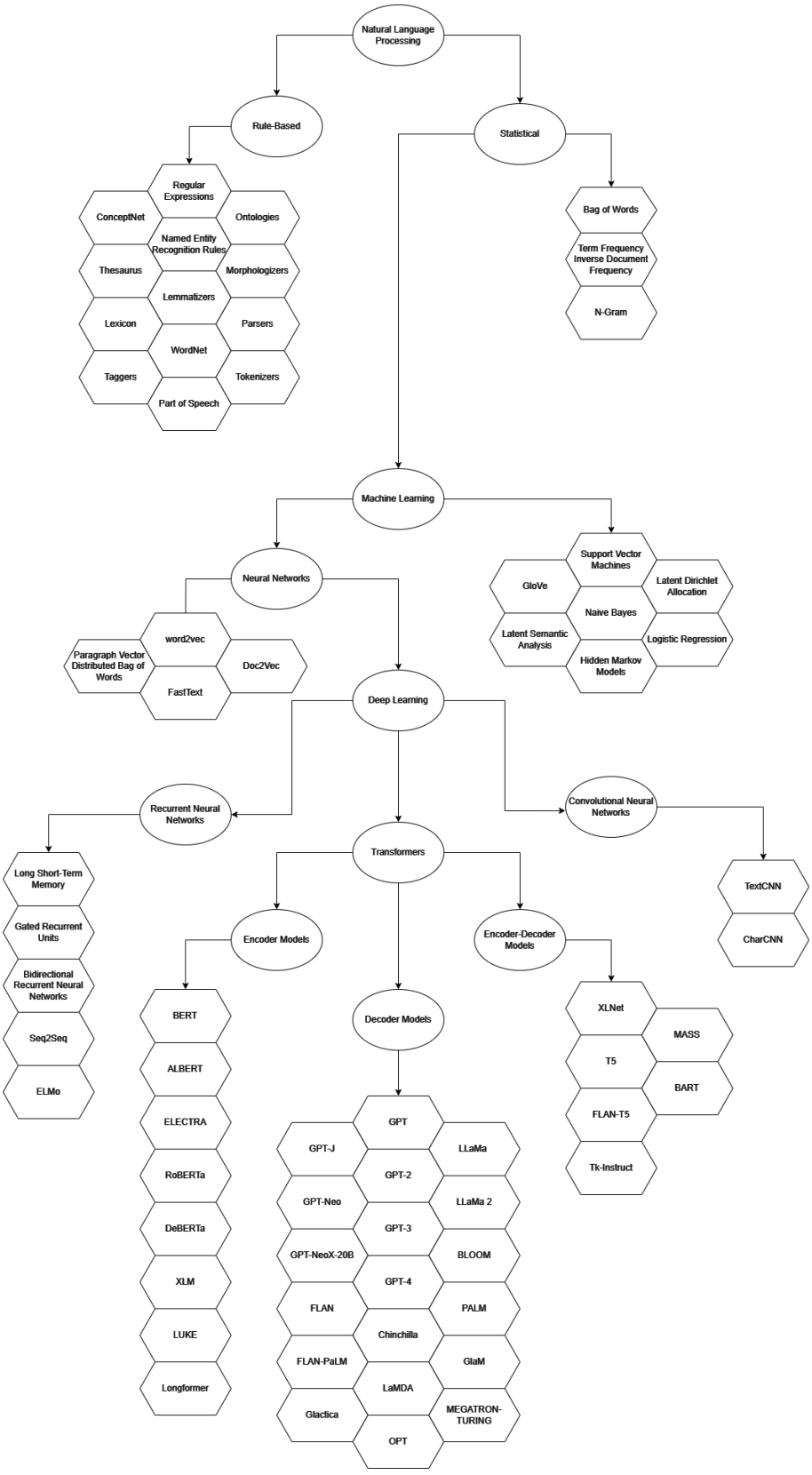


Fig. 1. Map of NLP techniques and models

3.2. Bag of Words

The Bag of Words (BoW) methodology is a simple yet effective way to represent textual data using word frequency vectors. It was initially proposed for text document analysis [59], but similar approaches have found a place in other areas such as image analysis. In this method, the number of occurrences of each "bag" generated for each type of instance or word is recorded, without considering their order or grammar. Different approaches to represent BoWs are used, which can include hashmaps or sparse matrices [6]. BoWs have been used extensively in NLP, and it has shown to be effective when paired with other algorithms, especially machine learning and statistical ones.

3.3. TF-IDF

TF-IDF works by determining the relative frequency of words in a specific document compared to the inverse proportion of that word in the entire document corpus. Intuitively, this calculation determines how relevant a particular word is in a specific document. Words that are common in a single document or a small group of documents tend to have higher TF-IDF values than common words like articles and prepositions. The formal procedure for implementing TF-IDF has some minor differences in its various applications, but the general approach works as follows. Given a collection of documents D, a term t, an individual document $d \in D$, a function $f(t, d)$ that returns the frequency of term t in document d, and T_d different terms in d, we calculate:

$$TF(t, d) = \frac{f(t, d)}{\sum_{i=0}^{T_d} f(t_i, d)} \tag{1}$$

$$IDF(t, D) = log(\frac{|D|}{|\{d \in D | t \in d, \forall d \in D\}|}) \tag{2}$$

$$TF\text{-}IDF(t, d, D) = TF(t, d) \cdot IDF(t, D) \tag{3}$$

TF-IDF also works best when paired with other algorithms, as the vector, mathematical representation of documents allows the use of different models and approaches.

3.4. N-gram

N-grams are sequences of characters or words extracted from a text. They can be categorized into two main types, character-based and, word-based. A character-based N-gram is a set of n consecutive characters extracted from a word. The primary motivation behind this approach is that similar words will have a high proportion of common N-grams. Typical values for n are 2 or 3, corresponding to the use of bigrams or trigrams, respectively.

Let S be a sequence of text data, and S_i represent the i-th item in the sequence, where $i \in [0..|S|]$. An n-gram can be defined as a contiguous sequence of n items from S, denoted as S_i, S_{i+1}, ..., S_{i+n}, where $0 \leq i \leq (|S| - n)$ [122]. N-grams are useful when capturing local patterns in text.

[6] https://docs.scipy.org/doc/scipy/reference/sparse.html

3.5. LSA

LSA is a technique in natural language processing and information retrieval that uses singular value decomposition (SVD) to analyze and represent the relationships between terms and documents in a high-dimensional space. It's widely used for tasks such as document retrieval, information retrieval, and text analysis to discover the latent semantic structure in a corpus of text [40]. LSA has found success in topic modeling, similarity, clustering, and classification among others.

A Document-Term Matrix must be constructed, representing the frequency of a term t in a document d for every $t \in T$ where T are all the terms in a corpus D, and every $d \in D$. A TF-IDF, or any other weighted, matrix can also be used. A decomposition into three matrices U, Σ and V^T is done, where

- U: The left singular vectors matrix represents the relationships between documents.
- Σ : A diagonal matrix with singular values, representing the strength of the latent semantic concepts.
- V^T: The right singular vectors matrix represents the relationships between terms.

Dimensionality reduction can then be applied by keeping the top k singular values and their vectors. After this reduction, a semantic space is achieved, where all documents and terms are described by a vector in the space, capturing a latent semantic relationship between them.

3.6. LDA

Latent Dirichlet Allocation represents a three-tiered hierarchical Bayesian model, where every item in a collection is depicted as a finite mixture of underlying topics. These topics are further represented as an infinite mixture of topic probabilities, which, in the context of text modeling, offer a clear way to represent a document [16].

In LDA, a document, denoted as d, is a collection of $|d|$ words, where $d = w_0, w_1, ..., w_{|d|}$. For all documents d in the corpus D and all topics k in the set of topics K, LDA models each document as a distribution over topics, creating a matrix of size $(|D|, |K|)$ to represent the different latent topic proportions for each document.

Additionally, a distribution β is created for every topic k, which represents the probability distribution over words for that topic. The goal of LDA is to learn the optimal θ distributions for every document d and the optimal β distributions for every topic k.

The matrix β is constructed, where $\beta_{ij} = p(w^j = 1|z^i = 1)$ represents the probability of word w^j occurring in topic z^i. The ultimate objective of LDA is to discover these latent topic proportions and word distributions to represent the underlying structure of the document collection.

LDA has proven to be a performant model for topic modeling in a variety of applications and domains. Its ability to uncover latent topics within a corpus of text data has made it a valuable tool for tasks such as document clustering, content recommendation, and sentiment analysis.

3.7. word2vec

Although vectorial and spatial representation of texts were proposed previously in other models and methodologies, word2vec [34] is a model that achieved major improvements in NLP with the use of embeddings.

The generation of a word embedding through Word2Vec can be done in two different ways: Skip-Gram and Continuous Bag of Words. Both approaches are based on a different handling of input and output variables but essentially use the same neural network structure. They have similar architectures but different learning objectives. The skip-gram model uses the current word to predict the surrounding words, while the CBOW model predicts the current word based on the surrounding words [44]. Skip-gram training tends to give a better semantic representation of words. The idea of a skip-gram is to be able to, given a probability distribution, predict the surrounding words in a window for every word W_n. This was one of the objectives proposed with the word2vec model[34].

$$P(W_{n+j}|W_n) \tag{4}$$

where the word W_t is the center word. Then an objective function to be minimized for N words can be formulated for a window of size m surrounding a word W_n:

$$J'(\theta) = \prod_{n=0}^{N} \prod_{\substack{-m \leq j \leq m \\ j \neq 0}} P(W_{n+j}|W_n;\theta) \tag{5}$$

Where the probability is also dependent on some parameters θ

A loss function to be optimized can then be proposed as follows. This loss function is generally optimized using gradient descent:

$$J(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} logP(W_{n+j}|w_n) \tag{6}$$

3.8. GloVe

GloVe is an unsupervised learning algorithm used to obtain vector representations of words. It trains on global statistics derived from word-word co-occurrence in a corpus, and the resulting representations reveal interesting linear substructures within the word vector space.

The training objective of GloVe is to learn word vectors in such a way that their dot product equals the logarithm of the co-occurrence probability of words. This objective associates (the logarithm of) probability ratios of co-occurrence with vector differences in the word vector space, thanks to the mathematical equivalence of logarithmic ratios and vector differences. This allows the encoding of some form of meaning, as these ratios can capture semantic relationships [106] [7].

[7] https://nlp.stanford.edu/projects/glove/

3.9. Neural Networks and Deep Learning

Given the advancement of computational power, the use of more demanding models became possible. In this order, neural language modeling aims to improve the performance of multiple related learning tasks [37] by taking advantage of useful information among them [28]. Task relationship is based on understanding how different tasks are related and includes learning tasks such as supervised tasks, which include classification and regression tasks, unsupervised tasks, including clustering tasks, semi-supervised tasks, active learning tasks, reinforcement learning tasks, online learning tasks, and multi-view learning tasks [166]. In the same way, the concept of word embedding is expanded on further, improving search tasks, and content recommendation [34]. Sequence-by-sequence learning models with neural networks were included in the development of NLP in [131] through a translation task using a multilayer LSTM to assign the input sequence to a vector of fixed dimensionality, and then another deep LSTM to decode the target sequence of the vector.

On the other hand, Ref [76], summarizes the use of convolutional neural networks in the fields of Sentence Classification, Sentiment Analysis, Text Classification, Text Summary, Machine Translation, and Response Relations, as well as the use of recurrent neural networks (RNN) which, according to Ref [133] is defined as a type of artificial neural network where the connections between units form a directed cycle creating an internal state of the network that allows you to exhibit dynamic behavior, usually applied to sequential data such as text, time series, financial data, voice, audio, video among, others [76]. CNNs have been used extensively for NLP, as Refs [75,51,77,128,147,117] have all used CNNs successfully to solve NLP tasks.

With the growth in computational power, RNNs started to become viable. Long-Short Term Memory (LSTM) [62] and Bidirectional LSTM [54], were used to solve NLP tasks [68,168]. The sequential-like nature of natural language made it perfect for LSTMs, as they specialize in that. LSTMS, however, also had some drawbacks [167], but they still performed exceptionally.

Currently, RNNs have almost been super-seeded by transformers, as they do not offer a better performance and are more computationally expensive to use and train, but they may still offer unique benefits compared to other approaches.

ELMo: Embeddings from Language Models (ELMo) is a model that achieved great performance on various tasks. It represents one of the early large, pre-trained language models that demonstrated a significant improvement in the quality of contextualized word representations. ELMo is a model that relies on LSTM and biLSTM units to create these contextual embeddings. It achieved great performance on a broad range of NLP tasks and benchmarks, including Q
A, Textual entailment, NER, Semantic role labeling, and Sentiment analysis among others [107].

4. Transformers

In 2017 a breakthrough was made in the area of NLP, the transformer architecture was proposed [143], creating a path for some of the current state-of-the-art models to be realized. The original architecture made use of attention mechanisms exclusively, being able

to move away from recurrent and convolutional neural networks. Having such a simple, but scalable, architecture would make transformers the preferred architecture for solving NLP tasks and create state-of-the-art models. Transformers provided improvements in two important aspects without a trade-off: they were more computationally inexpensive, requiring less time to train, and being more parallelizable, while also obtaining better results on standard NLP tasks and NLP benchmarks.

Without the need to use an RNN, the inherent need for sequential computational steps was not needed, so a full token sequence could be computed without the need to precompute previous steps. This allows parallelization in sentence processing, making training more efficient on devices such as GPUs and TPUs. This would result in two of the most attractive aspects of transformers: they are highly scalable and can be easily trained on specialized hardware; they can be pre-trained on large unlabeled text corpus, in a self-supervised manner, to generate language understanding, and then fine-tuned to solve specific NLP tasks while achieving state-of-the-art performance [143]. Generally, the amount of data necessary to perform the fine-tuning on a transformer-based model that has been pre-trained is also considerably lower than if a model was trained from scratch.

The benefits of the transformer architecture made it so researchers started to use it to create more complex, bigger, or specific, derived architectures, while also utilizing different training or fine-tuning data. Currently, transformers represent the state-of-the-art architecture for NLP.

4.1. The Transformer Architecture

The original transformer architecture has different mechanisms that make it so effective. First, a preprocessing pipeline is used, in which a trainable tokenizer is used.

The tokenizer outputs are passed along a learned embedding layer to create a word embedding of the input tokens. This embedding layer generally works similarly to a lookup table, in which the input is an index, related to the tokens, and the output is the embedding for that token, which may differ in dimensionality and may include contextual information. The initialization matrix of this embedding layer can come from other trained word vectors like GloVe or word2vec, or with a random distribution. This embedding layer will then be trained during backpropagation.

A positional encoding is used to maintain information about the position of the tokens in a sequence. In the original transformer architecture a positional encoding layer with the same shape as the embedding layer was proposed so that they could be added. The following functions were used, for a position n, a constant k, and an input dimensionality d_model. The value of k was 10000 in the original architecture [143].

$$PE(n, 2i) = sin(\frac{n}{k^{2i/d_{model}}}) \tag{7}$$

$$PE(n, 2i + 1) = cos(\frac{n}{k^{2i/d_{model}}}) \tag{8}$$

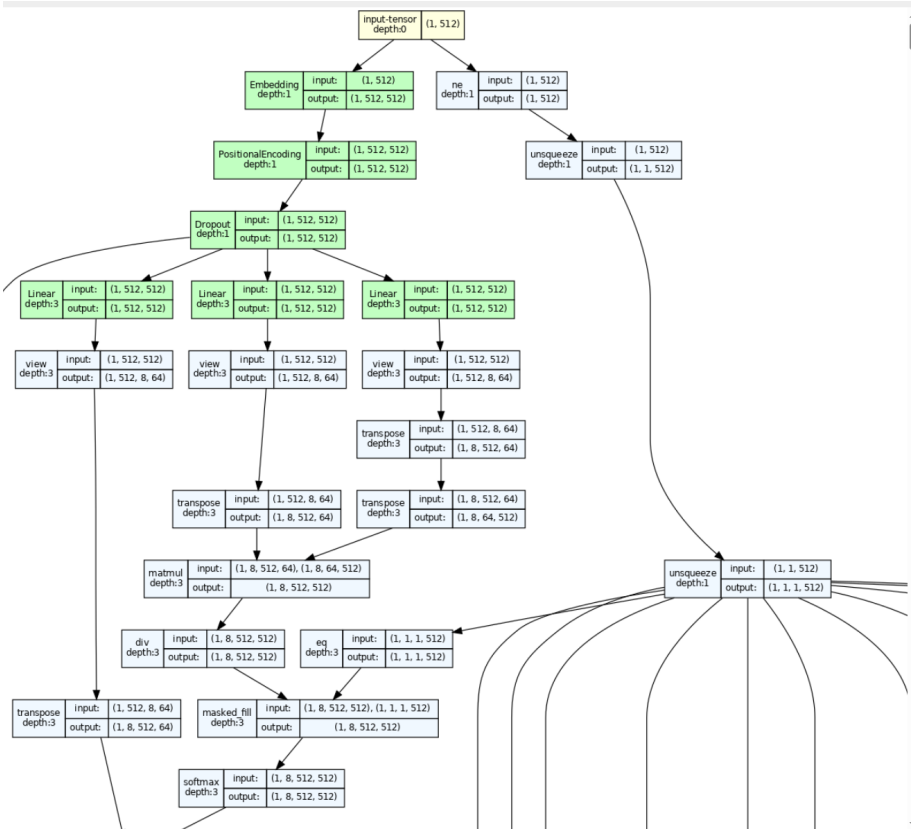


Fig. 2. Computation graph of the transformer architecture

Other models may use a trained positional encoding layer. The positional encoding is achieved with varying degrees of success when using a trained layer, as some models may not actually be learning an absolute or relative positional encoding of the sequence. This may depend on the training objective used when training the model, with better positional representation achieved on models with an autoregressive nature, in comparison to models that use masked modeling [148].

Afterwards, one of the most important points in the transformer architecture is used, the attention heads. Attention was a mechanism proposed some years prior to the transformer architecture [11], and it was used in conjunction with RNNs. The use of an attention head in a transformer expenses it from needing to use a kind of recurrent or convolutional architecture. An attention head takes 3 inputs, a query Q, a key K, and a value V, where [143]

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{9}$$

A version of attention that can be run in parallel, called Multi-Head Attention, was also proposed. The function takes three arguments: Q, K, and V, which are matrices that

correspond to a sequence of query vectors, key vectors, and value vectors, respectively. W_i^Q, W_i^K, and W_i^V are weight matrices for the i-th attention head. They project the input Q, K, and V matrices into a different subspace suitable for the i-th head's attention operation. By having separate weight matrices, each head can potentially learn to focus on different parts of the input or capture different types of dependencies in the data.

$$MultiHead(Q, K, V) = [A_0; ...; A_h]W^O \tag{10}$$

$$where A_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

The intention of using attention heads is that the attention mechanism allows transformers to capture the relations between the tokens in a sequence by providing a spatial similarity of the value related to the query, while also assigning different levels of importance to different parts of the input sequence dynamically. Multi-Head Attention is able to attend to information that may be encoded in different spaces, and may attend to, for example, relationships with different length dependencies[143].

While a transformer model that uses both the encoder block and the decoder block uses attention mechanisms, they can be further categorized into self-attention and attention.

Encoder Block : The encoder block employs self-attention. Self-attention allows each token in the input sequence to attend to all other tokens within the same input sequence. This mechanism captures relationships and dependencies among tokens in the input. In this case, $Q = V = K$, where Q, V, and K are the query, value, and key vectors. The encoder block will then be composed of the following components:

- Self-Attention Layer: As mentioned, the self-attention mechanism allows each token in the input sequence to attend to all other tokens. Since $Q = K = V$, this is specifically known as self-attention because each token is effectively attending to itself and other tokens within the same sequence. This helps the model understand the context and inter-token relationships.
- Feed-Forward Neural Network: After self-attention, the output goes through a Feed-Forward Neural Network (FFNN) that is applied position-wise to each token's representation. The FFNN typically consists of two linear transformations with an activation function in between.
- Residual Connections: Each of the two layers (self-attention and FFNN) has a residual connection that adds the input of the layer to its output, followed by layer normalization. This skip-connection helps with gradient flow and allows the network to learn more effectively by providing an alternative pathway for gradient propagation.
- Layer Normalization: Layer normalization is applied after the residual connection, which normalizes the outputs across the features for each data instance, helping in stabilizing the learning process.

Decoder Block : The decoder block also uses self-attention, but it serves a slightly different purpose. Here, self-attention enables each token in the target sequence to attend to other tokens within the same target sequence. This is crucial for modeling dependencies within the output sequence.

In addition to self-attention, the decoder block utilizes attention between the encoder and decored blocks. This type of attention allows the decoder to focus on information from the encoder's output. This attention enables the decoder to consider relevant information from the input sequence while generating the output sequence. The decoder block will then have the following components:

- Masked Self-Attention Layer: The early self-attention layer in the decoder is modified to prevent each token from attending to subsequent tokens in the sequence. This is known as masked self-attention and is necessary because, during training, the decoder should only be allowed to consider tokens it has already generated.
- Encoder-Decoder Attention Layer: This layer is responsible for attending to the encoder's output. Here, the Q vectors come from the previous masked self-attention layer's outputs, whereas the K and V vectors are the outputs from the encoder stack. This mechanism helps the decoder focus on different parts of the encoder's output which is important for tasks such as translation where the alignment between input and output sequence elements is not known in advance.
- FFNN: Like the encoder block, the decoder also contains a position-wise FFNN.
- Residual Connections and Layer Normalization: Each sub-layer within the decoder (the masked self-attention, encoder-decoder attention layer, and FFNN) includes a residual connection followed by layer normalization.

Attention mechanisms are a fundamental component of transformers, and they encompass both self-attention, for capturing relationships within the same sequence, and attention, for connecting information between different sequences, such as the encoder's output and the decoder's input.

It is important to note that all of these layers are trained with backpropagation. With current frameworks, such as Pytorch [104,103] or JAX[17], automatic differentiation can be used to obtain the needed gradients for training. The autograd those frameworks use, along with graph execution storage, has also been a key point for the efficient training of different DL architectures.

The first steps of the transformer architecture computation may look like fig 2, using PyTorch and torchview [8], while the transformer architecture may look as follows:

```
Transformer(
  (encoder_embedding): Embedding(10000, 512)
  (decoder_embedding): Embedding(10000, 512)
  (positional_encoding): PositionalEncoding()
  (encoder_layers): ModuleList(
    (0-5): 6 x EncoderLayer(
      (self_attn): MultiHeadAttention(
        (W_q): Linear(in_features=512, out_features=512, bias=True)
        (W_k): Linear(in_features=512, out_features=512, bias=True)
        (W_v): Linear(in_features=512, out_features=512, bias=True)
        (W_o): Linear(in_features=512, out_features=512, bias=True)
      )
      (feed_forward): PositionWiseFeedForward(
        (fc1): Linear(in_features=512, out_features=2048, bias=True)
        (fc2): Linear(in_features=2048, out_features=512, bias=True)
        (relu): ReLU()
      )
      (norm1): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
```

[8] https://github.com/mert-kurttutan/torchview

```
      (norm2): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False)
    )
  )
  (decoder_layers): ModuleList(
    (0-5): 6 x DecoderLayer(
      (self_attn): MultiHeadAttention(
        (W_q): Linear(in_features=512, out_features=512, bias=True)
        (W_k): Linear(in_features=512, out_features=512, bias=True)
        (W_v): Linear(in_features=512, out_features=512, bias=True)
        (W_o): Linear(in_features=512, out_features=512, bias=True)
      )
      (cross_attn): MultiHeadAttention(
        (W_q): Linear(in_features=512, out_features=512, bias=True)
        (W_k): Linear(in_features=512, out_features=512, bias=True)
        (W_v): Linear(in_features=512, out_features=512, bias=True)
        (W_o): Linear(in_features=512, out_features=512, bias=True)
      )
      (feed_forward): PositionWiseFeedForward(
        (fc1): Linear(in_features=512, out_features=2048, bias=True)
        (fc2): Linear(in_features=2048, out_features=512, bias=True)
        (relu): ReLU()
      )
      (norm1): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
      (norm2): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
      (norm3): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False)
    )
  )
  (fc): Linear(in_features=512, out_features=5000, bias=True)
  (dropout): Dropout(p=0.1, inplace=False)
)
```

4.2. Usage Growth and Abstract Analysis

The main problems of the Information Explosion registered in the bibliometric context during the last decades is that Scholarly Visibility and efficient searches are largely determined by the keyword selection criteria used by the authors. The success, precision, and adequacy of the choice of these descriptors with respect to the content of published scientific articles represent a matter of great importance for researchers [50].

In order to extend the analytical implications beyond keyword processing, previous studies have addressed a detailed analysis of the title or abstract [41], however, traditional methods applied this analysis manually, which makes it difficult to process large numbers of articles and, consequently, to understand the internal structure of the global academic debate on topics of broad or transversal scope.

Abstract analysis is similar to bibliometric keyword analysis. Unlike the keywords, the abstract offers more information about the evaluated research [114]. By their nature, keywords are limited, both for authors (author keywords) and for journals (keyword plus®). As a consequence, bibliometric analyzes based solely on keywords can incorporate biases that can cause Type II Errors.

At times, abstract analysis has been an unassisted prior conceptualization mechanism [102,121] used by researchers to facilitate their decision-making to help guide the subsequent bibliometric analysis carried out with specific text mining tools [36,142]. Abstract analysis has been used as a content evaluation mechanism in areas such as business and

management [121] or project management [93], among others, and has been applied as an inductive technique [121], registering important limitations on the possibilities of automation and the implications of the findings obtained.

This study uses custom software that allows the automated analysis of the abstracts of the evaluated articles (N=75527), offering an important evolution that improves the process and the actual analysis of the content, overcoming the reported limitations and allowing a robust and rigorous analysis of the academic literature. The dataset was obtained from Web of Science, indexed sources.

A Bag of Words (BOW) was created on all the abstracts from the 75527 articles analyzed. A dataset with the most common English words [9], from the Google Web Trillion Word Corpus, was used to eliminate the first 500 words from the BOW, as they provided little value to the analysis since they were not directly related to NLP. Tab 1 was generated with some words directly related to particular models, architectures, or approaches.

This dataset will be made publicly available, with a caveat on the distribution of the abstracts. A SHA256 hash of the abstract will be provided along with other data, such as authors, DOI, title, and publication date. The raw abstracts will not be shared due to concerns of licensing and copyright holding, as not all of the papers have open access [25]. The BERT and SetFit models are also available [24].

Data from Tab 1 comes from all the NLP related articles that had an abstract available, spanning 6 decades of research. If we select only the papers published in the last 5 years, a different picture is obtained. As expected the usage of previous techniques falls down as time progresses, but even then, some of the usages of more recent techniques, like LSTM, have not changed as fast as transformers based models. BERT, transformers, pre-trained, encoder, embedding, bidirectional, and deep are some of the terms that saw a significant change in usage.

[9] Dataset taken from https://www.kaggle.com/datasets/rtatman/english-word-frequency

[10] Since BERT was introduced in 2018 it had 0 mentions

Table 1. Models

Word	Usage	Word	Usage	Word	Usage
All time usage					
neural	16324	random	3052	transformer	2047
deep	12581	pretrained	3034	rulebased	1951
networks	9568	recurrent	3011	fuzzy	1904
attention	8274	latent	3004	linear	1901
vector	6563	cnn	2967	regression	1894
statistical	5569	lstm	2846	encoder	1808
rules	5546	ner	2746	wordnet	1744
embeddings	5404	means	2554	endtoend	1693
graph	5365	inference	2497	naive	1601
graphs	5365	bidirectional	2417	word2vec	1480
embedding	5031	multimodal	2280	clusters	1476
clustering	3971	svm	2224	bayes	1468
convolutional	3518	pos	2201	markov	1436
bert	3195	sequences	2107		
Usage before 2018					
neural	5522	pos	1397	embedding	903
graphs	4397	wordnet	1397	clusters	884
statistical	4147	svm	1384	bayes	874
rules	4049	fuzzy	1352	lstm	833
networks	3820	linear	1247	cnn	819
vector	3578	inference	1234	regression	727
deep	3103	markov	1173	bidirectional	464
clustering	2454	recurrent	1169	word2vec	431
attention	2114	rulebased	1156	endtoend	419
graph	1803	sequences	1126	pretrained	184
embeddings	1694	ner	1078	encoder	145
means	1669	multimodal	962	transformer	21
latent	1572	naive	960	bert	0
random	1523	convolutional	916		
Usage after 2018					
neural	10802	bidirectional	1953	sequences	981
deep	9478	recurrent	1842	graphs	968
attention	6160	ner	1668	means	885
networks	5748	encoder	1663	svm	840
embedding	4128	random	1529	pos	804
embeddings	3710	clustering	1517	rulebased	795
graph	3562	rules	1497	linear	654
bert	3195	latent	1432	naive	641
vector	2985	statistical	1422	bayes	594
pretrained	2850	multimodal	1318	clusters	592
convolutional	2602	endtoend	1274	fuzzy	552
cnn	2148	inference	1263	wordnet	347
transformer	2026	regression	1167	markov	263
lstm	2013	word2vec	1049		
Change in usage after 2018					
bert	NA[10]	graph	1.976	clusters	0.67
transformer	96.476	neural	1.956	naive	0.668
pretrained	15.489	regression	1.605	clustering	0.618
encoder	11.469	recurrent	1.576	svm	0.607
embedding	4.571	ner	1.547	pos	0.576
bidirectional	4.209	networks	1.505	means	0.53
deep	3.054	multimodal	1.37	linear	0.524
endtoend	3.041	inference	1.024	fuzzy	0.408
attention	2.914	random	1.004	rules	0.37
convolutional	2.841	latent	0.911	statistical	0.343
cnn	2.623	sequences	0.871	wordnet	0.248
word2vec	2.434	vector	0.834	markov	0.224
lstm	2.417	rulebased	0.688	graphs	0.22
embeddings	2.19	bayes	0.68		

Some of the terms selected are very similar, like "embedding" and "embeddings", or refer to similar architectures like "recurrent" and "lstm". We can group and lemmatize

those terms and obtain a clearer picture. 4 groups were created based on the similarity of the words included in them, The groups will include the following terms:

- Group 1: bert, transformer, pretrained, encoder, decoder, attention, gpt
- Group 2: lstm, cnn, recurrent, convolutional, siamese
- Group 3: svm, means, bayes, random, graph, mean, regression, logistic, linear, vector, markov
- Group 4: fuzzy, rules, rulebased, valence, vader, pattern, matching, statistical, feature

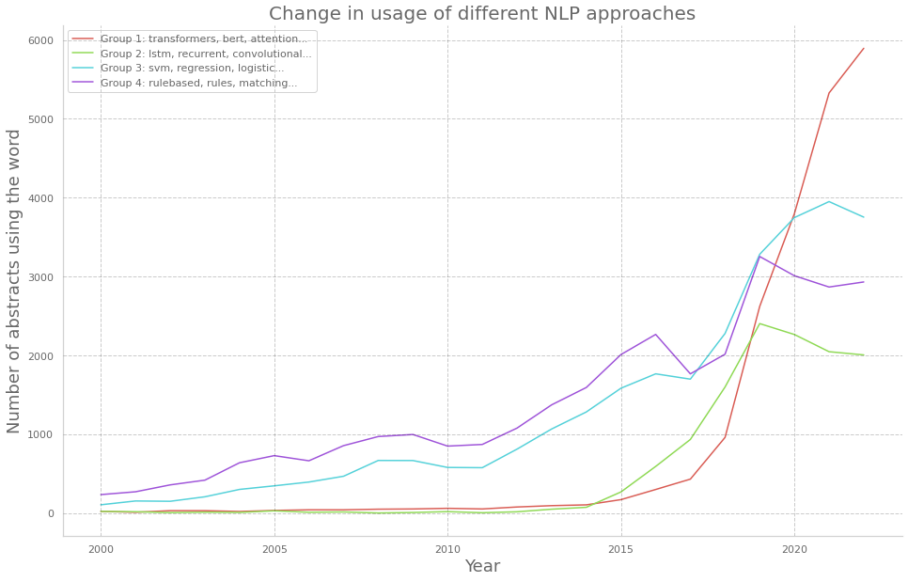


Fig. 3. Change in usage of different NLP approaches

It can be seen in fig 3 that the group related to transformers has had a massive growth in the past years, while the usage of other approaches seems to be slowing down. Further work can be done to take into account all the words and possible clusters from different NLP approaches, and there are limitations to this section, but this serves as another point to understand the impact that transformers, attention models, and BERT have had in the NLP research space.

We then propose the year 2018 as a turning point in NLP research. Even though transformers were proposed a year prior, in 2018 a paper was published, in which BERT was introduced [43]. BERT is a model, based on the transformer architecture that uses attention mechanisms to learn contextual relations between tokens to generate language understanding. Following the benefits of transformers, BERT is trained on large datasets of unlabeled text, using self-supervised training, and is then fine-tuned for specific NLP tasks. It introduced some novelties to transformers, such as being able to extract deep bidirectional representations, meaning that tokens before and after would affect the meaning of a token being paid attention to. BERT had 2 original variations, $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$,

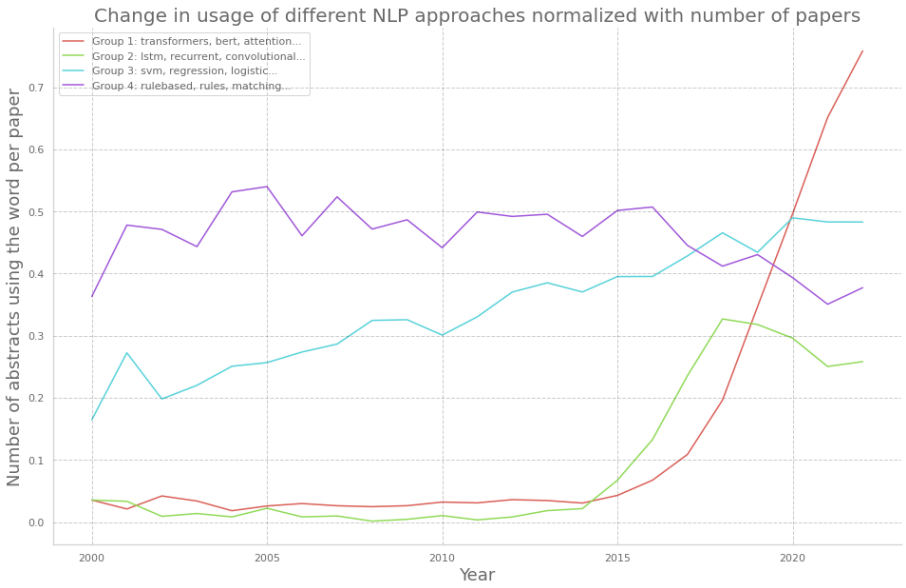


Fig. 4. Change in usage of different NLP approaches normalized with papers per year

where the main difference is the size of the model. $\text{BERT}_{\text{BASE}}$ had 12 layers, 12 attention heads, a total of 110M parameters, and a hidden layer size of 768. $\text{BERT}_{\text{LARGE}}$ had 24 layers, 16 attention heads, a total of 340M parameters, and a hidden layer size of 1024. The output of $\text{BERT}_{\text{BASE}}$ contains an embedding for every token in the input, with an embedding size of 768, and the output of $\text{BERT}_{\text{LARGE}}$ also contains an embedding for every token, with an embedding size of 1024. A trade-off has to be made when choosing what models to use for solving tasks, and in the case of the 2 BERT variations the trade-off is performance for computational expense. Even with transformers being so relatively efficient, bigger models appear to achieve better results, while also being computationally more expensive. In general, $\text{BERT}_{\text{LARGE}}$ is expected to achieve better performance than $\text{BERT}_{\text{BASE}}$ on tasks that require more fine-grained modeling of the input data, but at the cost of requiring more energy and memory to use.

he BERT model, in conjunction with the transformer architecture, has seemingly created a new wave of research within the field of NLP. As illustrated in Table 1, BERT has been cited 3195 times since its introduction in 2018, demonstrating greater relevance than other models like LSTMs or CNNs, which were proposed several years earlier. It is noteworthy that the data analyzed predominantly originates from recent publications, as depicted in Figure 5. Yet, this does not detract from the conclusion that BERT has significantly impacted NLP usage and research. BERT, alongside transformer and attention-based models, have established themselves as pivotal architectures. The trend in the adoption of these terms, as shown in Figures 6, 7, and 8, indicates that their usage has either increased or remained steady since their inception. A positive value in the graph indicates that the usage of the word increased, while a negative value indicates that the usage of the word decreased. A value of 0 in the graph indicates that the change remained the same for

that period. It is clear that research papers implementing the BERT model inherently use transformer and attention mechanisms, as BERT itself is built upon these architectures.

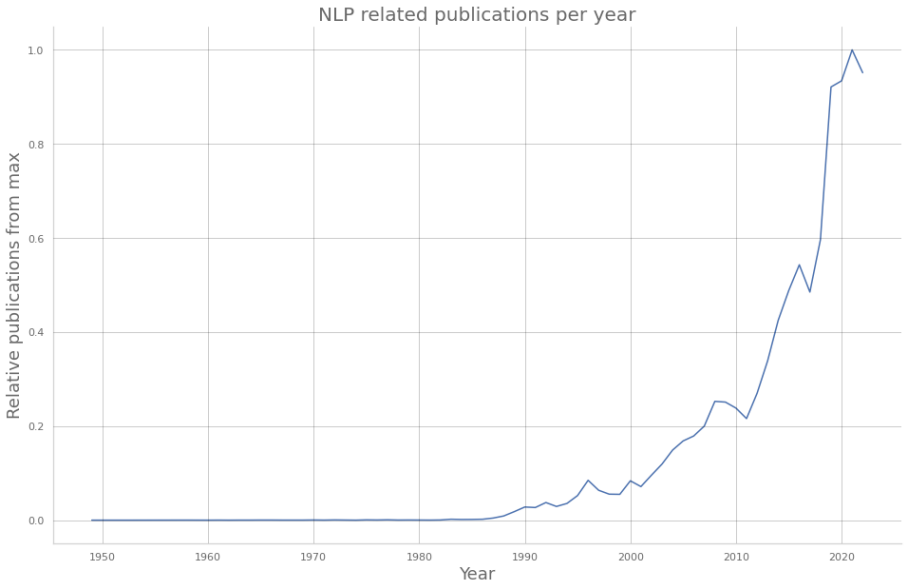


Fig. 5. NLP related publications per year

Analyzing the times BERT was mentioned in the articles since its introduction in 2018 in Table 3 furthers the understanding of its impact on the research community. It can be observed that the usage, or at least the research related to BERT, is at its highest point. These mentions include variations on the original BERT model, such as DistilBERT [116], BioBERT [83] or SciBERT [13]. BERT's versatility and its utilization of the transformer architecture and pre-training, fine-tuning approach make it accessible to a wide range of teams. This accessibility is facilitated by the availability of high-performance computing hardware or cloud-based computing services. These factors contribute to the ability of researchers to achieve state-of-the-art performance on a variety of natural language processing tasks, without the need for an excessive amount of labeled data or access to specialized computing resources.

Table 2. Number of articles mentioning BERT per year

Year	Papers mentioning BERT
2019	150
2020	434
2021	778
2022	802

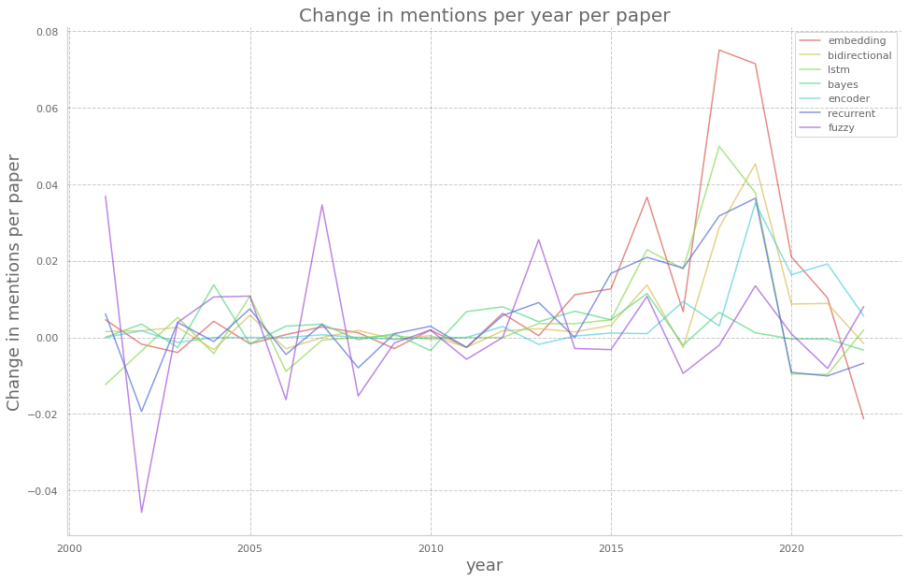


Fig. 6. Change in word usage per paper per year. A positive change means the word had more usage in the year prior by y per paper

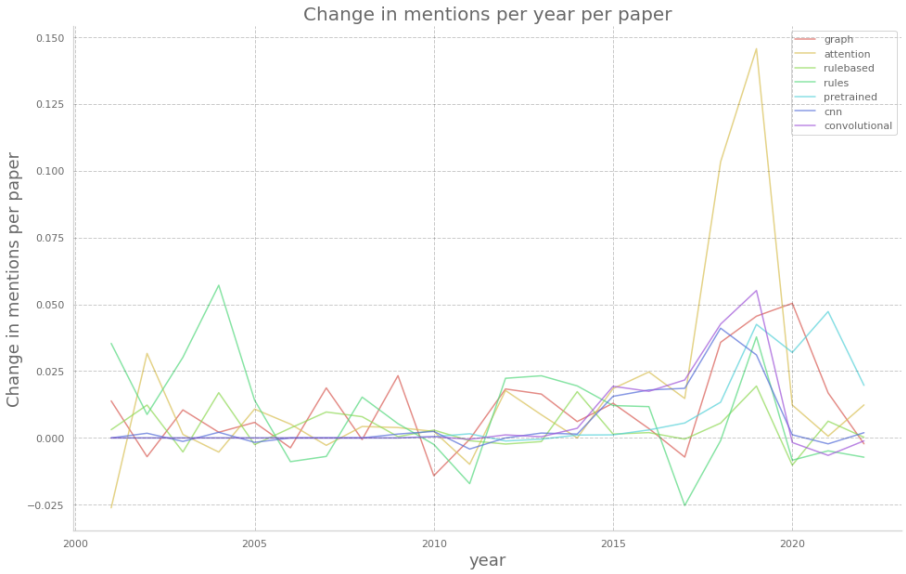


Fig. 7. Change in word usage per paper per year

A combination of these factors appears to have made BERT a good option for researchers. It provided state-of-the-art results in most NLP tasks, while only requiring

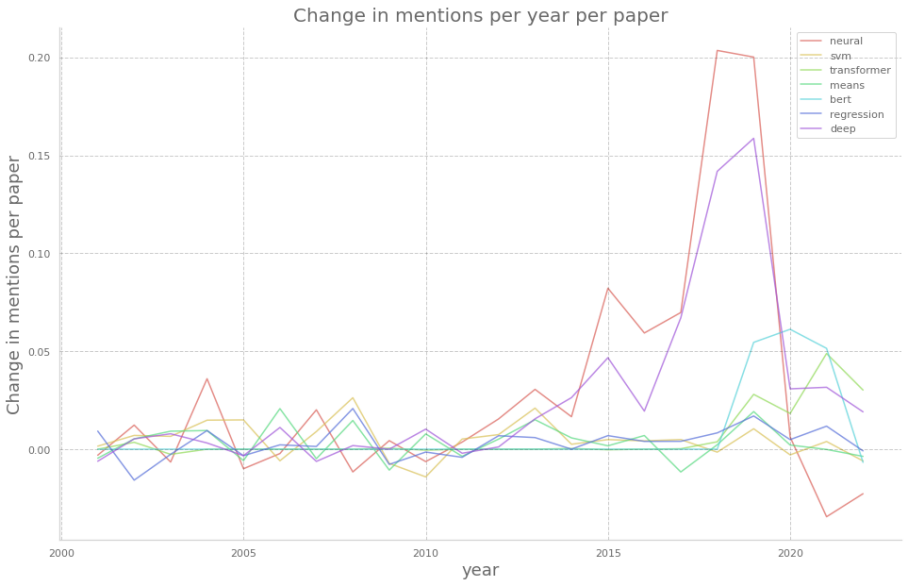


Fig. 8. Change in word usage per paper per year

comparatively small datasets for fine-tuning while being computationally cheap enough to be able to be run on smaller specialized hardware or on cloud computing services. The open-sourcing of the whole pre-trained model was also a likely factor for its widespread usage. Summarizing the reasons why researchers have been attracted to BERT:

- State-of-the-art performance: BERT has achieved state-of-the-art performance on a wide range of NLP tasks, outperforming previous models on benchmarks such as the GLUE and SQuAD datasets.
- Pre-training and fine-tuning: BERT's architecture is designed to be pre-trained on large amounts of unlabeled text data and then fine-tuned on smaller labeled datasets for specific tasks. This approach allows researchers to leverage the information contained in the pre-trained model, reducing the amount of labeled data needed for fine-tuning, and the computational cost of training a model from scratch.
- Versatility: BERT can be fine-tuned for a wide range of NLP tasks, including text classification, NER, and question answering, by simply changing the architecture of the final layer of the model.
- Bidirectionality: BERT's architecture uses a bidirectional transformer, which means that the model takes into account the context from both the left and the right side of a given word when generating embeddings. This allows improved performance in tasks where understanding the context is crucial, such as question answering.
- Large model size: BERT models are large, with hundreds of millions of parameters, which gives them the ability to model complex patterns and relationships within the text. This makes them suitable for handling large and complex datasets.
- Accessibility: BERT models are widely available, and pre-trained models, along with the fine-tuning code, can be easily accessed and used in many NLP tasks.

Even though BERT has had a big impact on the research community, other models with other architectures, pre-training datasets, training objectives, sizes, and overall design changes have been proposed. Those models achieve better performances in different tasks and may be more suitable for different situations and objectives.

4.3. Further Analysis of Abstracts

An experiment to show the capabilities of different models and approaches was conducted. A task was proposed to showcase how one might start to solve an NLP task, and what benefits and trade-offs the different models might provide in terms of performance, cost, and training data required. The training, and following labeling of abstracts, can also be used as further proof of the impacts transformers have had in the NLP field. An increase in papers using DNNs, especially the transformer architecture, is evident when looking at the labeled abstracts through time.

A more in-depth analysis of the abstracts was conducted to further understand the change in usage of different NLP techniques. All the abstracts were categorized into 5 different categories, based on which model or technique they used in the study. Four approaches were proposed for the categorization. The categories are:

- 0) The abstract does not mention a particular model or technique. Some papers analyzing frameworks, surveys, papers centered on the computer vision component of NLP, and dataset proposals among others fall into this category.
- 1) A model based on rules or symbolic analysis is used.
- 2) An approach using statistical methods is used. This includes BoWs, N-Grams, and TF-IDF, along with other machine learning techniques like SVMs, Logistic Regression, LDA, and others. Shallow neural network models like word2vec also belong in this category.
- 3) Approaches that use Deep Learning and other Deep Neural Network architectures such as RNNs, CNNs, and LSTM are included in this category.
- 4) The approach proposed uses transformer based models, like BERT, GPT, T5, and others. Although transformers are a type of DNN, they will be differentiated from other architectures included in category 3.

Since an abstract can contain conflicting categories, meaning that it could, for example, include both categories 3 and 4, the highest category is to be selected. Most abstracts would fall into category 1, as most of them include some form of rule based preprocessing. All of the abstracts in Category 4 would also belong in Category 3, as transformers are a subset of DNNs, and all DNNs would also fall into Category 2 as they use statistical methods in training and in architecture design. The classification of the abstracts was not done in a definitive manner, it was done to revise the capabilities of different models and their ability to adapt to the classification criteria that we proposed.

A small dataset of 384 labeled abstracts was created to train and test all the models. A split of 250 abstracts was dedicated to training, while the other 134 abstracts were used for testing. The distributions of the training dataset can be seen in Fig 9:

TF-IDF approach. A vector representation of the abstracts was created using TF-IDF. A K-nearest neighbors model was trained using 250 of the labeled abstracts. Afterwards, all the 75527 abstracts were categorized into one of the 5 categories and organized by year.

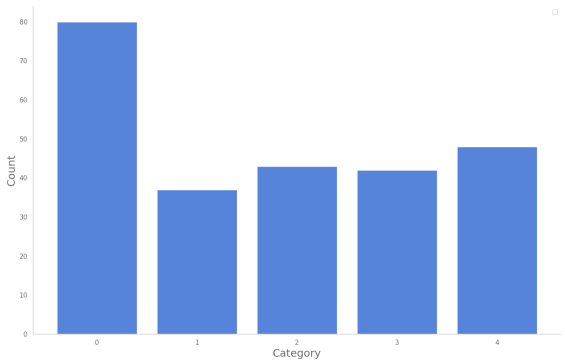


Fig. 9. Distribution of training dataset

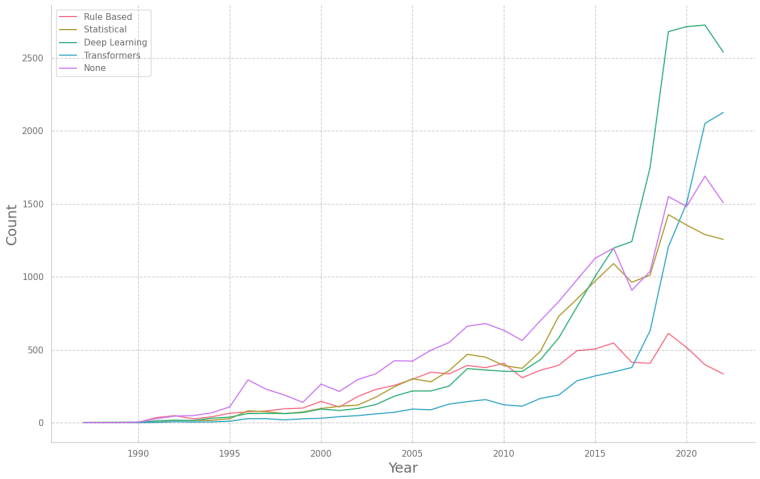


Fig. 10. TF-IDF KNN Predictions per Year

BERT model approach. A BERT based model was fine-tuned using 250 labeled abstracts. The fine-tuning used a loss to categorize each abstract into one of the 5 categories. Afterwards, all the 75527 abstracts were categorized into one of the 5 categories using the fine-tuned model.

The results immediately give BERT an edge, as any paper categorized in Category 4 before 2017 would have been miscategorized. BERT also obtained a significantly higher accuracy on the testing dataset. Another model like RoBERTa, or even some encoder-decoder models like FLAN-T5 could achieve better results with the same training param-

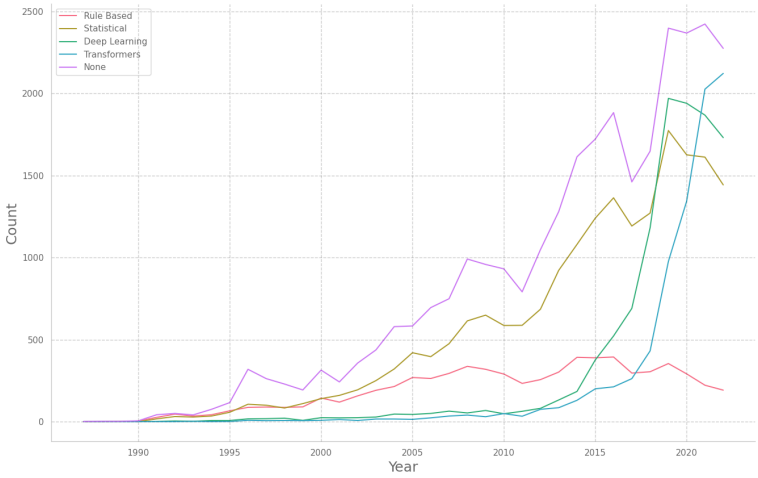


Fig. 11. BERT Predictions per Year

eters and resources, other than computer power required, but BERT was used as a baseline for a fine-tuning approach.

SetFit few-shot approach. Sentence Transformer Fine-tuning (SetFit) is a framework for few-shot fine-tuning, using contrastive learning, Pattern Exploiting Training (PET), and Parameter Efficient Fine-Tuning (PEFT) training [139]. Using Sentence [111] MPNet [127] [11] as the base model, the training was done on only 15 labeled examples per category, for a total of 75 labeled examples, in comparison to the 250 labeled examples of the previous approaches.

LLM few-shot approach. A Llama-2-70b model was used to categorize all the testing dataset abstracts. Not all the 75527 abstracts were categorized, as using this model is by far the most computationally expensive approach. A prompt was created using only one abstract per category and then passed to the LLama2-70b model along with the new abstract to categorize.

This approach was the least time consuming in human hours, as only 5 abstracts in total were categorized for the few-shot prompting. This approach also obtained similar accuracy on the test dataset compared to BERT.

Analysis Results. The results of this analysis of the abstracts reinforces the argument for the transformers architecture being an inflection point in NLP. The tradeoffs of the three different are more evident in this analysis. The TF-IDF approach was the less performant,

[11] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

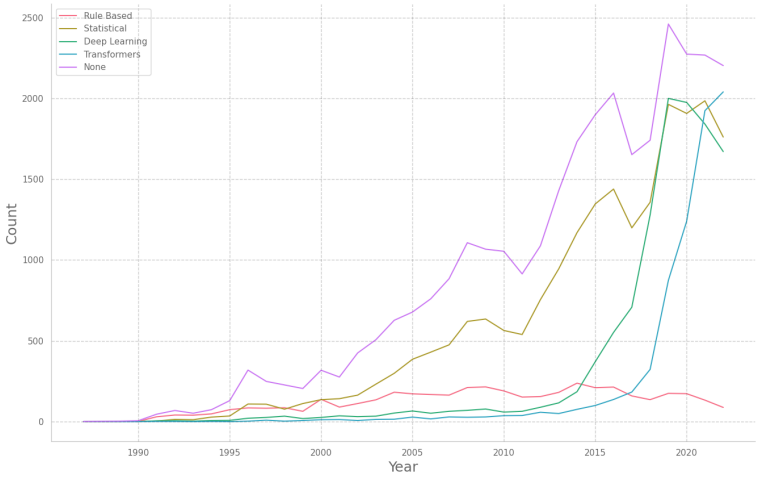


Fig. 12. SetFit Predictions per Year

but also the less computationally expensive. The BERT approach was a good balance between computer expense and performance. The few-shot Llama-2-70b approach was the most computationally expensive, but also obtained a good performance, similar to the one of BERT, and did not require extensive labeling of abstracts. On a classification task that has categories that are so loosely defined, it makes sense that models that are good at reasoning, like Llama-2, will have a good performance by default. Some examples of the answers Llama-2 gave are included 9. The dataset included [25] has the classification given by the models tested. Note that the classification should be used as a starting point for further use, as it may not be highly accurate.

Table 3. Different approaches on text classification

Approach	Accuracy	Labeled Samples	Computer power - Device Used
TF-IDF / KNN	0.4701	250	low - cpu
BERT	0.7164	250	medium - NVIDIA V100
SetFit	0.5970	75	medium - NVIDIA V100
Llama 2	0.7089	5	high - 2x NVIDIA A100

With the results into consideration, the following advantages and disadvantages for the models can be outlined:

TF-IDF with KNN: This model, employing classical NLP techniques, served as a baseline to gauge performance against more traditional methods. It was the least computationally demanding and least performant, highlighting the advancements in NLP techniques over time.

BERT-based model: Representing advanced methods, BERT utilized fine-tuning to categorize abstracts achieving higher accuracy. BERT's results underscore the superiority of current state-of-the-art models and their balance of computational cost and performance.

SetFit few-shot approach: This approach demonstrated that with few-shot learning, it is possible to train models efficiently with a minimal amount of labeled data. While not as performant as BERT, it offers a middle ground for projects with limited data availability.

LLM few-shot approach: Using Llama-2-70b, this method explored the capabilities of large language models, which require significant computational resources but minimize human labor in terms of data labeling, setup, and training logistics. Despite its expense, this method competes in accuracy with BERT, showing promise for high-efficiency applications.

Each model provided unique insights, with BERT emerging as a strong all-around option for balancing cost, data requirements, and performance. Llama-2-70b showcased the potential for achieving similar accuracy with less labeled data but at a greater computational cost. The findings illustrate that the choice between models depends on the specific needs and constraints of the project at hand. Models with lower computational costs may suffice for certain tasks, while advanced models may be necessary for applications demanding high accuracy and efficiency. This experiment serves as a foundation for selecting appropriate models in NLP tasks and for understanding the ongoing advancements within the field.

All the models could improve their performance using different techniques and higher quality data, but this serves as a good baseline to contrast the advantages and disadvantages of the different approaches.

It is important to note that the technique used with Llama-2-70b, based on prompting rather than training, appears to be gaining traction, and with models that allow a greater context size and better understanding, it might show even better results in the near future.

4.4. Extra considerations for transformers

While the architecture of Transformer models is a critical determinant of their performance, other factors also play significant roles in achieving optimal results. Understanding and fine-tuning these aspects can significantly enhance a Transformer's ability to generalize and excel on a wide range of tasks

Regularization Techniques Overfitting is a common challenge with complex models, including Transformers. Regularization techniques such as dropout, label smoothing, and weight decay can be implemented to encourage the model to develop a more general representation and prevent over reliance on specific training examples.

Attention Mechanism Adjustments The attention mechanism is the heart of Transformer models. Adjusting the attention heads, exploring different types of attention (such as local, global, or sparse attention), and experimenting with the attention flow can lead to improvements in model performance and efficiency.

Tokenizers As mentioned earlier, DL models, including BERT, typically process input data that has been tokenized and represented as numerical vectors, rather than raw text. This is a common practice in NLP as it allows the model to more easily manipulate and analyze the input data. The tokenizer used for BERT was originally Wordpiece [157], a subword tokenizer, meaning that it breaks words up into smaller units called subwords or word pieces, rather than using fixed-size word tokens like many other tokenizers. Wordpiece breaks down words into smaller units based on the most frequent character sequences in the training data. It then inserts two special tokens, [CLS] and [SEP], into the input sequence. [CLS] is placed at the beginning of every input, while [SEP] is inserted between every pair of input sequences. These tokens are used by BERT to mark the start and end of sentences and to distinguish between different sentences within an input. Computational word representations and tokenizers are still a point of improvement and difference between models. Some of the more commonly used tokenizers are [157], [78] and byte pair encoding (BPE) [120]

A good tokenizer is crucial for the performance of a model, as it will dictate what are the minimum elements that the language will hold. A balance of simple objects, such as "a", "I", "am", or "is", and more complicated objects like "individual", "account" or "program" is imperative for both efficient and performant generation of tokens.

The tokens generated by the tokenizer will then be transformed into embeddings. Initially, each token gets mapped to an integer, to then be mapped as a vector representing the token on an embedding layer of the transformer. These embeddings hold the spatial information of the token, so they can be processed by the model.

The paragraphs at 9 were generated using GPT-3, and it will be used to showcase how the tokenizers of different models would split text. Each of the words separated by a space represents a token.

Datasets Datasets are a fundamental part of the pre-training and fine-tuning of models. A good, refined and rich set of training data is essential for the performance of complex models such as transformers. There is an extensive list of datasets aimed at different tasks and a variety of datasets that compile massive amounts of text from different sources. GLUE [146], a collection that includes data from other datasets, such as [149,153,126], QQP[12], STS-B[13] among others, has been popular for its variety. The Pile, an 825GiB dataset constructed on 22 subsets [47] has been used to train models like [15]. BookCorpus [169], a dataset containing a large number of data from books, used to train BERT, ALBERT, and XLNet among others. SQuAD 2.0 [110] has also been used to fine-tune models like BERT. Other relevant datasets include RedPajama [38] or RefinedWeb[105].

Benchmarks Benchmarks are among the most reliable methods for assessing the improvements a model brings. Many datasets are commonly used as benchmarks to evaluate a model's performance. Some examples are RACE [80], SQuAD 2.0, GLUE, XTREME [65], or BIG-bench [129].

It is important to note that no perfect benchmark has been developed yet, and there are questions over their reliability and usability as a tool to measure the actual performance of a model [94,49].

[12] data.quora.com/First-Quora-Dataset-Release-Question-Pairs

[13] http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark

5. Transformers in Practice: Model Variants

5.1. Some Variations on BERT

All of the next models improve on some aspects of BERT, All of them being encoder-only models.

ALBERT: A Lite BERT achieved state-of-the-art performance on the GLUE, RACE, and SQuAD benchmarks, being better than BERT while also having fewer parameters than $\text{BERT}_{\text{LARGE}}$, making training faster and requiring less memory usage. It uses three main techniques that differ from BERT, being factorized embedding parameterization, cross-layer parameter sharing, and sentence-order prediction as an objective [82]

DistilBERT: It is a variation on the BERT model, focusing on reducing the computational power and memory required to run it while being faster and still maintaining many of BERT's capabilities. It uses a model compression technique called knowledge distillation [21], in which a model, the "student" is trained on a model, the "teacher", that will be giving a distillation loss to the "student". [116]

ELECTRA: A new method for pre-training, replacing token detection, is proposed as a more efficient alternative to previous ones, including MLM. The model proposed with this approach, ELECTRA, outperforms other models, like BERT or XLNet, using a similar model size. By predicting if a token was or not replaced, these efficiency gains can be obtained. [35]

DeBERTa: It improves on the BERT and RoBERTa architectures, by introducing an enhanced masked decoder to improve on positional data, and a disentangled attention mechanism, making it so each word is represented by two separated vectors encoding their content and positions. [60]

RoBERTa: Under the finding that BERT was undertrained, a newer model that improves on it is developed. RoBERTa was trained using more data, more training time, longer batches, removing the Next Sentence Prediction (NSP) objective, using longer sequences, and changing the masking objective dynamically. [90]

5.2. Decoder models

Decoder models, known for their auto-regressive nature, play a pivotal role in various generative tasks, thanks to their unique architecture that enables sequential generation. This auto-regressive property means that these models generate output tokens one by one, conditioned on the previously generated tokens, allowing them to capture intricate dependencies and context in the data they produce. While decoder models boast impressive capabilities, they are not without their limitations. The auto-regressive generation process

can be slow, as each token's generation is dependent on previous tokens, leading to increased inference times, especially for long sequences. Additionally, this sequential generation approach can make it challenging to parallelize the inference process efficiently. Despite their limitations [87], they have demonstrated their capability to create the current state-of-the-art models. Some of these challenges are being currently researched.

Some relevant models are Chinchilla [63], which works on the scaling of Large Language Models (LLMs), and how to optimize pre-training on them, making it more efficient; GPT-3 [19], which achieves a few-shot performance comparable to those models that are fine-tuned; OPT [163], an open-source model based on GPT-3, having similar performance while being more energy efficient on the development stage; Galactica [134], a model focused on scientific tasks, achieving state-of-the-art performance on scientific benchmarks; BLOOM [156], an open access multilingual model; PaLM [32], a dense model, having a massive size of 540B parameters, to study the impact of scaling on language models; GLaM [45], that uses sparsity to scale LLMs, reaching an impressive 1.2T parameters, while still being less computationally expensive to train than GPT-3.

Most of the current state-of-the-art models fall into the decoder only architecture. It includes models like GPT4, LLama 2, Mistral 7b [72], Zephyr 7b, PaLM 2 or Claude 2 [14]

6. Results: General guidance

6.1. Performance

With models like Chinchilla [63] and Galactica [134] it can be observed that the quality and quantity of data used in pre-training plays an important role in the performance of the model. This makes it so that the size of the model, including parameters, number of attention heads, layers, and other hyper-parameters, are not the only important components when deciding on what model to use in regard to performance. A model like Galactica could be more suited to working with scientific data. Some domain-specific versions of some models have been published, including :

- SciBERT, a model based on BERT, pre-trained on a large multi-domain corpus of scientific publications to label scientific data [13].
- BioBERT, a model with largely the same architecture as BERT, pre-trained on large-scale biomedical corpora to perform text mining on biomedical text [83].
- FinBERT, a model based on BERT, trained to solve NLP tasks related to finances [7].
- LEGAL-BERT, states that the guidelines for pre-training BERT-like models may not be as suitable for the legal domain, proposing an investigation on available strategies for specialized domains. It was pre-trained on legislation, court cases, and contracts [27].
- ClinicalBERT, pre-trained on clinical data and fine-tuned to predict hospital re-admissions [66].
- DSGPT, proposing a domain-specific generative pre-training method, and then creating a model based on it, applied to E-commerce title and review summarization [164].

[14] https://www.anthropic.com/index/claude-2

The languages on which LLMs are pre-trained also have an impact on their performance. English seems to be the most common language for model releases, but some language-specific variations of some models have been created:

- Japanese: Refs [95,70,123,141] create variations on the BERT model to be able to handle different NLP tasks on the Japanese language. A variation on GPT-2 for story co-creation in Japanese was made in Ref [100]
- Spanish: Refs [3,26,145,48] create variations on BERT for the Spanish language. A model to solve the Reading Comprehension and Reasoning Explanation for Spanish was created [22] in Ref [113], using a variation on GPT-J.
- Chinese: Variations on the BERT architecture have been implemented in Ref [71,39,88].
- Multilingual: Some multilingual models have been proposed in Refs [159,161,156]

6.2. Running the Models

The idea of pre-training and then fine-tuning a model in NLP is to leverage the possible language understanding created by training the model on large datasets, and then fine-tune it to solve specific tasks. This is an approach that enables different groups that may not possess the computational power to train LLMs to still be able to use them in production or research. Although some computational power and memory will still be needed, cloud computing services provide a way to, without purchasing dedicated computing equipment and servers, be able to run the LLMs. There are some important characteristics that may indicate how much computational resources will be needed to run the models. The quantity of parameters a dense model has is generally a good way to estimate the computational resources needed to run the model: bigger models, like OPT, or BLOOM, will require more resources than GPT-Neo or GPT2.

In some cases using a smaller model may be needed, but techniques such as distillation [61,73,116], or module substitution [158] among others [53,57,130], can provide similar performance while keeping the model small, like in the case on DistilBERT and BERT. Although the size of the state-of-the-art models seems to growing [109], most likely faster than the growth in computational power, even considering High Performance Computing (HPC) and distributed computing [144,67,97], smaller models still can achieve good results on most NLP tasks.

6.3. Finding newer models

As mentioned prior, newer models have some limitations as to who can train and use them. As it seems, big groups, be it open-source communities, corporations, or institutions, are the ones creating the new state-of-the-art models. Most of these organizations have communication channels and blogs [15], along with other regular publications from AI research centers or experts.

[15] https://huggingface.co/blog, https://ai.googleblog.com/, https://www.eleuther.ai/, https://ai.facebook.com/blog/, https://bigscience.huggingface.co/blog, https://www.deepmind.com/, https://news.microsoft.com/source/topics/ai/

Model (Including variations)	Parameters	Hidden State Size	Layers	Attention Heads	Main association	Availability[16]	Paper publication date[17]
Encoder Architecture							
BERT	110M - 340M	768 - 1024	12 - 24	12 - 16	Google	Open - Model versions available	2018-10-11
ALBERT	11M - 223M	768 - 4096	12 - 24	12 - 64	Google	Open - Model versions available	2019-09-26
DistilBERT	66M	768	6	12	HuggingFace	Open - Model versions available	2019-10-02
ELECTRA	110M - 335M	256 - 1024	12 - 24	4 - 16	Google	Open - Model versions available	2020-03-23
RoBERTa	125M - 355M	768 - 1024	12 - 16	12 - 16	Meta	Open - Model versions available	2019-07-26
LUKE	483M	1024	24	16	Studio Ousia	Open - Model versions available	2020-10-02
Longformer	149M-435M	768-1024	12-24	12-16	Allen Institute	Open - Model versions available	2020-04-10
Decoder Architecture							
LLaMa	6.7B - 65.2B	4096 - 8192	32 - 80	32 - 64	Meta	Available under non commercial, research license	2023-02-27
LLaMa 2[18]	7B - 70B	4096 - 8192	32 - 80	32 - 64	Meta	Open - Model Versions Available	2023-07-18
GPT	110M	768	12	12	OpenAI	Open - Model versions available	2018-06-10
GPT2	117M - 1542M	768 - 1600	12 - 48	12 - 25	OpenAI	Open - Model versions available	2019-07-28
GPT3	125M - 175B	768 - 12288	12 - 96	12 - 96	OpenAI	Closed - Access via API	2020-05-28
GPT-J	6B	4096	28	16	EleutherAI	Open - Model versions available	2021-07-07
GPT-Neo	125M - 2.7B	2048	24	16	EleutherAI	Open - Model versions available	2021-05-21
GPT-NeoX-20B	20B	6144	44	64	EleutherAI	Open - Model versions available	2022-04-14
OPT	125M - 175B	768 - 12288	12 - 96	12 - 96	Meta	Open - Model versions available	2022-05-02
Chinchilla	70M	8192	80	64	DeepMind	Closed	2022-05-29
LaMDA	2B - 137B	2560 - 8192	10 - 64	40 - 128	Google	Closed-Architecture implementation available	2022-06-20
Galactica	125M - 120B	768 – 10240	12 - 96	12 - 80	Meta	Open - Model versions available	2022-11-16
BLOOM	550M - 176B	1024 - 14336	24 - 70	16 - 112	BigScience	Open - Model versions available	2022-11-09
PaLM	8B - 540B	4096 - 18432	32 - 118	16 - 48	Google	Closed-Architecture implementation available	2022-04-05
GlaM	1.9B - 1.2T[19]	3072 - 65536	12 - 64	12 - 128	Google	Closed	2021-12-13
MEGATRON-TURING	530B	20480	105	128	NVIDIA-Microsoft	Closed - Codebase available	2022-01-28
GLM-130B	130B	12288	70	96	Tsinghua U	Open – Model versions available	2022-10-05
FLAN	137B	8192	64	128	Google	Closed	2021-09-03
FLAN-PaLM	8B-540B	4096-18432	32-118	16-48	Google	Closed	2022-04-05
Falcon	180B	14848	80	64	TII[20]	Open – Model versions available	2023-10-12
Encoder-Decoder Architecture							
XLNet	110M - 340M	768 - 1024	12 - 24	12 - 16	CMU, Google	Open - Model versions available	2019-06-19
T5	60M - 11B	512 - 1024	6 - 24	8 - 128	Google	Open - Model versions available	2019-10-23
Tk-Instruct	60M - 11B	512 - 1024	6 - 24	8 - 128	Various Institutions[21]	Open - Model versions available	2022-04-16
FLAN	60M - 11B	512 - 1024	6 - 24	8 - 128	Google	Open - Model versions available	2021-09-03

7. Future

In the past five years, significant advancements have been made in the field of NLP. There are different directions that newer models seem to be taking to improve performance and computational expense. Making bigger models appears to be a good way to improve performance, but the computational cost makes it so those models use more energy and take longer to train. For these massive models, specialized methods and hardware may be needed to train them [125,97,67].

Using bigger, cleaner datasets can be another way to improve models, without the need to make them bigger. Processing larger datasets can still signify an increase in the energy needed to pre-train the models, but the impact is smaller than increasing the size of the model [134,63].

Decoder-only, auto-regressive, generative models seem to be the new approach to achieving state-of-the-art results on many tasks, even those that were previously dominated by BERT-like, fine-tuned models. GPT-3, Galactica, and others prove that auto-regressive models appear to be few-shot learners [19,134], while also showing some emergent capabilities [150]. The drawback of these generative models is that they have been considerably larger than the encoder-only models, taking more energy and memory to utilize. Parameter growth can be seen in fig 13.

[16] Most open models are available at Huggingface [1] and the transformers library [154]

[17] Publications on developer blogs or other communications could have been made earlier. The date of the largest model is considered

[18] LLaMa 2 has been made open [138]

[19] GLaM has ~145M-96.6B activated parameters

[20] Technology Innovation Institute

[21] Allen Institute for AI, Univ. of Washington, Arizona State Univ., Sharif Univ. of Tech., Tehran Polytechnic, PSG College of Tech., IIT Kharagpur, Univ. of Amsterdam, UC Berkeley, Columbia Univ. , Factored AI, Govt. Polytechnic Rajkot, Microsoft Research, Stanford Univ., Zycus Infotech, Univ. of Massachusetts Amherst, National Inst. of Tech. Karnataka, TCS Research, IIT Madras, National Univ. of Singapore, Johns Hopkins Univ.

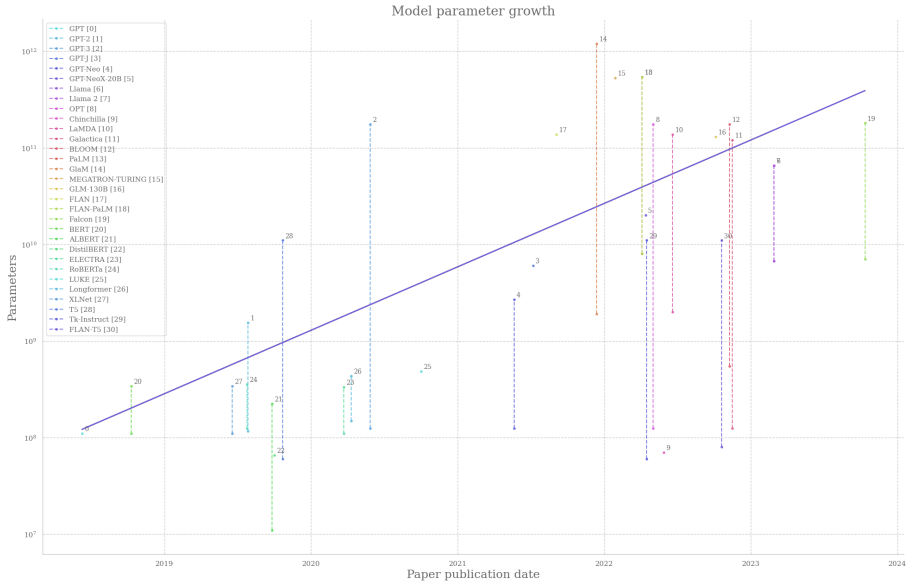


Fig. 13. Model size growth

A wave of research on model to human, consumer, interaction has been developing. Concerns over the safety and abuse of models have been studied [20,14]. LaMDA is a model that improves on this human interaction, achieving more safety and factual grounding [136]. Models like InstructGPT [101], which most likely paved the way to OpenAI's ChatGPT, have used techniques involving Reinforcement Learning (RL), specifically reinforcement learning from human feedback (RLFH) [33] to try to solve alignment issues.

Further advancements in solving the limitations of LLM usage have been made. Foundational models like LLaMa [137], which achieve good performance even with fewer parameters, that can be run on consumer hardware, have been created. Open, permissive versions of LLaMa have been replicated [52]. Improved models that use RLHF, building on LLaMa and other models have also been created [138,132,29,79,92].

Even with the efforts from open-source communities, with models like BLOOM, having direct access to massive models with billions or even trillions of parameters is difficult due to their size. Artificial Intelligence as a Service (AIaaS) may be a way to put high-performance models on production without the need to worry about multi-accelerator setups or expensive equipment. Nevertheless, researchers are conducting investigations into smaller models and compression techniques, yielding promising results and the potential to deploy LLMs on low power devices.

In recent months, there has been remarkable progress in the development of newer models that have achieved state-of-the-art performance. One such model is GPT-4 [99], which has demonstrated human-level performance on numerous benchmarks, while outperforming most other models. GPT-4 has the advantage of being multimodal, giving it extended capabilities. This model, currently available under an API, will most likely be the new standard to beat for new coming models. Not many details about its architecture

have been given under official channels, so it is hard to make statements about it, which seems to be the new trend for new non open-source models, as product competitivity starts to grow.

Another notable achievement is attributed to PaLM 2 [6], representing a significant advancement in language modeling, incorporating techniques that enhance its understanding and generation of text

The open-source space of LLMs has also gained traction. Newer models like LLama 2 [138], Falcon [5], Mistral 7B [72], INCITE [22], MPT [135] have achieved promising results.

Improvements in the efficiency of training and usage of LLMs have been made. PEFT techniques such as LoRA [64], QLoRA [42], Prefix-Tuning [85], P-Tuning [89] and AdaLoRA [162] are some methods to efficiently fine-tune LLMs.

New frameworks, projects and tools have been created in order to help those that are building with NLP. LangChain [23] provides different tools to reach a production-ready app faster. LlamaIndex [24] provides ways to connect, structure and retrieve data. Haystack [25] also provides different tools to integrate LLMs in NLP apps. Vector databases and vector stores have also gained popularity to structure NLP projects and applications. Some examples include FAISS [74], pgvector for Postgres [26], Chroma [27] and Vespa [28]

Some challenging areas that require improvement in LLMs are enhancing their reasoning and factual capabilities, despite RLHF and Chain-of-Tought [151] providing a partial pathway to address these issues. Long term memory and context length are also areas in which improvements can be made. Retrieval-Augmented Generation (RAG) [84] has also been used to provide a form of memory to LLMS, and to generate text that is more informative, factually correct, and contextually relevant.

There has also been extensive, yet not conclusive, research on the different implications of these new NLP models and approaches. Scholars can develop new lines of research around ethics in AI [98], the design and supervision of algorithms [112] , according to a particular emphasis on business practices [9]. A gap that researchers must close in the future is the exploration of ethical challenges of LLMs such as GPT-4 through experiments and robust diagnostic models [170].

NLP has emerged as a significant field of research, and it continues to expand. Many organizations have already adopted, and will continue to integrate, the techniques and models discussed in this study. Consequently, it is essential to anticipate the future landscape of NLP, recognizing current trends and the insights they offer. Bearing this in mind, it is expected that NLP will be increasingly applied across various industries, with its effects becoming more evident.

---

[22] https://huggingface.co/togethercomputer/RedPajama-INCITE-7B-Base
[23] https://github.com/langchain-ai/langchain
[24] https://github.com/run-llama/llama_index
[25] https://github.com/deepset-ai/haystack
[26] https://github.com/pgvector/pgvector
[27] https://github.com/chroma-core/chroma
[28] https://github.com/vespa-engine/vespa

8. Conclusion

With this work, we have given some insight into the area of NLP. With this information, researchers, organizations, and individuals will be able to delve into NLP with current models, especially LLMs, and their different approaches. Further research can also be conducted in a guided way, having more context of where current NLP trends came from, and where they may be going.

It is important to note that some of the information in this paper will most likely become outdated in a short time, because, as previously discussed, the field of NLP is moving at an ever increasing speed. The general recommendations will, however, remain useful for researchers, individuals, and institutions that want to approach the field of NLP.

References

1. Hugging face - the ai community building the future., https://huggingface.co/
2. Agarwal, M.: An overview of natural language processing. International Journal for Research in Applied Science and Engineering Technology (IJRASET) 7(5), 2811–2813 (2019)
3. Akhtyamova, L.: Named entity recognition in spanish biomedical literature: Short review and bert model. In: 2020 26th Conference of Open Innovations Association (FRUCT). pp. 1–7. IEEE (2020)
4. Allen, J., Hunnicutt, M.S., Klatt, D.H., Armstrong, R.C., Pisoni, D.B.: From text to speech: The MITalk system. Cambridge University Press (1987)
5. Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., Noune, B., Pannier, B., Penedo, G.: Falcon-40B: an open large language model with state-of-the-art performance (2023)
6. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J.H., Shafey, L.E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Abrego, G.H., Ahn, J., Austin, J., Barham, P., Botha, J., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C.A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, A., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P., Ros, A.C., Roy, A., Saeta, B., Samuel, R., Shelby, R., Slone, A., Smilkov, D., So, D.R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Y., Xu, K., Xu, Y., Xue, L., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S., Wu, Y.: Palm 2 technical report (2023)
7. Araci, D.: Finbert: Financial sentiment analysis with pre-trained language models (2019), https://arxiv.org/abs/1908.10063
8. Plaza-del Arco, F.M., Molina-González, M.D., Urena-López, L.A., Martín-Valdivia, M.T.: Comparing pre-trained language models for spanish hate speech detection. Expert Systems with Applications 166, 114120 (2021)
9. Attard-Frost, B., De los Ríos, A., Walters, D.R.: The ethics of ai business practices: a review of 47 ai ethics guidelines. AI and Ethics 3(2), 389–406 (2023)

10. Backus, J.W., Bauer, F.L., Green, J., Katz, C., McCarthy, J., Naur, P., Perlis, A.J., Rutishauser, H., Samelson, K., Vauquois, B., et al.: Revised report on the algorithmic language algol 60. The Computer Journal 5(4), 349–367 (1963)
11. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2016)
12. Bao, T., Ren, N., Luo, R., Wang, B., Shen, G., Guo, T.: A bert-based hybrid short text classification model incorporating cnn and attention-based bigru. Journal of Organizational and End User Computing (JOEUC) 33(6), 1–21 (2021)
13. Beltagy, I., Cohan, A., Lo, K.: Scibert: Pretrained contextualized embeddings for scientific text. CoRR abs/1903.10676 (2019), http://arxiv.org/abs/1903.10676
14. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. p. 610–623. FAccT '21, Association for Computing Machinery, New York, NY, USA (2021), https://doi.org/10.1145/3442188.3445922
15. Black, S., Leo, G., Wang, P., Leahy, C., Biderman, S.: Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow (Mar 2021), https://doi.org/10.5281/zenodo.5297715
16. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
17. Bradbury, J., Frostig, R., Hawkins, P., Johnson, M.J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., Zhang, Q.: JAX: composable transformations of Python+NumPy programs (2018), http://github.com/google/jax
18. Brown, R.W.: Linguistic determinism and the part of speech. The Journal of Abnormal and Social Psychology 55(1), 1 (1957)
19. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020), https://arxiv.org/abs/2005.14165
20. Buchanan, B., Lohn, A., Musser, M., Sedova, K.: Truth, lies, and automation. Center for Security and Emerging Technology 1(1), 2 (2021)
21. Bucilua, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 535–541 (2006)
22. Cabezudo, M.A.S., Diestra, D., López, R., Gómez, E., Oncevay, A., Alva-Manchego, F.: Overview of recores at iberlef 2022: Reading comprehension and reasoning explanation for spanish. Procesamiento del Lenguaje Natural 69, 281–287 (2022)
23. Cambria, E., White, B.: Jumping nlp curves: A review of natural language processing research [review article]. IEEE Computational Intelligence Magazine 9(2), 48–57 (2014)
24. Canchila, S., Meneses-Eraso, C., Casanoves-Boix, J., Cortés-Pellicer, P., Castelló-Sirvent, F.: Indexed NLP article classifier models (2023), https://doi.org/10.7910/DVN/X0FTLZ
25. Canchila, S., Meneses-Eraso, C., Casanoves-Boix, J., Cortés-Pellicer, P., Castelló-Sirvent, F.: Indexed NLP Article Metadata Dataset (2023), https://doi.org/10.7910/DVN/5YIGNG
26. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. Pml4dc at iclr 2020(2020), 1–10 (2020)
27. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: Legal-bert: The muppets straight out of law school (2020), https://arxiv.org/abs/2010.02559

28. Chen, S., Zhang, Y., Yang, Q.: Multi-task learning in natural language processing: An overview. arXiv preprint arXiv:2109.09138 (2021)
29. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/
30. Chomsky, N.: Three models for the description of language. IRE Transactions on Information Theory 2(3), 113–124 (1956)
31. Chomsky, N.: Systems of syntactic analysis. The Journal of Symbolic Logic 18(3), 242–256 (1953)
32. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., Fiedel, N.: Palm: Scaling language modeling with pathways (2022), https://arxiv.org/abs/2204.02311
33. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf
34. Church, K.W.: Word2vec. Natural Language Engineering 23(1), 155–162 (2017)
35. Clark, K., Luong, M., Le, Q.V., Manning, C.D.: ELECTRA: pre-training text encoders as discriminators rather than generators. CoRR abs/2003.10555 (2020), https://arxiv.org/abs/2003.10555
36. Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F.: Scimat: A new science mapping analysis software tool. Journal of the American Society for Information Science and Technology 63(8), 1609–1630 (2012)
37. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. p. 160–167. ICML '08, Association for Computing Machinery, New York, NY, USA (2008), https://doi.org/10.1145/1390156.1390177
38. Computer, T.: Redpajama: An open source recipe to reproduce llama training dataset (April 2023), https://github.com/togethercomputer/RedPajama-Data
39. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z.: Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29, 3504–3514 (2021)
40. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science 41(6), 391–407 (1990)
41. Derrick, G., Jonkers, K., Lewison, G.: Characteristics of bibliometrics articles in library and information sciences (lis) and other journals. In: Proceedings, 17th international conference on science and technology indicators. STI. pp. 449–551 (2012)
42. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms (2023)
43. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018), http://arxiv.org/abs/1810.04805

44. Di Gennaro, G., Buonanno, A., Palmieri, F.A.: Considerations about learning word2vec. The Journal of Supercomputing pp. 1–16 (2021)
45. Du, N., Huang, Y., Dai, A.M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A.W., Firat, O., Zoph, B., Fedus, L., Bosma, M., Zhou, Z., Wang, T., Wang, Y.E., Webster, K., Pellat, M., Robinson, K., Meier-Hellstern, K., Duke, T., Dixon, L., Zhang, K., Le, Q.V., Wu, Y., Chen, Z., Cui, C.: Glam: Efficient scaling of language models with mixture-of-experts (2021), https://arxiv.org/abs/2112.06905
46. Fedorova, E., Druchok, S., Drogovoz, P.: Impact of news sentiment and topics on ipo underpricing: Us evidence. International Journal of Accounting & Information Management 30(1), 73–94 (2022)
47. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al.: The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020)
48. García-Pablos, A., Perez, N., Cuadros, M.: Sensitive data detection and classification in spanish clinical text: Experiments with bert. arXiv preprint arXiv:2003.03106 (2020)
49. Gardner, M., Merrill, W., Dodge, J., Peters, M., Ross, A., Singh, S., Smith, N.A.: Competency problems: On finding and removing artifacts in language data. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 1801–1813. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021), https://aclanthology.org/2021.emnlp-main.135
50. Gbur Jr, E.E., Trumbo, B.E.: Key words and phrases—the key to scholarly visibility and efficiency in an information explosion. The American Statistician 49(1), 29–33 (1995)
51. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1243–1252. PMLR (06–11 Aug 2017), https://proceedings.mlr.press/v70/gehring17a.html
52. Geng, X., Liu, H.: Openllama: An open reproduction of llama (May 2023), https://github.com/openlm-research/open_llama
53. Gordon, M.A., Duh, K., Andrews, N.: Compressing bert: Studying the effects of weight pruning on transfer learning (2020), https://arxiv.org/abs/2002.08307
54. Graves, A., Fernández, S., Schmidhuber, J.: Bidirectional lstm networks for improved phoneme classification and recognition. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005. pp. 799–804. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)
55. Grice, H.P.: Logic and conversation. In: Speech acts, pp. 41–58. Brill (1975)
56. Grishman, R.: Information extraction: Techniques and challenges. In: Pazienza, M.T. (ed.) Information Extraction A Multidisciplinary Approach to an Emerging Information Technology. pp. 10–27. Springer Berlin Heidelberg, Berlin, Heidelberg (1997)
57. Guo, F.M., Liu, S., Mungall, F.S., Lin, X., Wang, Y.: Reweighted proximal pruning for large-scale language representation (2019), https://arxiv.org/abs/1909.12486
58. Haralick, R.M., Ripken, K.: An associative-categorical model of word meaning. Artificial Intelligence 6(1), 75–99 (1975)
59. Harris, Z.S.: Distributional structure. WORD 10(2-3), 146–162 (1954), https://doi.org/10.1080/00437956.1954.11659520
60. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention (2020), https://arxiv.org/abs/2006.03654
61. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015), https://arxiv.org/abs/1503.02531
62. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9, 1735–80 (12 1997)

63. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G.v.d., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J.W., Vinyals, O., Sifre, L.: Training compute-optimal large language models (2022), https://arxiv.org/abs/2203.15556
64. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021)
65. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization (2020), https://arxiv.org/abs/2003.11080
66. Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission (2019), https://arxiv.org/abs/1904.05342
67. Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, M.X., Chen, D., Lee, H., Ngiam, J., Le, Q.V., Wu, Y., Chen, Z.: Gpipe: Efficient training of giant neural networks using pipeline parallelism (2018), https://arxiv.org/abs/1811.06965
68. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging (2015), https://arxiv.org/abs/1508.01991
69. Ignat, O., Jin, Z., Abzaliev, A., Biester, L., Castro, S., Deng, N., Gao, X., Gunal, A., He, J., Kazemi, A., Khalifa, M., Koh, N., Lee, A., Liu, S., Min, D.J., Mori, S., Nwatu, J., Perez-Rosas, V., Shen, S., Wang, Z., Wu, W., Mihalcea, R.: A phd student's perspective on research in nlp in the era of very large language models (2023)
70. Iwasaki, Y., Yamashita, A., Konno, Y., Matsubayashi, K.: Japanese abstractive text summarization using bert. In: 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI). pp. 1–5 (2019)
71. Jia, C., Shi, Y., Yang, Q., Zhang, Y.: Entity enhanced BERT pre-training for Chinese NER. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6384–6396. Association for Computational Linguistics, Online (Nov 2020), https://aclanthology.org/2020.emnlp-main.518
72. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023)
73. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: Tinybert: Distilling bert for natural language understanding (2019), https://arxiv.org/abs/1909.10351
74. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data 7(3), 535–547 (2019)
75. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 655–665. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014), https://aclanthology.org/P14-1062
76. Khurana, D., Koli, A., Khatter, K., Singh, S.: Natural language processing: State of the art, current trends and challenges. Multimedia tools and applications pp. 1–32 (2022)
77. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751. Association for Computational Linguistics, Doha, Qatar (Oct 2014), https://aclanthology.org/D14-1181
78. Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. CoRR abs/1808.06226 (2018), http://arxiv.org/abs/1808.06226
79. Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.R., Stevens, K., Barhoum, A., Duc, N.M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A.,

Schuhmann, C., Nguyen, H., Mattick, A.: Openassistant conversations – democratizing large language model alignment (2023)
80. Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: Race: Large-scale reading comprehension dataset from examinations (2017), https://arxiv.org/abs/1704.04683
81. Lample, G., Conneau, A.: Cross-lingual language model pretraining (2019), https://arxiv.org/abs/1901.07291
82. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations (2019), https://arxiv.org/abs/1909.11942
83. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. CoRR abs/1901.08746 (2019), http://arxiv.org/abs/1901.08746
84. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks (2021)
85. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4582–4597. Association for Computational Linguistics, Online (Aug 2021), https://aclanthology.org/2021.acl-long.353
86. Liddy, E.D.: Natural language processing.(2001). Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc (2001)
87. Lin, C.C., Jaech, A., Li, X., Gormley, M.R., Eisner, J.: Limitations of autoregressive models and their alternatives. arXiv preprint arXiv:2010.11939 (2020)
88. Liu, W., Fu, X., Zhang, Y., Xiao, W.: Lexicon enhanced chinese sequence labeling using bert adapter. arXiv preprint arXiv:2105.07148 (2021)
89. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J.: Gpt understands, too (2021)
90. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019), http://arxiv.org/abs/1907.11692
91. Low, J.F., Fung, B.C., Iqbal, F., Huang, S.C.: Distinguishing between fake news and satire with transformers. Expert Systems with Applications 187, 115824 (2022)
92. Luo, H., Chuang, Y.S., Gong, Y., Zhang, T., Kim, Y., Wu, X., Fox, D., Meng, H., Glass, J.: Sail: Search-augmented instruction learning (2023)
93. Machado, F., Martes, C.D.: Project management success: A bibliometric analisys. Revista de Gestão e Projetos-GeP 6(1), 28–44 (2015)
94. Mishra, S., Arunkumar, A., Bryan, C., Baral, C.: A survey of parameters associated with the quality of benchmarks in nlp (2022)
95. Mutinda, F.W., Yada, S., Wakamiya, S., Aramaki, E.: Semantic textual similarity in japanese clinical domain texts using bert. Methods of Information in Medicine 60(S 01), e56–e64 (2021)
96. Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W.: Natural language processing: an introduction. J Am Med Inform Assoc 18(5), 544–551 (Sep 2011)
97. Narayanan, D., Harlap, A., Phanishayee, A., Seshadri, V., Devanur, N., Ganger, G., Gibbons, P., Zaharia, M.: Pipedream: generalized pipeline parallelism for dnn training. pp. 1–15 (10 2019)
98. Nasim, S.F., Ali, M.R., Kulsoom, U.: Artificial intelligence incidents & ethics a narrative review. International Journal of Technology, Innovation and Management (IJTIM) 2(2), 52–64 (2022)
99. OpenAI: Gpt-4 technical report (2023)

100. Osone, H., Lu, J.L., Ochiai, Y.: Buncho: Ai supported story co-creation via unsupervised multitask learning to increase writers' creativity in japanese. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. CHI EA '21, Association for Computing Machinery, New York, NY, USA (2021), https://doi.org/10.1145/3411763.3450391
101. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022), https://arxiv.org/abs/2203.02155
102. Park, J., Shin, H., Choi, H.J., Heo, J., Kim, W.H.: A rapid review of the use of appropriate technology in global health. Journal of Appropriate Technology 7(2), 235–247 (2021)
103. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS-W (2017)
104. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library
105. Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., Launay, J.: The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only (2023)
106. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014), https://aclanthology.org/D14-1162
107. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations (2018), https://arxiv.org/abs/1802.05365
108. Pollock, S.: A rule-based message filtering system. ACM Trans. Inf. Syst. 6(3), 232–254 (jul 1988), https://doi.org/10.1145/45945.214327
109. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pre-trained models for natural language processing: A survey. Science China Technological Sciences 63(10), 1872–1897 (2020)
110. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for squad (2018), https://arxiv.org/abs/1806.03822
111. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2019), http://arxiv.org/abs/1908.10084
112. Rodgers, W., Nguyen, T.: Advertising benefits from ethical artificial intelligence algorithmic purchase decision pathways. Journal of business ethics 178(4), 1043–1061 (2022)
113. De la Rosa, J., Fernández, A.: Zero-shot reading comprehension and reasoning for spanish with bertin gpt-j-6b (2022)
114. Saglam, M.: Key themes in brand reputation research: a bibliometric analysis with vosviewer software. Research Journal of Business and Management 9(1), 1–12 (2022)
115. Salminen, J., Mustak, M., Corporan, J., Jung, S.g., Jansen, B.J.: Detecting pain points from user-generated social media posts using machine learning. Journal of Interactive Marketing 57(3), 517–539 (2022)
116. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR abs/1910.01108 (2019), http://arxiv.org/abs/1910.01108
117. Santos, I., Nedjah, N., de Macedo Mourelle, L.: Sentiment analysis using convolutional neural network with fasttext embeddings. In: 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI). pp. 1–5 (2017)

118. Self, J.: Building expert systems, edited by f. hayes-roth d.a. waterman and d.b. lenat addison-wesley, london, 1983 (£27.50). Robotica 2(2), 119–119 (1984)
119. Sells, P., Shieber, S.M., Wasow, T.: Foundational issues in natural language processing: Introduction. Mit Press (1991)
120. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units (2015), https://arxiv.org/abs/1508.07909
121. Shah, Z.A., Khalidi, L.D.M.A.: Content analysis of abstracts of business administration and management sciences theses. Multicultural Education 7(12) (2021)
122. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE mobile computing and communications review 5(1), 3–55 (2001)
123. Shibayama, N., Shinnou, H.: Construction and evaluation of japanese sentence-bert models. In: Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation. pp. 249–256 (2021)
124. Shortliffe, E.: Mycin: A rule-based computer program for advising physicians regarding antimicrobial therapy selection pp. 739–739 (01 1976)
125. Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., Zhang, E., Child, R., Aminabadi, R.Y., Bernauer, J., Song, X., Shoeybi, M., He, Y., Houston, M., Tiwary, S., Catanzaro, B.: Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model (2022), https://arxiv.org/abs/2201.11990
126. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1631–1642. Association for Computational Linguistics, Seattle, Washington, USA (Oct 2013), https://aclanthology.org/D13-1170
127. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: Mpnet: Masked and permuted pre-training for language understanding (2020)
128. Song, P., Geng, C., Li, Z.: Research on text classification based on convolutional neural network. In: 2019 International Conference on Computer Network, Electronic and Automation (ICCNEA). pp. 229–232 (2019)
129. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A.W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A.S., Andreassen, A., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B.R., Loe, B.S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B.Y., Howald, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ramírez, C.F., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C.D., Potts, C., Ramirez, C., Rivera, C.E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, D., Khashabi, D., Levy, D., González, D.M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D.C., Yang, D., Lee, D.H., Shutova, E., Cubuk, E.D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodola, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E.A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E.E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G.I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G., Jaimovitch-López, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H., Schütze, H., Yakura, H., Zhang, H., Wong,

H.M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J.F., Simon, J.B., Koppel, J., Zheng, J., Zou, J., Kocoń, J., Thompson, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J.U., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Jones, J., Tenenbaum, J.B., Rule, J.S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K.D., Gimpel, K., Omondi, K., Mathewson, K., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Colón, L.O., Metz, L., Şenel, L.K., Bosma, M., Sap, M., ter Hoeve, M., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Quintana, M.J.R., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M.L., Hagen, M., Schubert, M., Baitemirova, M.O., Arnaud, M., McElrath, M., Yee, M.A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M.V., Peng, N., Chi, N., Lee, N., Krakover, N.G.A., Cameron, N., Roberts, N., Doiron, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N.S., Iyer, N.S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P.A.M., Doshi, P., Fung, P., Liang, P.P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P., Eckersley, P., Htut, P.M., Hwang, P., Milkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R.E., Gabriel, R., Habacker, R., Delgado, R.R., Milliere, R., Garg, R., Barnes, R., Saurous, R.A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., LeBras, R., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R., Lee, R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S.M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S.R., Schoenholz, S.S., Han, S., Kwatra, S., Rous, S.A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S.S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Shyamolima, Debnath, Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S.P., Lee, S.H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S.T., Shieber, S.M., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Telleen-Lawton, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V., Prabhu, V.U., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z.J., Wang, Z., Wu, Z.: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models (2022), https://arxiv.org/abs/2206.04615

130. Sun, S., Cheng, Y., Gan, Z., Liu, J.: Patient knowledge distillation for bert model compression (2019), https://arxiv.org/abs/1908.09355

131. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. p. 3104–3112. NIPS'14, MIT Press, Cambridge, MA, USA (2014)

132. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023)

133. Tarwani, K.M., Edem, S.: Survey on recurrent neural network in natural language processing. Int. J. Eng. Trends Technol 48(6), 301–304 (2017)

134. Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., Stojnic, R.: Galactica: A large language model for science (2022)
135. Team, M.N.: Introducing mpt-7b: A new standard for open-source, commercially usable llms (2023), www.mosaicml.com/blog/mpt-7b, accessed: 2023-05-05
136. Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H.S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhao, V., Zhou, Y., Chang, C.C., Krivokon, I., Rusch, W., Pickett, M., Srinivasan, P., Man, L., Meier-Hellstern, K., Morris, M.R., Doshi, T., Santos, R.D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E., Le, Q.: Lamda: Language models for dialog applications (2022), https://arxiv.org/abs/2201.08239
137. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023)
138. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023)
139. Tunstall, L., Reimers, N., Jo, U.E.S., Bates, L., Korat, D., Wasserblat, M., Pereg, O.: Efficient few-shot learning without prompts (2022)
140. Turing, A.M.: Computing machinery and intelligence. Springer (2009)
141. Ueda, N., Kawahara, D., Kurohashi, S.: Bert-based cohesion analysis of japanese texts. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 1323–1333 (2020)
142. Van Eck, N.J., Waltman, L.: Citation-based clustering of publications using citnetexplorer and vosviewer. Scientometrics 111, 1053–1070 (2017)
143. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017), https://arxiv.org/abs/1706.03762
144. Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., Rellermeyer, J.S.: A survey on distributed machine learning. ACM Comput. Surv. 53(2) (mar 2020), https://doi.org/10.1145/3377454
145. Vunikili, R., Supriya, H., Marica, V.G., Farri, O.: Clinical ner using spanish bert embeddings. In: IberLEF@ SEPLN. pp. 505–511 (2020)
146. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding (2018), https://arxiv.org/abs/1804.07461
147. Wang, W., Gang, J.: Application of convolutional neural network in natural language processing. In: 2018 International Conference on Information Systems and Computer Aided Education (ICISCAE). pp. 64–70 (2018)
148. Wang, Y.A., Chen, Y.N.: What do position embeddings learn? an empirical study of pretrained language model positional encoding (2020)
149. Warstadt, A., Singh, A., Bowman, S.R.: Neural network acceptability judgments (2018), https://arxiv.org/abs/1805.12471

150. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W.: Emergent abilities of large language models (2022)
151. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023)
152. Weizenbaum, J.: Eliza—a computer program for the study of natural language communication between man and machine. Commun. ACM 9(1), 36–45 (jan 1966), https://doi.org/10.1145/365153.365168
153. Williams, A., Nangia, N., Bowman, S.: A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1112–1122. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), https://aclanthology.org/N18-1101
154. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-Art Natural Language Processing. pp. 38–45. Association for Computational Linguistics (10 2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6
155. Woods, W.A.: Progress in natural language understanding: An application to lunar geology. In: Proceedings of the June 4-8, 1973, National Computer Conference and Exposition. p. 441–450. AFIPS '73, Association for Computing Machinery, New York, NY, USA (1973), https://doi.org/10.1145/1499586.1499695
156. Workshop, B., :, Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., Tow, J., Rush, A.M., Biderman, S., Webson, A., Ammanamanchi, P.S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A.V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P.O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A.F., Alfassy, A., Rogers, A., Nitzav, A.K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D.I., Radev, D., Ponferrada, E.G., Levkovizh, E., Kim, E., Natan, E.B., De Toni, F., Dupont, G., Kruszewski, G., Pistilli, G., Elsahar, H., Benyamina, H., Tran, H., Yu, I., Abdulmumin, I., Johnson, I., Gonzalez-Dios, I., de la Rosa, J., Chim, J., Dodge, J., Zhu, J., Chang, J., Frohberg, J., Tobing, J., Bhattacharjee, J., Almubarak, K., Chen, K., Lo, K., Von Werra, L., Weber, L., Phan, L., allal, L.B., Tanguy, L., Dey, M., Muñoz, M.R., Masoud, M., Grandury, M., Šaško, M., Huang, M., Coavoux, M., Singh, M., Jiang, M.T.J., Vu, M.C., Jauhar, M.A., Ghaleb, M., Subramani, N., Kassner, N., Khamis, N., Nguyen, O., Espejel, O., de Gibert, O., Villegas, P., Henderson, P., Colombo, P., Amuok, P., Lhoest, Q., Harliman, R., Bommasani, R., López, R.L., Ribeiro, R., Osei, S., Pyysalo, S., Nagel, S., Bose, S., Muhammad, S.H., Sharma, S., Longpre, S., Nikpoor, S., Silberberg, S., Pai, S., Zink, S., Torrent, T.T., Schick, T., Thrush, T., Danchev, V., Nikoulina, V., Laippala, V., Lepercq, V., Prabhu, V., Alyafeai, Z., Talat, Z., Raja, A., Heinzerling, B., Si, C., Taşar, D.E., Salesky, E., Mielke, S.J., Lee, W.Y., Sharma, A., Santilli, A., Chaffin, A., Stiegler, A., Datta, D., Szczechla, E., Chhablani, G., Wang, H., Pandey, H., Strobelt, H., Fries, J.A., Rozen, J., Gao, L., Sutawika, L., Bari, M.S., Al-shaibani, M.S., Manica, M., Nayak, N., Teehan, R., Albanie, S., Shen, S., Ben-David, S., Bach, S.H., Kim, T., Bers, T., Fevry, T., Neeraj, T., Thakker, U., Raunak, V., Tang, X., Yong, Z.X., Sun, Z., Brody, S., Uri, Y., Tojarieh, H., Roberts, A., Chung, H.W., Tae, J., Phang, J., Press, O., Li, C., Narayanan, D., Bourfoune, H., Casper, J., Rasley, J., Ryabinin, M., Mishra, M., Zhang, M., Shoeybi, M., Peyrounette, M., Patry, N., Tazi, N., Sanseviero, O., von Platen, P., Cornette, P., Lavallée, P.F., Lacroix, R., Rajbhandari, S., Gandhi, S., Smith, S., Requena, S., Patil, S., Dettmers, T., Baruwa, A., Singh, A., Cheveleva, A., Ligozat, A.L., Subramonian, A., Névéol, A., Lovering, C., Garrette, D., Tunuguntla, D., Reiter, E., Taktasheva, E., Voloshina, E., Bogdanov, E.,

Winata, G.I., Schoelkopf, H., Kalo, J.C., Novikova, J., Forde, J.Z., Clive, J., Kasai, J., Kawamura, K., Hazan, L., Carpuat, M., Clinciu, M., Kim, N., Cheng, N., Serikov, O., Antverg, O., van der Wal, O., Zhang, R., Zhang, R., Gehrmann, S., Mirkin, S., Pais, S., Shavrina, T., Scialom, T., Yun, T., Limisiewicz, T., Rieser, V., Protasov, V., Mikhailov, V., Pruksachatkun, Y., Belinkov, Y., Bamberger, Z., Kasner, Z., Rueda, A., Pestana, A., Feizpour, A., Khan, A., Faranak, A., Santos, A., Hevia, A., Unldreaj, A., Aghagol, A., Abdollahi, A., Tammour, A., HajiHosseini, A., Behroozi, B., Ajibade, B., Saxena, B., Ferrandis, C.M., Contractor, D., Lansky, D., David, D., Kiela, D., Nguyen, D.A., Tan, E., Baylor, E., Ozoani, E., Mirza, F., Ononiwu, F., Rezanejad, H., Jones, H., Bhattacharya, I., Solaiman, I., Sedenko, I., Nejadgholi, I., Passmore, J., Seltzer, J., Sanz, J.B., Dutra, L., Samagaio, M., Elbadri, M., Mieskes, M., Gerchick, M., Akinlolu, M., McKenna, M., Qiu, M., Ghauri, M., Burynok, M., Abrar, N., Rajani, N., Elkott, N., Fahmy, N., Samuel, O., An, R., Kromann, R., Hao, R., Alizadeh, S., Shubber, S., Wang, S., Roy, S., Viguier, S., Le, T., Oyebade, T., Le, T., Yang, Y., Nguyen, Z., Kashyap, A.R., Palasciano, A., Callahan, A., Shukla, A., Miranda-Escalada, A., Singh, A., Beilharz, B., Wang, B., Brito, C., Zhou, C., Jain, C., Xu, C., Fourrier, C., Periñán, D.L., Molano, D., Yu, D., Manjavacas, E., Barth, F., Fuhrimann, F., Altay, G., Bayrak, G., Burns, G., Vrabec, H.U., Bello, I., Dash, I., Kang, J., Giorgi, J., Golde, J., Posada, J.D., Sivaraman, K.R., Bulchandani, L., Liu, L., Shinzato, L., de Bykhovetz, M.H., Takeuchi, M., Pàmies, M., Castillo, M.A., Nezhurina, M., Sänger, M., Samwald, M., Cullan, M., Weinberg, M., De Wolf, M., Mihaljcic, M., Liu, M., Freidank, M., Kang, M., Seelam, N., Dahlberg, N., Broad, N.M., Muellner, N., Fung, P., Haller, P., Chandrasekhar, R., Eisenberg, R., Martin, R., Canalli, R., Su, R., Su, R., Cahyawijaya, S., Garda, S., Deshmukh, S.S., Mishra, S., Kiblawi, S., Ott, S., Sang-aroonsiri, S., Kumar, S., Schweter, S., Bharati, S., Laud, T., Gigant, T., Kainuma, T., Kusa, W., Labrak, Y., Bajaj, Y.S., Venkatraman, Y., Xu, Y., Xu, Y., Xu, Y., Tan, Z., Xie, Z., Ye, Z., Bras, M., Belkada, Y., Wolf, T.: Bloom: A 176b-parameter open-access multilingual language model (2022), https://arxiv.org/abs/2211.05100

157. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR abs/1609.08144 (2016), http://arxiv.org/abs/1609.08144
158. Xu, C., Zhou, W., Ge, T., Wei, F., Zhou, M.: Bert-of-theseus: Compressing bert by progressive module replacing (2020), https://arxiv.org/abs/2002.02925
159. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer (2020), https://arxiv.org/abs/2010.11934
160. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. CoRR abs/1906.08237 (2019), http://arxiv.org/abs/1906.08237
161. Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., Tam, W.L., Ma, Z., Xue, Y., Zhai, J., Chen, W., Zhang, P., Dong, Y., Tang, J.: Glm-130b: An open bilingual pre-trained model (2022), https://arxiv.org/abs/2210.02414
162. Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., Zhao, T.: Adaptive budget allocation for parameter-efficient fine-tuning (2023)
163. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L.: Opt: Open pre-trained transformer language models (2022), https://arxiv.org/abs/2205.01068
164. Zhang, X., Jiang, Y., Shang, Y., Cheng, Z., Zhang, C., Fan, X., Xiao, Y., Long, B.: Dsgpt: Domain-specific generative pre-training of transformers for text generation in e-commerce

title and review summarization. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (jul 2021), https://doi.org/10.1145%2F3404835.3463037

165. Zhang, Y., Liang, R., Qi, Y., Fu, X., Zheng, Y.: Assessing graduate academic scholarship applications with a rule-based cloud system. In: Cheng, E.C.K., Wang, T., Schlippe, T., Beligiannis, G.N. (eds.) Artificial Intelligence in Education Technologies: New Development and Innovative Practices. pp. 102–110. Springer Nature Singapore, Singapore (2023)

166. Zhang, Y., Yang, Q.: An overview of multi-task learning. National Science Review 5(1), 30–43 (2018)

167. Zhao, J., Huang, F., Lv, J., Duan, Y., Qin, Z., Li, G., Tian, G.: Do RNN and LSTM have long memory? In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 11365–11375. PMLR (13–18 Jul 2020), https://proceedings.mlr.press/v119/zhao20c.html

168. Zhou, C., Sun, C., Liu, Z., Lau, F.C.M.: A c-lstm neural network for text classification (2015), https://arxiv.org/abs/1511.08630

169. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books (2015), https://arxiv.org/abs/1506.06724

170. Zhuo, T.Y., Huang, Y., Chen, C., Xing, Z.: Exploring ai ethics of chatgpt: A diagnostic analysis. arXiv preprint arXiv:2301.12867 (2023)

9. Appendix

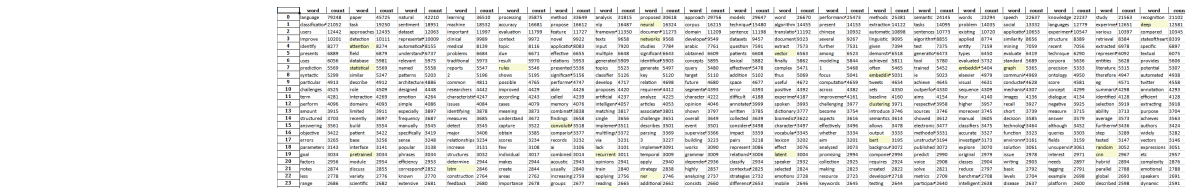


Fig. 14. Word count in abstracts



Fig. 15. Word cloud of abstracts

- BERT:

  [CLS] a token ##izer is a program that takes a given sentence or phrase and splits it into its component parts , or token ##s . these token ##s are typically words or sub ##words , such as individual characters or strings of characters . for example , " ducks " might be token ##ized into " duck " and " s . " token ##izer ##s also take into account plural forms of words , so that a sentence like " the boys are ducks " might be token ##ized into " the " , " boy " , " s " , " are " , and " duck " (with the " s " representing the plural form of " boy ") . token ##izer ##s are used in a variety of applications , such as text processing and natural language processing . [SEP] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]}

- BLOOM:

  <pad> <pad> <pad> ... A token izer is a program that takes a given sentence or phrase and splits it into its component parts , or tokens . These tokens are typically words or sub words , such as individual characters or strings of characters . For example , " duc ks " might be token ized into " duc k" and "s . " Token izers also take into account plural forms of words , so that a sentence like "The boys are duc ks " might be token ized into "The " , "b oy " , " s" , " are " , and " duc k" (with the " s" representing the plural form of "b oy "). Token izers are used in a variety of applications , such as text processing and natural language processing .

- GPT-NeoX:

  A token izer is a program that takes a given sentence or phrase and splits it into its component parts , or tokens . These tokens are typically words or sub words , such as individual characters or strings of characters . For example , " du cks " might be token ized into " du ck " and " s ." Token izers also take into account plural forms of words , so that a sentence like " The boys are ducks " might be token ized into " The ", " boy ", " s ", " are ", and " du ck " (with the " s " representing the plural form of " boy "). Token izers are used in a variety of applications , such as text processing and natural language processing . [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]

- LUKE:

  <s> A token izer is a program that takes a given sentence or phrase and splits it into its component parts , or tokens . These tokens are typically words or sub words , such as individual characters or strings of characters . For example , " du cks " might be token ized into " du ck " and " s ." Token izers also take into account plural forms of words , so that a sentence like " The boys are ducks " might be token ized into " The ", " boy ", " s ", " are ", and " du ck " (with the " s " representing the plural form of " boy "). Token izers are used in a variety of applications , such as text processing and natural language processing . </s> [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]

- XLNet:

  A token izer is a program that takes a given sentence or phrase and split s it into its component parts , or token s . These token s are typically words or sub word s , such as individual characters or strings of characters . For example , " d uck s " might be token ized into " d uck " and " s . " To ken izer s also take into account plural forms of words , so that a sentence like " The boys are duck s " might be token ized into " The " , " boy " , " s " , " are " , and " d uck " (with the " s " representing the plural form of " boy ") . To ken izer s are used in <sep> <cls>

9.1. Llama-2 experimenmts

The prompt used on LLama-2-70b to categorize the abstracts

```
You are a research scientist that specializes in Natural Language Processing. You are tasked with classifying
      different abstracts into one of 5 categories, depending on the newest technology or model the researchers used.
       The categories are:

0) The researchers did not explicitly mention a model or technique used, or it is not inmediately clear what they used
      .
1) Rule Based
2) Statistical
3) Neural Networks
4) Transformers
You are to use the highest category that fits the abstract (if the abstract fits in 2 and 3 you would choose 3). You
      can only choose one category.
The language used in these abstracts may be highly technical, so read every word very carefully.

Here is an example for category 3
{example abstract}

Here is an example for category 4
{example abstract}

Here is an example of category 2
{example abstract}

Here is an example for category 0
{example abstract}

Here is an example of category 1
{example abstract}

Here is an example of category 3
{example abstract}

The authors must be explicit about the model they are using, otherwise you should categorize with category 0

Here is how you should answer:
This abstract belongs to category []. Replace the [] with the category. Afterwards you should rationalize and recheck
      your answer. Also recheck if the researchers were really obvious about what they were using, or if you were
      making assumptions. finally, the last line should be what category you would put the abstract again, with the
      same format as before

Please categorize the next abstract
```

True category: 1. Llama-2 started with the correct category (1), and then miscategorized it as category 0. However, an argument could be made that Llama-2 was correct with this reasoning and the abstract was actually mislabeled.

```
This abstract belongs to category 1 (Rule-Based). The authors mention using HMM Toolkit (HTK) to construct the system,
       which suggests a rule-based approach to speech recognition. Additionally, they mention building a Mongolian
      acoustic model and collecting a large size corpus to construct the language model, which further supports a
      rule-based approach.

However, upon closer examination, it seems that the authors may have used a combination of techniques, including
      statistical modeling and machine learning, to develop their speech recognition system. Therefore, I would
      reevaluate my initial categorization and place this abstract in category 0 (The researchers did not explicitly
      mention a model or technique used), as it is unclear whether the approach used was purely rule-based or
      combined with other techniques.

Therefore, the final answer would be:
This abstract belongs to category 0.
```

True category: 3. Here Llama-2 appears to reach an incorrect conclusion based on the use of other techniques not related to the transformer architecture, classifying the abstract incorrectly.

```
Category: 3 (Neural Networks)

Rationalization: The abstract mentions the use of a neural network, specifically a reliability-based automatic repeat
      request (RB-ARQ), to improve the performance of a brain-computer interface (BCI) for communication.
      Additionally, the authors aim to combine RB-ARQ with error correction based on error-related potentials (ErrPs)
       to further enhance performance. The use of neural networks and error correction techniques suggests that the
      research falls under category 3.

Recheck: Upon rereading the abstract, I noticed that the authors actually used a combination of techniques, including
      signal averaging, RB-ARQ, and ErrP-based error correction. While the use of neural networks is still evident,
      the inclusion of other techniques means that the research may not exclusively fall under category 3. Therefore,
       I would revisit my initial answer and consider categorizing it as 4 (Transformers (BERT, GPT...)), as the
      authors seem to be combining different techniques to achieve their goals.

Final answer: After reevaluating the abstract, I believe it belongs to category 4
```

Santiago Canchila is Computer Science and Artificial Intelligence student with honorific recognition at Universidad Sergio Arboleda. He has imparted classes in Machine Learning and Blockchain Technologies. He has also built many projects and platforms with Artificial Intelligence as the main focus, with ample expertise in data processing and ML model building.

Carlos Eduardo Meneses is PhD student in Innovation and Business, Economist from the Nueva Granada Military University, Specialist in Urban and Regional Economics and Master's in Business Administration MBA from Prime Business School of the Sergio Arboleda University, candidate for a Master's in Urban and Regional Economics, PhD student in Innovation and Business and currently serves as Vice Dean of the School of Economics at the Sergio Arboleda University. Winner of the Sergio Arboleda Excellence Award in the Knowledge Management modality, passionate about the digital world, new economies and decentralized value systems, he has dedicated his research on issues related to the Collaborative Economy that leads with satisfaction to structuring and directing the International Collaborative Economy program of the Sergio Arboleda University in Colombia and Madrid Spain, as well as the opening of elective subjects such as Blockchain and Collaborative Economy at the same University and the creation of the new undergraduate program in Innovation and Economic Technology at the School of Economy. He currently teaches classes in the following fields: Professor of Business Economics Master of Business Administration, Professor of Economic Environment Analysis Specialization in Marketing Management, Professor of Strategic Analysis of Socioeconomic Environments (Bogotá, Barranquilla and Santa Marta), Collaborative Economy Teacher, Professor of Economics of Technological Platforms and Blockchain Economics teacher.

Javier Casanoves-Boix is lecturer at the Department of Marketing of the Faculty of Economics of from the Alfonso X El Sabio University (Madrid, Spain). His research attention has focused on the areas of Marketing and Branding. He has published papers in several refereed journals and editorial books. highlighting "Revista de Investigación Educativa", "International Visual Culture Review" and " UCJC Business & Society Review".

Pascual Cortés-Pellicer is an Associate Professor in the Department of Business Management of the Polytechnic University of Valencia (UPV), Spain. He has published several academic papers at internationally renowned journals and conference proceedings. His teaching and research interests include reverse logistics, sustainability, circular economy and supply chain management. In the 2021 year, He obtained the University Cel award to the best PhD thesis in the logistics field.

Fernando Castelló-Sirvent is an Economist and Associate Professor in the Department of Business Management at the Polytechnic University of Valencia (UPV), Spain. He holds a Ph.D. in Business Management from the University of Valencia. His research interests encompass artificial intelligence and its applications, sustainable transitions, strategy, crisis management, think tanks and public policy, entrepreneurship, and teaching innovation. He has extensive teaching experience in strategy, business information systems, international trade, international economics, and Spanish economics. For over two decades, he served as a senior strategy advisor to family businesses and has led numerous research and knowledge transfer projects.

Received: February 17, 2023; Accepted: January 06, 2024.

Computer Science and Information Systems 21(3):1147–1177 https://doi.org/10.2298/CSIS230302033M

MCClusteringSM: An approach for the Multicriteria Clustering Problem based on a Credibility Similarity Measure

Lugo Medrano Cesar[1] Gastelum Chavira Diego Alonso[2] Valdez Lafarga Octavio[3]
Velarde Cervantes Jose Luis[4]

[1] Doctoral Program in Business Analytics, Universidad Autonoma de Occidente
Blvd. Lola Beltran y Blvd Rolando Arjona, CP80020 Culiacan, Mexico
cesar.lugo@uadeo.mx
[2] Department of Economic-Administrative Sciences, Universidad Autonoma de Occidente
Blvd. Lola Beltran y Blvd Rolando Arjona, CP80020 Culiacan, Mexico
diego.chavira@uadeo.mx
[3] Department of Technology and Engineering, Universidad Autonoma de Occidente
Blvd. Lola Beltran y Blvd Rolando Arjona, CP80020 Culiacan, Mexico
octavio.valdezl@uadeo.mx
[4] Faculty of Computer Science, Universidad Autonoma de Sinaloa
Josefa Ortiz de Dominguez SN, Cd. Universitaria, CP80013 Culiacan, Mexico
jl.velarde17@info.uas.edu.mx

Abstract. Multicriteria clustering problem has been studied and applied scarcely. When a multicriteria clustering problem is tackled with an outranking approach, it is necessary to include preferences of decision makers on the raw dataset, e.g., weights and thresholds of the evaluation criteria. Then, it is necessary to conduct a process to obtain a comprehensive model of preferences represented in a fuzzy or crisp outranking relation. Subsequently, the model can be exploited to derive a multicriteria clustering. This work presents an exhaustive search approach using a credibility similarity measure to exploit a fuzzy outranking relation to derive a multicriteria clustering. The work includes two experimental designs to evaluate the performance of the algorithm. Results show that the proposed method has good performance exploiting fuzzy outranking relations to create the clusterings.

Keywords: MCDA, Multicriteria Clustering Problem, Similarity Measure.

1. Introduction

It is known today that one of the operational tools used to solve decision problems is Multicriteria Decision Analysis (MCDA). It is a field of Operations Research that addresses complex decision problems, some of them with multiple criteria in conflict. In MCDA, decision-makers can perform a choice, within a set of decision alternatives based on their preferences, by ranking or sorting them [35].In the choice problem, a restricted number of potential alternatives is chosen, as small as possible, which justifies eliminating the rest. Likewise, in the ranking problem, the decision maker can rank the alternatives from the best to the worst, with the possibility of ties and incompatibilities between them. Finally, in the sorting problem, each alternative is assigned to a set of pre-defined ordered

categories. However, there can be an additional problem called clustering, where grouping similar objects into homogeneous groups (clusters) is necessary. These clusters are unknown a priori, and the objective is to obtain them using the information contained in the elements to be clustered [1]. The clustering problem has been addressed in the statistics and data analytics contexts. Data analytics focuses on data processing to extract and highlight useful information that would otherwise be impossible to know. Two approaches to analyzing data for grouping objects have been developed: supervised classification and unsupervised classification. Olteanu [31] refers to the first approach as a supervised grouping process that relies on a priori information regarding the groups or classes to place the objects. For the second one, there is no benefit from any knowledge of the structure of the data. The purpose of clustering is to group objects based on the natural structure of the data using measures of similarity. In the context of MCDA, the clustering problem pertains to the need to group alternatives based on the similarity measures obtained from understanding the preferences of decision makers. In that sense, Boujelben [6] defined the multicriteria clustering problem as a combination of classical clustering with MCDA, where clusters of alternatives are obtained based on criteria whose clusters are undefined a priori.

Corrente et al. [12] affirm that classification/sorting and clustering is a fundamental topic in artificial intelligence (AI), with several machine learning (ML) algorithms that are available for classification and clustering tasks, and recently these problems have adopted a constructive decision aiding perspective in contrast to the statistical pattern recognition approach typically adopted in AI/ML.

The multicriteria clustering techniques allow understanding, in a better way, the structure of a problem [18]. These authors presented an illustrative example: during the exploratory phase of a problem characterized by a lot of alternatives, one could submit to the decision-maker the representative elements of the different clusters (instead of the whole dataset) to simplify the decision process. Multicriteria clustering methods can also be used to help decision-makers build categories in a sorting context.

Though the multicriteria clustering problem has been studied and applied in specific areas, there is still room for improvement in existing clustering algorithms. One limitation of some MCDA clustering methods is the use of a cutting level λ. This cutting level is arbitrarily set to exploit a fuzzy outranking relation to create the clusters. The resulting clusters can increase or decrease due to the cutting level value. The number of clusters increases due to the increment of the incomparability between alternatives when the cutting level is near one. The number of clusters decreases because the indifference between alternatives is increased when the cutting level is near zero. The proposed method in this work avoids this situation by using a credibility similarity measure to obtain the clustering.

On the other hand, literature regarding MCDA indicates that the clustering problem has not received considerable interest, as well as in the search for other forms of clustering since a wide range of decision-making problems in various areas of engineering and management such as logistics and transport, urban planning, environmental assessment, energy efficiency analysis, and financial decision-making, require the assignment of a set of decision options (alternatives) to groups [12]. Thus, this work proposes an approach to address instances of the multicriteria clustering problem without using an arbitrary cutting level. The method is an exhaustive search approach that exploits a fuzzy outranking relation to derive clusters of alternatives in the sense of a partition, i.e., each alternative

belongs to a unique cluster. The work is organized as follows: Section 2 provides a theoretical and conceptual framework of multicriteria clustering, including the foundation of outranking relations, the nature of a clustering problem and its complexity, a brief literature review, and the introduction of some important concepts about similarity measures. Section 3 presents the formulation of the multicriteria clustering problem and the proposed approach to address it. An empirical study to assess the proposed approach and a comparison with a multicriteria clustering procedure is described in Section 4. The results are presented in Section 5. Concluding general remarks and future perspectives are given in Section 6.

2. Theoretical and Conceptual Framework

2.1. Outranking Relations

In Multicriteria Decision Analysis, outranking relation theory (ORT) is a well-known area of knowledge. ORT has been widely used by MCDA researchers, mainly in Europe [17]. Bernard Roy set MCDA foundations during the late 1960s through the development of the ELECTRE (ELimination Et Choix Traduisant la REalité family of methods) [36].

An outranking relation is a binary relation that enables the decision maker to assess the strength of the outranking character of an alternative x over an alternative y. This strength increases if there are enough arguments to confirm that *"x is at least as good as y"* while there is no strong evidence to refute this statement.

Given two alternatives, x, and y, alternative x outranks y if *i*) there is a qualified majority of weighted criteria on which x is performing at least as good as y, and *ii*) there is no criterion on which y seriously outperforms x.

The outranking approaches operate in two major stages: construction and exploitation. The first involves modeling the decision-maker's preferences on criteria to construct an outranking relation among the considered alternatives. In contrast, the second stage involves the exploitation of the outranking relation to evaluating the alternatives for choice, ranking, or sorting purposes and, recently, for clustering. There are different procedures for constructing outranking relations, the most used being the ELECTRE and PROMETHEE [8] methods. Olteanu [31] also mentions the RUBIS method defined in a bipolar setting. ELECTRE I, for the choice problem, and ELECTRE III, for the ranking problem, are among the most popular outranking methods. In the construction stage, ELECTRE I builds a crisp outranking relation; meanwhile, ELECTRE III builds a fuzzy outranking [7]. Sometimes, a cutting level λ is used to transform fuzzy outranking relations into crisp ones. According to Linkov et al. [26], the cutting level is a technical parameter representing the sum of weights for the criteria that must be in concordance with the outranking relation to hold. A high cutting level means a high proportion of the criteria need to be the same or better than an alternative for the outranking relation to hold; a low cutting level means a lower proportion of the criteria are required for outranking relations to hold.

Defining the characteristics of fuzzy outranking relations and how to model decision-makers' preferences is beyond this work's scope. Still, readers can see Bouyssou [7], Fodor and Roubens [20], and Roy [37] for an explanation of them.

2.2. The Clustering Problem

There are different types of clustering, such as partition, hierarchy, pyramid, etc., the first two being the most frequent. This work is related to the partition type. In a partitional clustering, the universe of elements is divided into mutually exclusive subsets, where each element belongs to only one subset, so the pairwise intersection is empty. It is known it has an exponential complexity, where the number of partitions of a set is given by the Bell number [29]. The Stirling number of the second kind can be used to know the number of ways to divide a set of n elements within non-empty k clusters [32],

$$S(n,k) = \frac{1}{k!}\sum_{i=0}^{k} -1^{i}\binom{n}{k}(k-i)^{n} \tag{1}$$

For instance, given the set $A = \{1, 2, 3\}$, it can be partitioned into *i*) one cluster with the three elements: $C = \{1, 2, 3\}$, *ii*) two clusters in three ways to assign the elements: $C = \{\{1, 2\}, \{3\}\}, C = \{\{1, 3\}, \{2\}\}, C = \{\{2, 3\}, \{1\}\}$, and *iii*) three clusters in one way to assign the elements: $C = \{\{1\}, \{2\}, \{3\}\}$. Table 1 shows the different ways of grouping a set of three elements (n) with values of the number of clusters $k = 1$, $k = 2$, and $k = 3$.

Table 1. Different ways of grouping a set of three elements

Stirling number (n, k)	Different Ways	Groupings
$S(3, 1)$	1	$C = \{1, 2, 3\}$
$S(3, 2)$	3	$C = \{\{1, 2\}, \{3\}\}, C = \{\{1, 3\}, \{2\}\}, C = \{\{2, 3\}, \{1\}\}$
$S(3, 3)$	1	$C = \{\{1\}, \{2\}, \{3\}\}$

Notes: n = The number of elements, k = The number of clusters, and C= the clustering.
The first column corresponds to Stirling number with different values of n and k. The second column contains the number of ways to group a set of n elements in k clusters, obtained by the Stirling number. The third column shows the number of ways to set the three elements within non-empty k clusters.

The sum of all the possible clusters can be obtained with the Bell number formula:

$$Bell_n = \sum_{k=1}^{n} S(n,k) \tag{2}$$

The number of ways in which clusters can be constructed grows exponentially. It depends on the number of alternatives (*n*) and desired clusters (*k*). Thus, the clustering problem is non-trivial. It is shown in Table 2.

More ways to assign a set of elements in partitions are shown in Table 3, showing the complexity of the clustering problem itself.

2.3. A Literature Review of Multicriteria Clustering

Multicriteria clustering can be distinguished from classical clustering because it is a preference similarity-oriented problem where clusters should be conceived in preference

Table 2. Possible partitions for six elements within non-empty k clusters

Elements	Clusterings						Partitions
	1 cluster	2 clusters	3 clusters	4 clusters	5 clusters	6 clusters	
1	1	-	-	-	-	-	1
2	1	1	-	-	-	-	2
3	1	3	1	-	-	-	5
4	1	7	6	1	-	-	15
5	1	15	25	10	1	-	52
6	1	31	90	65	15	1	203

Notes: The Element column corresponds to the number of elements (n) in the group. The Clusterings column contains the Stirling number $S(n, k)$ in function to the number of elements n and the number of clusters k. The Partition column shows the total of possible clusters obtained with the Bell number formula.

Table 3. Number of partitions for up to 18 elements

n	partitions Bell (n)	n	Partitions Bell (n)
1	1	10	115975
2	2	11	678570
3	5	12	4213597
4	15	13	27644437
5	52	14	190899322
6	203	15	1382958545
7	877	16	10480142147
8	4140	17	82864869804
9	21147	18	682076806159

Source: Own elaboration based on [40].

proximity [19]. On this basis, general references about multicriteria clustering include mainly the next four aspects: *i*) the highlight methods in MCDA clustering, *ii*) how clusters are being evaluated in preference relations, *iii*) how the set preference relations are, and *iv*) if metaheuristics are used.

The next works considered all the previous four aspects: First, De Smet and Guzman [16] proposed an extension of the well-known k-means algorithm to the multicriteria framework. This extension relies on the definition of a multicriteria distance based on the preference structure defined by the decision maker. Then, De Smet and Eppe [15] developed a method that builds clusters and relations between these clusters based on a binary outranking matrix. An extension of the k-means algorithm is presented and tested on artificial data sets. Later, Baroudi and Safia [1] proposed a new clustering approach based on the definition of a new distance that considers the problem's multicriteria nature. This distance uses the preference relations of the PROMETHEE outranking method and the Sokal and Michener index so widely used in the classification field.

In this sense, Fernandez et al [19] proposed a clustering method based on a valued indifference relation inspired by outranking methods. They suggested a method based on comparing cluster centers and clusters' average net flow scores. Baroudi and Safia [2] tackle the problem of defining relations between clusters in multicriteria decision aid

clustering. Meyer and Olteanu [29] formally defined the problem of clustering in MCDA using notions that are native to this field alone and highlighted the different structures which it is tried to uncover through this process.

Baroudi and Safia [3] also Sarrazin et al [38] developed a model based on the FlowSort sorting procedure and the PROMETHEE I outranking method. Rosenfeld et al [34] presented a study to draw attention to clustering algorithms. They consider algorithms must be compared with performance indicators or criteria, leading to asymmetric preference relations.

Other contributions need to be considered too, do not include the four aspects completely, nevertheless add to the research in this field. For instance, Bisdorff [5] proposed an ELECTRE-like approach for clustering judges from their L-valued pairwise proximities in preference judgments. Cailloux et al [10] presented multicriteria clustering procedures to discover data structures from a multicriteria perspective. De Smet [14] addressed the ordered multicriteria clustering problem, detecting ordered categories. The k-means procedure and the underlying idea of the FLOWSORT method inspired the algorithm. Boujelben [6] worked on the multicriteria-ordered clustering problem too. He defined a preference profile to measure the preferential quality of the clusters and similarity and inconsistency profiles to analyze the clusters on each criterion. Chen et al. [11] proposed a total ordered clustering algorithm, which considers the preference degree between any two alternatives. Afterward, Liu et al. [27] introduced a multicriteria ordered clustering algorithm based on PROMETHEE and K-Medoids clustering algorithms. Daneshvar et al. [13] presented a multicriteria clustering method by combining k-means algorithm and PROMETHEE technique; the parameters of the problem are the cluster separator profiles which genetic algorithm (GA) is used to optimize them. Ishizaka et al. [22] proposed a new hierarchical multicriteria clustering based on PROMETHEE, where the number of clusters is not specified. On the other hand, Kandakoglu et al. [23] studied the project portfolio decision problem with uncertain multiple criteria evaluations, decision makers' preference information, and resource constraints. They proposed a new methodology based on multicriteria ordered clustering to deal with this problem. Valencia et al. [41] used clustering analysis with the multicriteria method TOPSIS [21] to group urban agricultural sites. TOPSIS was implemented directly into the clustering analysis process to generate weights for sustainability indexes before applying the k-means clustering algorithm. Later, Bashir et al. [4] motivated by the partial net outranking flow/profile of the PROMETHEE and the Fuzzy c-means clustering, presented a multicriteria-ordered profile clustering algorithm.

2.4. Similarity Measures

The concept of similarity is elemental to the clustering process. The level of similarity between two objects can be measured in many ways, which is firmly related to the nature of the elements themselves [31]. The term similarity can be related to a function used to compare elements of any type. Usually, its input is two elements, and the output is a value between 0 and 1. A value equal to zero means the elements are completely dissimilar. On the other hand, a value equal to one means the two elements are identical. The similarity is related to distance, where zero implies the two elements are identical, and one implies they are completely dissimilar, i.e., it is opposite to the similarity measure [43].

In that sense, a general way of measuring the similarity between two elements is by using proximity measures. These measures can be defined by a dissimilarity function *D*, also named a distance function, or using a similarity function S. A large value for *D* between two elements means that the objects are dissimilar. In contrast, a small value refers that they are similar. The opposite statements are made for the similarity function *S* [31].

Distance measures. Different clustering methods, e.g., k-means, use distance measures to determine the similarity or dissimilarity between any pair of elements. It is functional to indicate the distance between two instances x_i and x_j as $d(x_i, x_j)$. A valid distance measure should be symmetric and obtain a zero value in the case of identical vectors [33].

According to [1] and [42], usually, a similarity measure has the following properties:

$$\begin{aligned} \forall a_i, a_j \in A : d(a_i, a_j) \rightarrow \mathbb{R}^+ \text{with} \\ d(a_i, a_i) = 0, \text{identity}, \\ d(a_i, a_j) \geq 0, \text{non-negativity}, \\ d(a_i, a_j) = d(a_j, a_i), \text{symmetry}, \end{aligned}$$

also, when next properties are true

$$\begin{aligned} d(a_i, a_j) = 0 \Rightarrow a_i = a_j, \text{uniqueness}, \\ \forall a_i, a_j, a_z \in A : d(a_i, a_j) \leq d(a_i, a_z) + d(a_j, a_z), \end{aligned} \tag{3}$$

then, the similarity measure is called distance. Because not all similarity measures meet the symmetry or inequality of the triangle or both, not all similarity measures are distances.

[24] establish that the distance measure can be divided into two groups. In the first group, there are the metric measures, which must have the properties above. In the second group, there is the semi-metric, which does not follow the fifth property. These measures do not make possible an appropriate ordering of points over a metric space. For three points in this space, the sum of the distances from *i* to *j* and from *j* to *z* can be shorter than the distance between *i* and *z*.

Similarity functions. An alternative concept that can be used instead of distance is the similarity function $s(x_i, x_j)$ which compares the two vectors x_i and x_j. When these vectors are somehow "similar", the function has a large value and the largest value for identical vectors. This function should be symmetrical (namely $s(x_i, x_j) = s(x_j, x_i)$). Sometimes, methods for calculating the "distances" in the case of binary and nominal attributes can be considered as similarity functions rather than distances [33]

3. Formulation of a Multicriteria Clustering Problem and the Proposed Approach

3.1. Problem Formulation

Let $A = \{a_1, a_2, \ldots, a_m\}$ be a finite set of alternatives considered by a decision-maker, which is valued by a set of criteria $G = \{g_1, g_2, \ldots, g_n\}$, some of them in conflict with each other. Besides, let $S_A^\sigma \subseteq A \times A \rightarrow [0, 1]$ be a fuzzy outranking relation that integrates the preferences of that decision-maker on the multiple criteria that describe the elements of A.

The problem is to exploit S_A^σ to obtain a set of partitions P_A using a credibility similarity measure; where alternatives belonging to a specific cluster $C_i \in P_A$ are similar to each other; at the same time, they must not be similar to alternatives that belong to another cluster $C_j \in P_A$, so that P_A is a partition of A and the clusters $C_i \in P_A$ reflect the best compromise between the conflicting objectives, "discriminate the dissimilar alternatives," and "group the most similar alternatives". These partitions must be the most consistent possible to the information included in S_A^σ. Thus, this work aims to address the multicriteria clustering problem where a decision-maker's preferences are involved.

In the following subsections, an approach to partition the set of alternatives based on the information contained in the fuzzy outranking relation is presented. This method is called *MCClusteringSM - Multicriteria Clustering based on a Credibility Similarity Measure*. Specifically, the method belongs to the family of methods that exploit outranking relations, such as the exploitation phases of ELECTRE I, ELECTRE III, and ELECTRE TRI, to tackle the choice, ranking, and sorting problems, respectively, as described in [37]. The general scheme of the MCClusteringSM is presented in Figure 1.

3.2. The Clustering Method to Exploit a Fuzzy Outranking Relation

This part introduces the proposed method to exploit a fuzzy outranking relation to solving the multicriteria clustering problem. First, the general procedure to obtain the clusters is presented. Next, an illustrative example of this procedure is exposed. Finally, the general procedure of MCClusteringSM is described.

General procedure to obtain the clusters. Here, we provide in detail the steps to obtain the cluster. Additionally, an illustrative example is described.

The general steps to obtain the partitions by using MCClusteringSM are described below:

Step 1: Obtaining the Degrees of Credibility Similarity (DCS).
First, according to each credibility level that corresponds to the alternatives, a difference between each pairwise alternative is calculated over the intensity that outranks one alternative to another; this is the intensity which i outranks j is c_{ij} and the intensity with which j outranks i is c_{ji} , the subtraction needed is over these two values, considering the result in absolute value, this difference d can be noted as d_{ij} or d_{ji}:

$$d_{ij} = |c_{ij} - c_{ji}| \quad d_{ji} = |c_{ji} - c_{ij}| \quad d_{ij} = d_{ji} \tag{4}$$

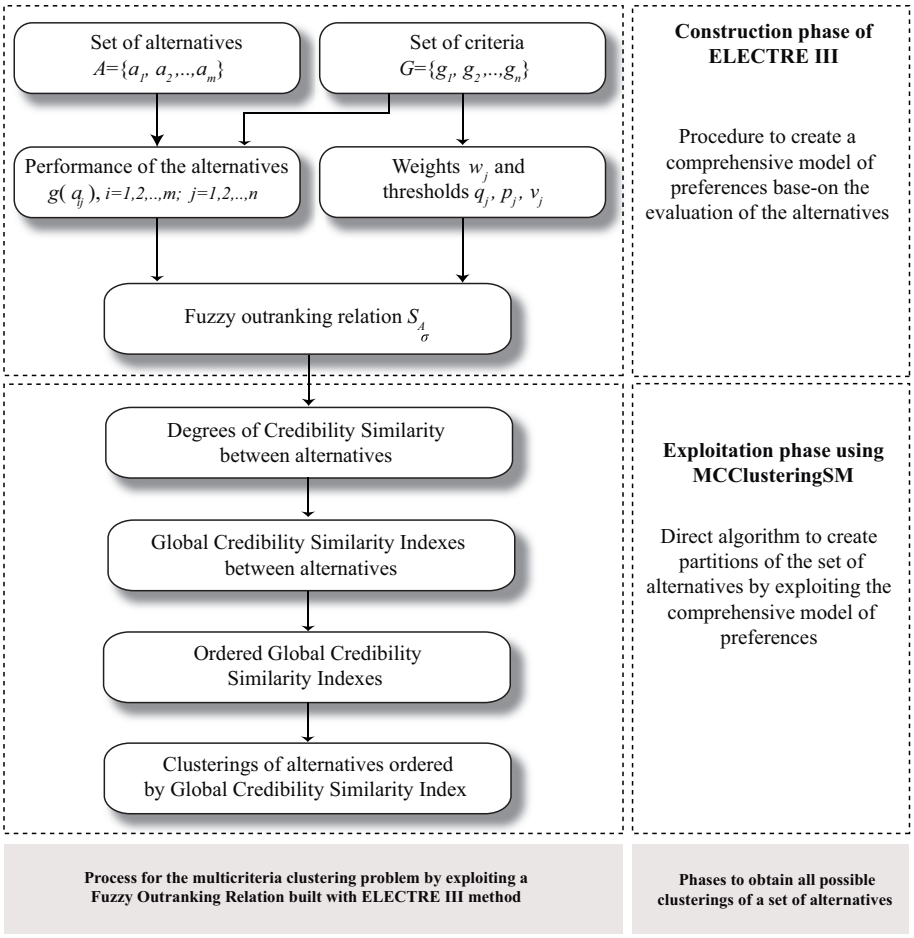


Fig. 1. The general approach for the multicriteria clustering problem Using ELECTRE III and MCClusteringSM

Source: Own elaboration based on [25]

Then, depending on the degree of credibility to calculate will be the differences to use in Equation 4; for a simple relation between two alternatives, the equation to calculate the degree of credibility similarity S is:

$$S_{ij} = 1 - \frac{d_{ij}}{c_{ij} + c_{ji}} \tag{5}$$

Equation 5 is the degree of credibility similarity of a cluster composed of two alternatives. Now, if a cluster of three alternatives or more can be formed, the equation would change in this way:

For the $M = \binom{N}{2}$ possible pairwise comparisons over the N alternatives of a subset n. Setting d_m as the mth level of credibility difference between each pairwise i, j of

alternatives, where $m = 1, 2, \ldots, M$, then, the DCS is:

$$S_n = 1 - \frac{\sum_{m=1}^{M} |c_{ijm} - c_{jim}| \in d_m}{\sum_{m=1}^{M} |c_{ijm} + c_{jim}| \in d_m} \tag{6}$$

For instance, the degree of credibility similarity of a cluster composed of the alternatives a, b, and c is:

$$S_{abc} = 1 - \frac{|c_{ab} - c_{ba}| + |c_{ac} - c_{ca}| + |c_{bc} - c_{cb}|}{|c_{ab} + c_{ba}| + |c_{ac} + c_{ca}| + |c_{bc} + c_{cb}|} \tag{7}$$

From Equation 6, a degree of one means the highest intensity of similarity between the alternatives, which suggests a similarity between the alternatives. However, a degree of zero means the lowest level of similarity, and this suggests that these alternatives should not be in the same cluster.

Some limitations have been found in this proposed similarity measure, which is because it does not fully follow the properties of a metric. In strict order, it should be called a semi-metric; one of these limitations is that it cannot be fully relied upon to represent dissimilarities in an Euclidean space without appropriate transformation [9]. Therefore, a transformation is proposed with good results for the research objectives. The new values of the transformed metric could still represent the data but will be more amenable to analysis or comparison. As an initial proposal, for this study, the metric was squared for it to represent dissimilarities in a Euclidean space better:

$$S_{ij} = 1 - \left[\frac{d_{ij}}{c_{ij} + c_{ji}}\right]^2 \tag{8}$$

This same power transformation is applied for three or more alternatives:

$$S_n = 1 - \left[\frac{\sum_{m=1}^{M} |c_{ijm} - c_{jim}| \in d_m}{\sum_{m=1}^{M} |c_{ijm} + c_{jim}| \in d_m}\right]^2 \tag{9}$$

Step 2: Compute the Global Credibility Similarity Index (GCSI).
Once the credibility similarity degrees are calculated, many scenarios can be presented depending on the number of clusters k required; the number of scenarios is equal to the number of ways to divide a set of alternatives within non-empty clusters. To obtain this, we can use Equation 1. Then we average the DCS of the partitions for each scenario to obtain a global credibility similarity index so that the best solution, including the clusters, will be the one with the highest index.

To set this, let's call W the number of ways to divide the set of alternatives A within non-empty clusters k, which is the Stirling number of the second kind. Then, we set n as the subset of alternatives that conforms to the k cluster in each scenario, where w is the wth way to divide the set of alternatives A within non-empty clusters k and $w = 1, 2, \ldots, W$, the GCSI is:

$$S_w = \sum_{n \in k} \frac{S_{nw}}{k} \tag{10}$$

Step 3: Ordering the Global Credibility Similarity Indexes.
To get the best clustering solution, according to the number of clusters required, ordering the global credibility similarity index is necessary to start the solution assignment from the best to the worst according to the index value.

An illustrative Example. Unlike classical methods, such as k-means, the alternatives are grouped based on the preference relations between alternatives given by the decision makers on the raw data set as thresholds and weights of the evaluation criteria. However, the cluster allocations are made directly from the data manifested in the model of preferences represented in a fuzzy outranking relation based on a credibility similarity measure calculated for all the possible combinations of grouped alternatives.

Let us now consider an illustrative example to see how the MCClusteringSM algorithm performs the calculation steps for the clustering process. The goal is to partition the set of alternatives $A = \{a, b, c, d\}$ into a given number of clusters in the best way.

Given the fuzzy outranking relation S_A^σ an input:

Fuzzy Outranking Relation

	a	**b**	**c**	**d**
a	1.00	0.90	0.70	1.00
b	0.80	1.00	0.70	0.80
c	0.30	0.30	1.00	0.30
d	0.00	0.30	0.30	1.00

Fig. 2. Fuzzy outranking relation of four alternatives
Source: Own elaboration

Step 1: Obtaining the Degrees of Credibility Similarity (DCS).

$$
\begin{aligned}
S_a &= 1 \\
S_b &= 1 \\
S_c &= 1 \\
S_d &= 1 \\
S_{ab} &= 1 - \left[\frac{|c_{ab} - c_{ba}|}{c_{ab} + c_{ba}}\right]^2 \\
&= 1 - \left[\frac{|0.9 - 0.8|}{0.9 + 0.8}\right]^2 \\
&= 0.9965 \\
S_{ac} &= 1 - \left[\frac{|0.7 - 0.3|}{0.7 + 0.3}\right]^2 \\
&= 0.84 \\
S_{ad} &= 1 - \left[\frac{|1 - 0|}{1 + 0}\right]^2 \\
&= 0 \\
S_{bc} &= 0.84 \\
S_{bd} &= 0.7933 \\
S_{cd} &= 1 \\
S_{abc} &= 1 - \left[\frac{|0.9 - 0.8| + |0.7 - 0.3| + |0.7 - 0.3|}{0.9 + 0.8 + 0.7 + 0.3 + 0.7 + 0.3}\right]^2 \\
&= 1 - (0.243)^2 \\
&= 0.9409 \\
S_{bcd} &= 0.8888 \\
S_{acd} &= 0.71 \\
S_{abd} &= 0.8227 \\
S_{abcd} &= 1 - \left[\frac{|0.9 - 0.8| + |0.7 - 0.3| + |1 - 0| + |0.7 - 0.3| + |0.8 - 0.3| + |0.3 - 0.3|}{0.9 + 0.8 + 0.7 + 0.3 + 1 + 0 + 0.7 + 0.3 + 0.8 + 0.3 + 0.3 + 0.3}\right]^2 \\
&= 1 - \left[\frac{2.4}{6.4}\right]^2 \\
&= 0.8594
\end{aligned}
$$

In an ordered way is:

$$
\begin{array}{lllll}
S_a = 1, & S_b = 1, & S_c = 1, & S_d = 1, & S_{cd} = 1, \\
S_{ab} = 0.9965, & S_{abc} = 0.9409, & S_{bcd} = 0.8888, & S_{abcd} = 0.8594, & S_{ac} = 0.84, \\
S_{bc} = 0.84, & S_{abd} = 0.8227, & S_{bd} = 0.7933, & S_{acd} = 0.71, & S_{ad} = 0
\end{array}
$$

Step 2: Compute the Global Credibility Similarity Index (GCSI).
Using Equation 10 to compute the Global Credibility Similarity Index, in this example, the possible numbers of clusters are from 2 to 3; using all the alternatives in one cluster or each alternative separated, this is $k = 2$ and $k = 3$.

For $k = 2$

$S_{cd_ab} = \frac{1+0.9965}{2} = 0.9985$

$S_{ac_bd} = \frac{0.84+0.7933}{2} = 0.81665$

$S_{bc_ad} = \frac{0.84+0}{2} = 0.42$

$S_{abc_d} = \frac{0.9409+1}{2} = 0.97045$

$S_{abd_c} = \frac{0.8227+1}{2} = 0.91135$

$S_{acd_b} = \frac{0.71+1}{2} = 0.855$

$S_{bcd_a} = \frac{0.8888+1}{2} = 0.9444$

For $k = 3$

$S_{a_b_cd} = \frac{1+1+1}{3} = 1$

$S_{a_c_bd} = \frac{1+1+0.7933}{3} = 0.9311$

$S_{a_d_bc} = \frac{1+1+0.84}{3} = 0.9466$

$S_{b_c_ad} = \frac{1+1+0}{3} = 0.66$

$S_{b_d_ac} = \frac{1+1+0.84}{3} = 0.9466$

$S_{c_d_ab} = \frac{1+1+0.9965}{3} = 0.9988$

Step 3: Ordering the Global Credibility Similarity Indexes.

For k=2, rank the clustering from highest to lowest:

1.$S_{cd_ab} = 0.99825$
2.$S_{abc_d} = 0.97045$
3.$S_{bcd_a} = 0.9444$
4.$S_{abd_c} = 0.91135$
5.$S_{acd_b} = 0.855$
6.$S_{ac_bd} = 0.81665$
7.$S_{bc_ad} = 0.42$

Thus, the best solution for$k$=2 is the cluster $C = \{\{c, d\}, \{a, b\}\}$.
For k=3, rank the clustering from highest to lowest:

1.$S_{a_b_cd} = 1$
2.$S_{c_d_ab} = 0.9988$
3.$S_{a_d_bc} = 0.9466$
4.$S_{b_d_ac} = 0.9466$
5.$S_{a_c_bd} = 0.9311$
6.$S_{b_c_ad} = 0.66$

The best solution for k=3 is the clustering $C = \{\{a\}, \{b\}, \{c, d\}\}$.

The General procedure of MCClusteringSM is presented in Algorithm 1. The input of MCClusteringSM is a fuzzy outranking relation, which will be partitioned according to a given number of clusters and alternatives. Each clustering obtained results from a partitional clustering process that considers preference relations based on the global credibility similarity index presented in Equation (10). MCClusteringSM was implemented in C language.

Algorithm 1: The MCClusteringSM procedure

Input : Fuzzy Outranking Relation $\boldsymbol{S_A^\sigma}$
Output: Partitional Clustering based on the credibility similarity index

1 Begin.
2 Set the number of clusters N_C and the number of alternatives N_A.
3 Compute the degrees of credibility similarity DCS between all possible combinations of the N_A alternatives.
4 Compute the Global Credibility Similarity Index $GCSI$ using DCS calculations from 1 to N_C.
5 Rank the $GCSI$ indexes in decreasing order to get all possible ordered clustering.
6 End.

4. Performance of MCClusteringSM

This section includes an empirical evaluation of MCClusteringSM and a comparison with a multicriteria clustering procedure.

This empirical evaluation aims to analyze how the direct method proposed performs when solving clustering problems with different structures and sizes. We intended to capture essential characteristics and analyze observations from the experiment execution. For this, with the help of an instance generator, we created a set of simulated reference sets, each of which is composed of classes of alternatives with different sizes and structures, and an associated fuzzy outranking relation that represents a model of preferences. This relation is constructed such that it can be utilized to generate cluster allocations of alternatives without inconsistencies according to its associated classes of alternatives. MCClusteringSM was executed to exploit each of the generated fuzzy outranking relations and evaluated if it successfully found the best solutions according to the fuzzy outranking relations supplied.

Besides, we attempted to evaluate the performance of MCClusteringSM compared to mccClusters [31]. It is a multicriteria clustering procedure implemented in the software Diviz [28]. This procedure computes clusters of alternatives based on the clustering typology: non-relational, relational, ordered, exclusive relational, and exclusive ordered multicriteria clustering. These evaluations were done to have a comparison point to analyze the performance of MCClusteringSM and to discuss its advantages and disadvantages.

4.1. Empirical evaluation of MCClusteringSM

Test criterion for evaluating the MCClusteringSM method. When a fuzzy outranking relation without cycles among alternatives is provided to a good clustering procedure, it should exploit this relation, preserving the cluster allocation of the alternatives according to preference relations. Thus, to evaluate the performance of MCClusteringSM, a fuzzy outranking relation without inconsistencies was provided to it. The best clustering solutions should be those that reflect the information contained in the fuzzy outranking relation. In MCClusteringSM, the best clustering solutions have Global Credibility Similarity Indexes of 1.0 or near 1.0, obtained with Equation 10.

Let us define $A = \{a_1, a_2, \ldots, a_m\}$ as a set of alternatives, $S_A^\sigma \subseteq A \times A \rightarrow [0, 1]$ as a fuzzy outranking relation, and C_A as an optimal clustered set of alternatives A, obtained with MCClusteringSM. Thus, according to this test criterion, the clustering of $(a_i, a_j) \in A \times A$ in S_A^σ should be reflected in C_A.

Experimental design. The experimental design for evaluating MCClusteringSM included a set of variables and a set of fuzzy outranking relations with different structures and sizes. This experimental design is described below.

The variables considered in this study were related to the number of clusters (V_1) and the number of alternatives (V_2). We selected small numbers for both variables due to the combinatorial nature of the problem. V_1 was defined with four values, and V_2 was defined with nine values. The first value of V_2 started with a value of six because of the maximum value of V_1, and it finished with a value of fourteen because MCClusteringSM is an exhaustive method. It last implies using high computing resources, in this case, a

large capacity of RAM (Random Access Memory). The variables of the experimental design are shown in Table 4.

Table 4. Variables of the experimental design

Variable	Value
V_1	2, 3, 4, 5
V_2	6, 7, 8, 9, 10, 11, 12, 13, 14

By assessing the performance of MCClusteringSM with the four values of V_1 and the nine values of V_2, we could achieve conclusions about the performance of our purpose in a multicriteria clustering context.

On the other hand, the general procedure to generate the fuzzy outranking relations of this test is presented in Algorithm 2. The output of it is a matrix that represents the

Algorithm 2: Instance generator

Input : Number of clusters N_C, Number of Alternatives N_A
Output: Fuzzy Outranking Relation $\boldsymbol{S_A^\sigma}$

1 Begin.
2 Form a vector V of alternatives of size N_A and randomly assign the N_C clusters to N_A alternatives, to ensure each cluster has at least an alternative.
3 Randomly complete V with the N_C clusters.
4 Create a matrix $\boldsymbol{S_A^\sigma}$ of size $N_A \times N_A$ and fill it with values $[0, 1]$ based on V, i.e., the clusters have the alternatives assigned on that basis.
5 End.

fuzzy outranking relation. The sizes of the matrix come from six to fourteen alternatives. Besides, a vector is randomly generated to indicate which alternative belongs to each cluster. Thus, the structure of the fuzzy outranking relation is based on such vector that represents a clustering. In this test, we wanted to know if the value of the clustering represented in the vector and implicitly in the fuzzy outranking relation is one of the best possible solutions, preferably, with a value of 1.0.

For each combination of the values of variables V_1 and V_2, Algorithm 2 was run 100 times. Thus, 3600 sets of references were created. This data generation was implemented in C language. The program was executed on an Asus computer with an Intel Core i5-10300H (2.5 GHz) processor, 16GB in RAM (DDR4 SDRAM), and a hard disk 250 GB 3G SAT 7.2K rpm LFF3. Figure 3 shows how a fuzzy outranking relation is generated with Algorithm 2.

Meanwhile, Algorithm 1 was implemented in C language and executed 3600 times. Instances of V_2 = 6, 7, 8, 10, 11, and 12 alternatives were executed in the Asus computer. Instances of V_2 = 13, 14 alternatives were run on a Lenovo server with processor Intel Xeon Gold 5218 (2.3 GHz) 8 cores, 128GB in RAM, and 764GB of hard disk.

In this experimental design, a response variable was defined to analyze the proposed method for the multicriteria clustering problem. It corresponds to the number of solu-

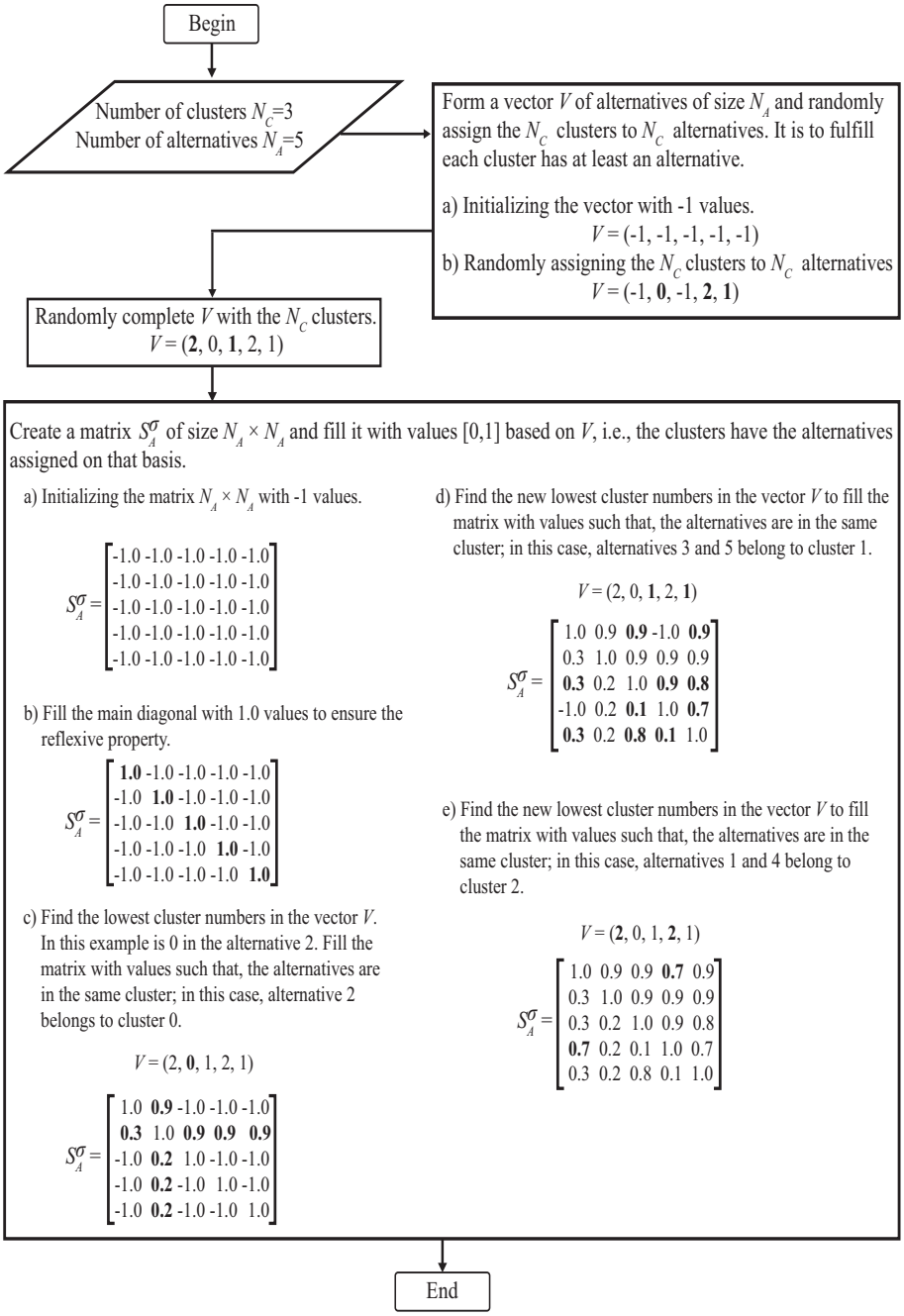


Fig. 3. A fuzzy outranking relation generated with Algorithm 2
Source: Own elaboration

tions without inconsistencies found per combination of variable 1 and variable 2 (V_1, V_2). Each combination (V_1, V_2) has 100 different fuzzy outranking relations of the 3600. For each run i-th of Algorithm 1 on a single fuzzy outranking relation, an auxiliary binary variable $y_i(V_1, V_2)$ is defined, which takes a value of one if Algorithm 1 finds the best solution without any inconsistency or zero otherwise. This last according to the source clustering generated in steps 3 and 4 of Algorithm 1, represented in the vector V. Thus, the response variable can be defined as $Y(V_1, V_2) = \Sigma_{i=1}^{100} y_i(V_1, V_2)$, which is the sum of all the 100 auxiliary binary variables per combination. The evaluation of the results of MCClusteringSM was done using an implementation of Algorithm 3 in C language.

Algorithm 3: Results evaluator

Input : Source clustering V, results of MCClusteringSM F
Output: Global Credibility Similarity Index of V in F

```
1  begin
2      set pos = 1
3      repeat
4          if V = F[pos] then
5              return Global Credibility Similarity Index of V ;
6          else
7              pos = pos + 1;
8          end
9      until the end of F is met;
10     return -1
11 end
```

4.2. Comparison of MCClusteringSM and Diviz mccClusters method

The study included a comparison with a multicriteria clustering method presented by Olteanu in [31]. This method is implemented in Diviz software [28] in the mccClusters component. This method was selected because it had good performance when exploiting fuzzy outranking relations to generate clusterings and by its availability. To use the mccClusters component, a workflow was designed in Diviz by adding the necessary components and parameters for the clustering generation.

Sample size and selection. Evaluating the mccClusters module with the 3,600 fuzzy outranking relations was impractical. We decided to use a sample whose size was calculated using the formula presented in [39] and defined as follows:

$$n = \frac{{Z_{\alpha/2}}^2 \cdot N \cdot p \cdot q}{E^2 \cdot (N-1) + {Z_{\alpha/2}}^2 \cdot p \cdot q} \tag{11}$$

where :

$$\begin{aligned} n &= \text{sample size,} \\ N &= \text{population size,} \\ Z_{\alpha/2} &= \text{critical value according to the confidence level selected,} \\ p &= \text{expected probability of the parameter to be evaluated,} \\ q &= 1 - p, \\ E &= \text{margin of error or imprecision allowed.} \end{aligned}$$

Thus, for a population of 3,600 fuzzy outranking relations, a 95% confidence level, a margin of error of 5.65%, and a probability p=0.50, the sample size was 277.14 fuzzy outranking relations:

$$n = \frac{1.96^2 \cdot 3600 \cdot 0.5 \cdot 0.5}{0.0565^2 \cdot (3600 - 1) + 1.96^2 \cdot 0.5 \cdot 0.5} = 277.14$$

Because the set of 3,600 fuzzy outranking relations was made up of 100 relations per each scenario of V_1 = 2, 3, 4, 5 clusters with V_2 = 6, 7, 8, 9, 10, 11, 12, 13, 14 alternatives; we decided to use a stratified sampling, composed of 36 strata, selecting randomly for practicality and convenience in each of them eight fuzzy outranking relations. Thus we selected 288 fuzzy outranking relations as a sample, mildly exceeding the minimum requirement established by the formula.

Clusters generation with Diviz mccClusters component. The mccClusters component computes clusters of alternatives based on the clustering typology: non-relational, relational, ordered, exclusive relational, and exclusive ordered multicriteria clustering [30]. In our case, the non-relational type was used because it groups a set of alternatives that are indifferent to each other, separating those that are not indifferent. In Data Analysis, each of these groups includes only one complementary relation (similarity and dissimilarity) [29]. Additionally, a fuzzy outranking relation is provided to the mccPreferenceRelation component. It creates a preference relation which is an input of the mccClusters component. It provides an option to define a cutting level from 0 to 1. If the cutting level is not used in an execution, this last is abruptly finished, and no result is produced. In this regard, Diviz issues an error message that this is probably due to a bug in the service. For this reason, the cutting level's default level (0.5) was used for this procedure.

Therefore, we exploited simulated fuzzy outranking relations to evaluate mccClusters, in the same way as MCClusteringSM, i.e., the success in finding the best solutions. The test included the same test criterion, variables, and response variable. For this test, a Diviz workflow was designed. It was composed of two file components for the alternatives and the fuzzy outranking relation, respectively. Also, it includes a mccPreferenceRelations component for the preference relations and a mccClusters for the final clustering. The Diviz workflow is shown in Figure 4.

The clusters generation with the mccClusters component was made considering a conversion. It was because the fuzzy outranking relations generated to test being constructed in plain text files. The conversion was made from plain text to the XMCDA format that Diviz uses. XMCDA is based on XML language for interoperability between different computer programs [28].

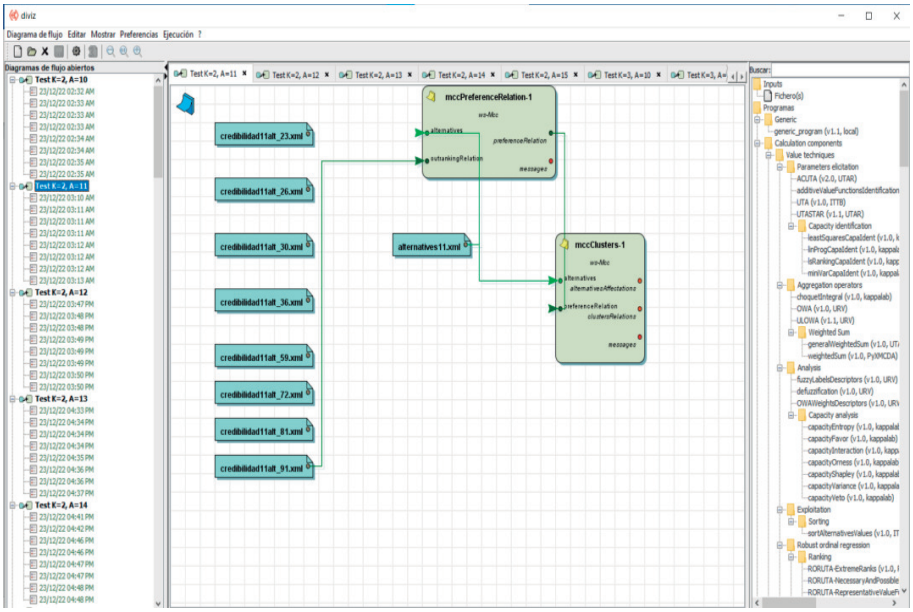


Fig. 4. Diviz workflow used for clusters generation with mccClusters component
Source: Own elaboration

Test criterion for evaluating the mccClusters component. In the same way, as in the test with MCClusteringSM, a solution without inconsistencies was considered just when the obtained clustering with mccClusters was the same as the reference clustering. This reference clustering is the vector V of alternatives of size N_A defined in step 5 of Algorithm 2. So, for our purpose, we can use the same logic for the response variable that in the previous test, considering that here the mccClusters module finds the best solution without inconsistency or zero otherwise. Then, the response variable can be defined as $Y(V_1, V_2) = \Sigma_{i=1}^{8} y_i(V_1, V_2)$, which is the sum of all the eight auxiliary binary variables per combination cause of the sampling used.

4.3. Empirical evaluation of MCClusteringSM considering a cutting level

In this section, we present another empirical evaluation to analyze the performance of MCClusteringSM. It included fuzzy outranking relations with different structures and sizes based on random cutting levels. Here, we attempted to capture the performance of MCClusteringSM with this new dataset and accomplish a comparison with the Diviz mccClusters method [30]. For this empirical evaluation, the fuzzy outranking relations were randomly generated so that there is a clustering of alternatives without inconsistencies for a given cutting level. The cutting level value was randomly obtained in a range of 0.5-0.75. The fuzzy outranking relations were created with a new version of the instance generator.

The test included the same test criterion, variables, values, and response variable as the empirical evaluation of section 4.2. However, this last differs in the number of replications. For each combination of the values of variables $V_1 = 2, 3, 4, 5$ clusters with

$V_2 = 6, 7, 8, 9, 10, 11, 12, 13, 14$ alternatives, the instance generator created 10 fuzzy outranking relations. Thus, 360 relations were created. This number is supported by the sample size calculated in subsection 4.2.1. The response variable was defined as $Y(V_1, V_2) = \Sigma_{i=1}^{10} y_i(V_1, V_2)$, which is the sum of all the ten binary variables per combination caused by the sampling used.

The MCClusteringSM was executed to exploit each of the generated fuzzy outranking relations. Then, we evaluated whether it successfully found the best solutions concerning the simulated fuzzy outranking relations. The valuations were done to get a reference point for the cutting level's inclusion in the input data generation. A solution without inconsistencies was considered just when the resulting clustering was the same as the one created by the instance generator.

Data generation using a cutting level $\lambda - cut$ The general procedure to generate the fuzzy outranking relations of this empirical evaluation is presented in Algorithm 4.

Algorithm 4: Instance generator considering a cutting level

Input : Number of clusters N_C, Number of Alternatives N_A
Output: Fuzzy Outranking Relation S_A^σ, vector V, and a cutting level $\lambda - cut$

1 Begin.
2 Form a vector V of alternatives of size N_A and randomly assign the N_C clusters to N_A alternatives, to ensure each cluster has at least an alternative.
3 Randomly complete V with the N_C clusters.
4 Randomly generate a cutting level $\lambda - cut$ $(0.5 \leqslant \lambda - cut \leqslant 0.75)$ with the N_C clusters.
5 Create a matrix S_A^σ of size $N_A \times N_A$ and fill it with values $[0, 1]$ based on V and $\lambda - cut$, i.e., the clusters have the alternatives assigned on that basis.
6 End.

The outputs of Algorithm 4 are a matrix, a vector, and a cutting level $\lambda - cut$. The first represents the fuzzy outranking relation according to the given cutting level $\lambda - cut$. The second one is an array indicating which alternative belongs to each cluster. The third one is a random cutting level between 0.5 and 0.75. The matrix and vector sizes depend on the number of alternatives (6-14), as in the previous test. The instance generator was implemented in C language and executed under the same conditions described for the previous test.

4.4. Comparison of MCClusteringSM and Diviz mccClusters method using a cutting level

As with MCClusteringSM, a cutting level was considered for the execution in the Diviz mccClusters component. It was done to compare the performance of both methods with this addition.

Like in the test with MCClusteringSM, a solution without inconsistencies was considered just when the resulting clustering with mccClusters was the same as the one created by the instance generator. For our purpose, the same test criterion, variables, and

values, and for the response variable, the same logic was applied as in the previous test for the Diviz mccClusters method. Still, the replication number was 10 instead of 8, considering what was done in section 4.3.1. Thus, the response variable was defined as $Y(V_1, V_2) = \Sigma_{i=1}^{10} y_i(V_1, V_2)$. This response variable is the sum of all the ten binary variables per combination caused by the sampling used.

Equally, as described in subsection 4.2.2., a fuzzy outranking relation is provided to the mccPreferenceRelation component, creating a preference relation. This relation is an input of the mccClusters component. Whereas in the last test with the Diviz mccClusters method, a cutting level was not used, we provided a cutting level generated by the instance generator for this new test. This cutting level needs to be supplied in the mccPreferenceRelation component shown in Figure 4.

5. Results

The carried-out experiments allowed us to portray the behavior of the proposed method. This section presents the analysis results of MCClusteringSM with the fuzzy outranking relations simulated to evaluate the inconsistencies regarding reference clusterings. Furthermore, a second analysis is presented between the multicriteria clustering method mccClusters with a sample of induced instances for comparison purposes with ours.

5.1. Results of MCClusteringSM vs. Induced Clusters

This experiment allowed us to observe the performance of the proposed method in a controlled environment. Specifically, according to the reference clustering, its accuracy in obtaining the best solutions for each simulated fuzzy outranking relation. Table 5 shows the results of our response variable against each possible combination of the independent variables, i.e., variables V_1: number of clusters and V_2: number of alternatives.

In all sets, MCClusteringSM found all the best solutions without inconsistencies in each one of the 100 fuzzy outranking relations evaluated per scenario. So, we can infer that the method was effective in 100% of the experiment scenarios.

For all the instances from six to fourteen alternatives, the credibility similarity metric obtained in the best solution was equal to one. A value of one means the highest intensity of similarity between the alternatives, which suggests a similarity between the alternatives. It also indicates the best solution with the highest intensity, indicating that the clusters were consistent with the fuzzy outranking relation provided.

Additionally, the results suggest that there is evidence to conclude that the size of the clusters and alternatives does not affect the performance of the similarity metric. However, the computational complexity of the method is a weakness in time and computational resources due to its combinatorial nature. Instances of V_2 = 6 to 12 alternatives, i.e., fuzzy outranking relations of size 6×6 to 12×12, required from 1GB up to 12GB of RAM. The required time in seconds to exploit each instance from 8 to 12 alternatives on the Asus computer was less than 0.5, 1, 2, 5, and 25, respectively. On the other hand, instances of V_2 = 13, 14 alternatives required around 26GB and 127GB of RAM, respectively. The required time for these cases on the Lenovo server was 47 seconds for the first one and 5 minutes for the second one.

Table 5. Results of MCClusteringSM vs. induced clusterings

Clusters	Alternatives	Solutions	Clusters	Alternatives	Solutions
	6	100		6	100
	7	100		7	100
	8	100		8	100
	9	100		9	100
2	10	100	4	10	100
	11	100		11	100
	12	100		12	100
	13	100		13	100
	14	100		14	100
	6	100		6	100
	7	100		7	100
	8	100		8	100
	9	100		9	100
3	10	100	5	10	100
	11	100		11	100
	12	100		12	100
	13	100		13	100
	14	100		14	100

Note: Solutions columns show the number of results without inconsistencies found per combination using MCClusteringSM.

5.2. Results of mccClusters module vs. Induced Clusters

The results of the mccClusters component executions, with the 288 fuzzy outranking relations, are presented in Table 6. As can be observed, this mccClusters only generated nineteen outcomes without inconsistencies in this experiment.

The cited nineteen outcomes without inconsistencies correspond to multicriteria clusterings with 6, 7, 8, and 9 alternatives in the four numbers of clusters tested. 63.15% of these solutions correspond to clusterings with 4 and 5 clusters. Therefore, the best performance of mccClusters was for clusterings with 4 and 5 clusters. For clustering with 4 clusters, it obtained 62.5%, 12.5%, 0%, 0%, 0%, 0%, 0%, 0%, and 0% of solutions without inconsistencies for 6, 7, 8, 9, 10, 11, 12, 13, and 14 alternatives respectively. For clustering with 5 clusters, the results were 25%, 37.5%, 12.5%, 0%, 0%, 0%, 0%, 0%, and 0% of solutions without inconsistencies for 6, 7, 8, 9, 10, 11, 12, 13 and 14 alternatives respectively. On the other hand, the worst performance of mccClusters was for clustering with 2 clusters. In this case, the results were 25%, 12.5%, 0%, 0%, 0%, 0%, 0%, 0%, and 0% of solutions without inconsistencies for 6, 7, 8, 9, 10, 11, 12, 13, and 14 alternatives respectively. Figure 5 illustrates these results.

In a global sense, the effectiveness of mccClusters was 6.59% over the 288 fuzzy outranking relations of the experiment. Unlike MCClusteringSM, the computational complexity in mccClusters method is not properly a weakness in time and computational resources due to the time duration to exploit each instance being less than 3 seconds, and Diviz software was easily executed on the Asus computer.

Table 6. Results of mccClusters module vs. Induced Clusters

Clusters	Alternatives	FOR [1] 1 2 3 4 5 6 7 8 [2]	Clusters	Alternatives	FOR 1 2 3 4 5 6 7 8
2	6	0 0 0 0 1 0 1 0	4	6	0 0 1 1 1 1 0 1
	7	0 0 0 0 0 0 0 1		7	0 0 0 0 1 0 0 0
	8	0 0 0 0 0 0 0 0		8	0 0 0 0 0 0 0 0
	9	0 0 0 0 0 0 0 0		9	0 0 0 0 0 0 0 0
	10	0 0 0 0 0 0 0 0		10	0 0 0 0 0 0 0 0
	11	0 0 0 0 0 0 0 0		11	0 0 0 0 0 0 0 0
	12	0 0 0 0 0 0 0 0		12	0 0 0 0 0 0 0 0
	13	0 0 0 0 0 0 0 0		13	0 0 0 0 0 0 0 0
	14	0 0 0 0 0 0 0 0		14	0 0 0 0 0 0 0 0
3	6	0 0 0 1 0 0 0 0	5	6	0 0 1 0 0 0 1 0
	7	0 0 0 0 1 0 0 0		7	0 0 0 0 1 1 0 1
	8	0 0 0 0 0 0 0 0		8	1 0 0 0 0 0 0 0
	9	0 0 0 0 1 1 0 0		9	0 0 0 0 0 0 0 0
	10	0 0 0 0 0 0 0 0		10	0 0 0 0 0 0 0 0
	11	0 0 0 0 0 0 0 0		11	0 0 0 0 0 0 0 0
	12	0 0 0 0 0 0 0 0		12	0 0 0 0 0 0 0 0
	13	0 0 0 0 0 0 0 0		13	0 0 0 0 0 0 0 0
	14	0 0 0 0 0 0 0 0		14	0 0 0 0 0 0 0 0

[1] Fuzzy Outranking Relation.

[2] In columns 1-8, a value of 1 means that the mccClusters component generated a solution without inconsistencies against the reference clustering. A value of 0 indicates the opposite.

5.3. Test Results of the Proposed Method vs. Induced Clusters considering a cutting level

This test allowed us to observe the behavior of the proposed method with the addition of a cutting level for constructing the fuzzy outranking relation. Particularly if clustering generated by the proposed method as the best solution corresponds to the induced clustering. Table 7 shows the results of our response variable defined in 4.3. against each possible combination of the independent variables, i.e., variables V_1: number of clusters and V_2: number of alternatives.

MCClusteringSM found 218 outcomes without inconsistencies and with the highest GCSI in each possible combination of the 360 fuzzy outranking relations evaluated. Thus, the method was strictly effective in 60.55% of the experiment scenarios.

About the mentioned 218 outcomes without inconsistencies, the distribution by cluster was 75, 65, 42, and 36 solutions for 2, 3, 4, and 5 clusters, respectively. For clustering with 2 clusters, MCClusteringSM found 83.33% of solutions without inconsistencies. The result of clustering with 3 clusters was 72.22% solutions without inconsistencies. For clustering with 4 clusters, it obtained 46.67% solutions without inconsistencies, and finally, for clustering with 5 clusters, the result was 40%. These results can be seen in detail corresponding to the solutions for each alternative value in Figure 6.

From the previous results, 342 solutions without inconsistencies given by MCClusteringSM were in the top 5 highest GCSI. Moreover, each one of these solutions was the

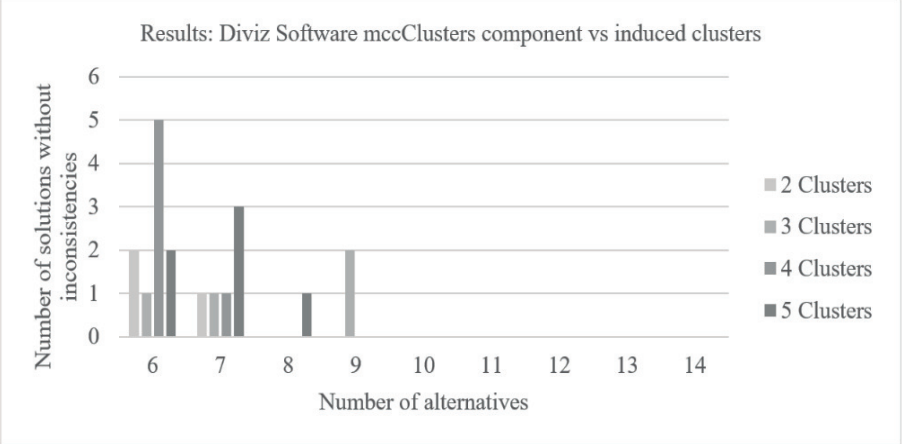


Fig. 5. Results of mccClusters for 288 fuzzy outranking relations
Source: Own elaboration

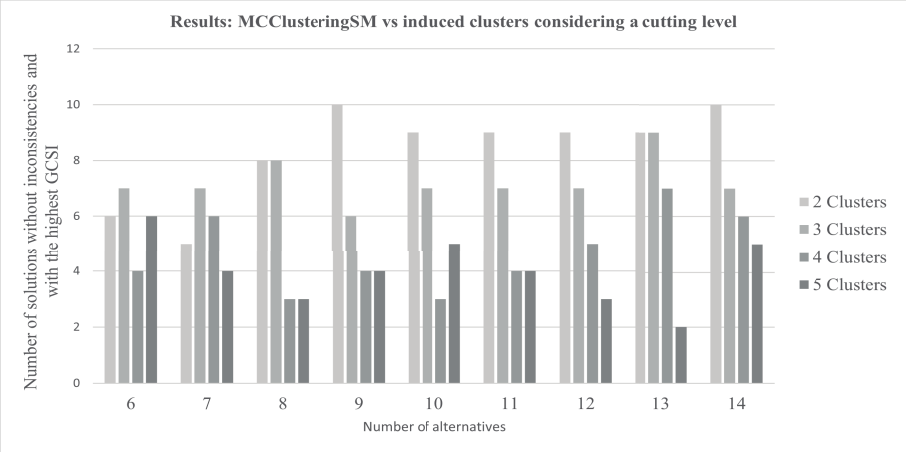


Fig. 6. Results of MCClusteringSM vs induced clusters for 360 fuzzy outranking relations
Source: Own elaboration

same as the reference clustering. Under this consideration, the method was effective in 95% of the scenarios, besides with values of credibility similarity metric from 0.988 to 1. Figure 7 illustrates these results.

Regarding the mentioned 342 outcomes, the distribution by cluster was 90, 84, 81, and 87 solutions for 2, 3, 4, and 5 clusters, respectively. In terms of percentage, it is 100%, 93.33%, 90%, and 96.66% of solutions without inconsistencies and within the top 5 highest GCSI for each cluster, respectively.

Table 7. MCClusteringSM vs. mccClusters using cut levels

Clusters	Alternatives	MCClusteringSM [1]										mccClusters [2]									
		1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
2	6	1	0	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1
	7	1	1	0	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
	8	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	10	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	11	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	12	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	13	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	6	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	7	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	8	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
	9	0	1	1	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1
	10	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1
	11	0	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1
	12	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	13	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	14	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1
4	6	0	1	1	0	0	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1
	7	1	0	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
	8	0	0	1	0	0	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1
	9	0	0	0	1	1	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1
	10	0	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
	11	1	1	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1
	12	1	1	1	0	0	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1
	13	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
	14	0	0	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1
5	6	0	1	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	7	1	0	0	0	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1
	8	0	0	0	0	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1
	9	1	0	1	0	0	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1
	10	1	1	0	0	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1
	11	1	0	1	0	0	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1
	12	1	0	0	0	0	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1
	13	0	0	0	1	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1
	14	0	1	1	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1

[1] In columns labeled from 1-10, a value of 1 means that MCClusteringSM generated a solution without inconsistencies against the reference clustering. A value of 0 indicates the opposite.

[2] In columns labeled from 1-10, a value of 1 means that mccClusters generated a solution without inconsistencies against the reference clustering. A value of 0 indicates the opposite.

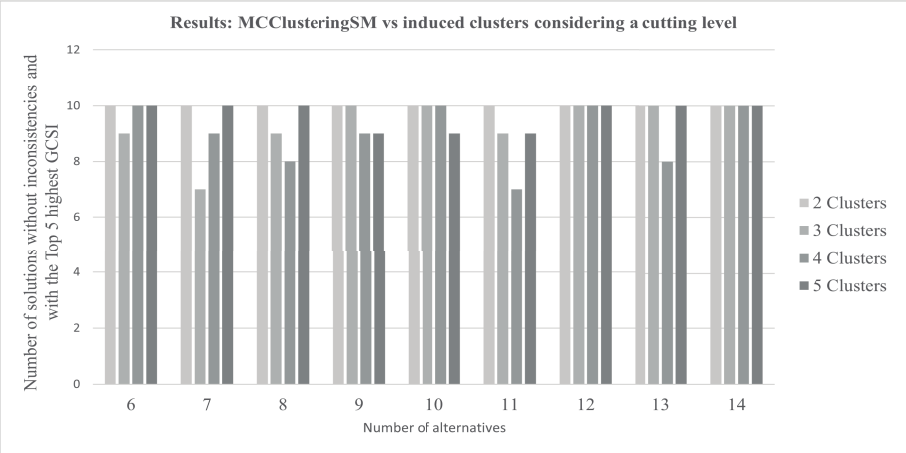


Fig. 7. Effectiveness of 95% of MCClusteringSM with 360 fuzzy outranking relations
Source: Own elaboration

5.4. Test mccClusters component results considering a cutting level

The results of the mccClusters component executions, with the 360 fuzzy outranking relations, are presented in Table 7. As can be observed, in all sets, this method found all the best solutions without inconsistencies in each of the 10 fuzzy outranking relations evaluated per scenario; namely, the clustering given by the Diviz mccClusters component as the best solution matches the induced clustering.

Furthermore, the results suggest that there is evidence to conclude that the size of the clusters and alternatives does not affect the performance of the mccClusters component. In a global sense, the effectiveness of mccClusters was 100% over the 360 fuzzy outranking relations of the experiment.

6. Concluding remarks and future perspectives

This paper addressed the problem of multicriteria clustering. We tackled this problem by using a method that includes a similarity measure to evaluate preferential information of the decision maker in a Multicriteria Decision Analysis context. The proposed method was designed to tackle the multicriteria clustering of other ones presented in the literature review of multicriteria clustering in section 2.3. There are several potential applications of multicriteria clustering, as many as there are in the choice, ranking, and sorting multicriteria problems. This problem has not received enough attention, but as methods are developed, this problem will be addressed more frequently.

Exploiting a preferences model to obtain clustering is a complex task. Nevertheless, based on the results, the proposed clustering procedure achieves good performances on the set of fuzzy outranking relations. Importantly, it outperforms the compared algorithm in finding the reference solutions. This approach can effectively exploit fuzzy outranking relations with up to 14 alternatives due to the computational complexity of deriving a

clustering. Due to MCClusteringSM being an exhaustive search method, it evaluates all possible forms of clustering and can obtain the best partitioning by generating clusters with the measure of similarity.

When the cutting level was used for the generation of the fuzzy outranking relations to be evaluated, outperformance with the compared algorithm was not achieved as in the first comparison made. However, 95% of the time, the reference solution was found in the top 5 best outcomes of our method, which indicates good effectiveness taking into consideration that MCClusteringSM does not use a cutting level for clustering construction whereas mccClusters component does. This $\lambda - cut$ value needs to be supplied and must be defined by the researcher from 0 to 1.

By using the similarity metric in the multicriteria clustering problem, an effort is being made to contribute to avoiding the cutting level λ, which sometimes is questionable and hard to assign it. In our approach, a metric is used directly to the input data, i.e., the credibility levels of the fuzzy outranking relation. It means that the clusterings are obtained using only the intrinsic information in the relation that already contains the decision maker's preferences. It avoids including a subjective decision when the cutting level is elicited to construct the clusters. The results indicate MCClusteringSM has good effectiveness even if it does not use a cutting level. On the other hand, mccClusters component also has good success but uses the subjective cutting level.

For future work, the proposed method should be improved regarding time and computational resources. Also, we consider that more validation tests are required using real and simulated datasets to highlight the efficiency of the MCClusteringSM algorithm. It will allow us to explore the method's limits by finding the maximum size within alternatives that can be solved with acceptable performance. Moreover, we will incorporate a cutting level in the simulated set generation to analyze any interesting characteristics from results against multicriteria clustering methods that use that cut level. We know the exhaustive characteristic of MCClusteringSM is its weakness because of the combinatorial nature of the problem. Thus, the proposed method will be extended using a metaheuristic based on evolutionary algorithms. Finally, applying the proposed method to a real problem is planned. The last would be interesting to value its behavior against other multicriteria clustering methods empirically.

References

1. Baroudi, R., Safia, N.B.: Towards multicriteria analysis: A new clustering approach. In: 2010 International Conference on Machine and Web Intelligence, ICMWI 2010 - Proceedings (2010)
2. Baroudi, R., Safia, N.B.: Towards the definition of relations between clusters in multicriteria decision aid clustering. In: Procedia Computer Science. vol. 17, pp. 134–140. Elsevier B.V. (2013)
3. Baroudi, R., Safia, N.B.: A multicriteria clustering approach based on similarity indices and clustering ensemble techniques. International Journal of Information Technology and Decision Making 13(4), 811–837 (2014)
4. Bashir, M.A., Muhiuddin, G., Rashid, T., Sardar, M.S.: Multicriteria Ordered the Profile Clustering Algorithm Based on PROMETHEE and Fuzzy c-Means. Mathematical Problems in Engineering 2023, 5268340 (2023), https://doi.org/10.1155/2023/5268340
5. Bisdorff, R.: Electre-like clustering from a pairwise fuzzy proximity index. European Journal of Operational Research 138(2), 320–331 (2002)

6. Boujelben, M.A.: A unicriterion analysis based on the promethee principles for multicriteria ordered clustering. Omega 69, 126–140 (2017), https://www.sciencedirect.com/science/article/pii/S0305048316305126
7. Bouyssou, D.: Outranking relations: Do they have special properties? Journal of Multi-Criteria Decision Analysis 5(2), 99–111 (1996)
8. Brans, J.P., Vincke, P., Vincke, P.H.: A Preference Ranking Organisation Method: (The PROMETHEE Method for Multiple Criteria Decision-Making) A PREFERENCE RANKING ORGANISATION METHODt (The PROMETHEE Method for Multiple Criteria Decision-Making). Tech. rep. (1985)
9. Buttigieg, P.L., Ramette, A.: A guide to statistical analysis in microbial ecology: A community-focused, living review of multivariate data analyses. FEMS Microbiology Ecology 90(3), 543–550 (2014)
10. Cailloux, O., Lamboray, C., Nemery, P.: A taxonomy of clustering procedures (01 2007)
11. Chen, L., Xu, Z., Wang, H., Liu, S.: An ordered clustering algorithm based on K-means and the PROMETHEE method. International Journal of Machine Learning and Cybernetics 9(6), 917–926 (2018)
12. Corrente, S., De Smet, Y., Doumpos, M., Greco, S., Zopounidis, C.: Classification, sorting and clustering methods based on multiple criteria: recent trends. Annals of Operations Research pp. 767–770 (2023)
13. Daneshvar, A., Homayounfar, M., Farahmandnejad, A.: Developing an Intelligent Multi Criteria Clustering Method Based on PROMETHEE. Journal of Industrial Management Perspective 9(Issue 4, Winter 2020), 41–61 (2020), https://jimp.sbu.ac.ir/article_87470.html https://jimp.sbu.ac.ir/article_87470_6bb05ceeda7f456d6992e185c421b7d9.pdf
14. De Smet, Y.: P2CLUST: An extension of PROMETHEE II for multicriteria ordered clustering. IEEE International Conference on Industrial Engineering and Engineering Management pp. 848–851 (2014)
15. De Smet, Y., Eppe, S.: Evolutionary Multi-Criterion Optimization. Evolutionary Multi-Criterion Optimization 5467(April), 380–392 (2009), http://www.springerlink.com/content/x79325217n618v74
16. De Smet, Y., Guzmán, L.M.: Towards multicriteria clustering: An extension of the k-means algorithm. European Journal of Operational Research 158(2), 390–398 (oct 2004)
17. Doumpos, M., Zopounidis, C.: Multicriteria Decision Aid Classification Methods. Kluwer Academic Publishiers (2004)
18. Doumpos, M., Figueira, J.R., Greco, S., Zopounidis, C. (eds.): New perspectives in multiple criteria decision making. Multiple Criteria Decision Making, Springer Nature, Cham, Switzerland, 1 edn. (2019)
19. Fernandez, E., Navarro, J., Bernal, S.: Handling multicriteria preferences in cluster analysis. European Journal of Operational Research 202(3), 819–827 (may 2010)
20. Fodor, J.C., Roubens, M.: Fuzzy preference modelling and multicriteria decision support. In: Theory and Decision Library A: (1994)
21. Hwang, Ching-Lai Yoon, K.: Multiple Attribute Decision Making. Lecture Notes in Economics and Mathematical Systems, Springer Berlin-Heidelberg, Berlin, Switzerland, 1 edn. (1981)
22. Ishizaka, A., Lokman, B., Tasiou, M.: A Stochastic Multi-criteria divisive hierarchical clustering algorithm. Omega (United Kingdom) 103, 102370 (2021), https://doi.org/10.1016/j.omega.2020.102370
23. Kandakoglu, M., Walther, G., Amor, S.B.: A robust multicriteria clustering methodology for portfolio decision analysis. Computers & Industrial Engineering 174, 108803 (2022), https://www.sciencedirect.com/science/article/pii/S0360835222007914
24. Legendre, L., Legendre, P.: Numerical ecology. Developments in Environmental Modelling, 3.– (01 1983)

25. Leyva López, J.C., Solano Noriega, J.J., Figueira, J.R., Liu, J., Gastélum Chavira, D.A.: Non-dominated sorting genetic-based algorithm for exploiting a large-sized fuzzy outranking relation. European Journal of Operational Research 293(2), 615–631 (2021), https://www.sciencedirect.com/science/article/pii/S0377221720310699
26. Linkov, I., Moberg, E., Trump, B.D., Yatsalo, B., Keisler, J.M.: Multi-Criteria Decision Analysis Case Studies in Engineering and the Environment. CRC Press, second edn. (2021)
27. Liu, X., Yu, H., Wang, G., Guo, L.: A multi-criteria ordered clustering algorithm based on PROMETHEE (2020)
28. Meyer, P., Bigaret, S.: Diviz: A software for modeling, processing and sharing algorithmic workflows in MCDA. Intelligent Decision Technologies 6(4), 283–296 (2012)
29. Meyer, P., Olteanu, A.L.: Formalizing and solving the problem of clustering in MCDA. European Journal of Operational Research 227(3), 494–502 (2013)
30. Olteanu, A.L.: Decision Deck XMCDA Web Services mccClusters (2013), https://www.decision-deck.org/ws/webservices_xmcda-v4/wsd-mccClusters-ws-Mcc.html#myPage
31. Olteanu, A.L.: On clustering in multiple criteria decision aid : theory and applications. Ph.D. thesis (06 2013)
32. Qi, F.: An Explicit Formula for the Bell Numbers in Terms of the Lah and Stirling Numbers. Mediterranean Journal of Mathematics 13(5), 2795–2800 (2016)
33. Rokach, L., Maimon, O.: Clustering Methods, pp. 321–352. Springer US, Boston, MA (2005), https://doi.org/10.1007/0-387-25465-X_15
34. Rosenfeld, J., Smet, Y.D., Debeir, O., Decaestecker, C.: Assessing partially ordered clustering in a multicriteria comparative context. Pattern Recognition 114, 107850 (2021), https://doi.org/10.1016/j.patcog.2021.107850
35. Roy, B.: Méthodologie multicritère d'aide à la décision. Politiques et Management Public (1986)
36. Roy, B.: Classement et choix en presence de points de vue multiples (La méthode Electre). Revue française d'informatique et de recherche opérationnelle (1968)
37. Roy, B.: The outranking approach and the foundations of electre methods. Theory and Decision (1991)
38. Sarrazin, R., De Smet, Y., Rosenfeld, J.: An extension of PROMETHEE to interval clustering. Omega (United Kingdom) 80, 12–21 (2018), https://doi.org/10.1016/j.omega.2017.09.001
39. Scheaffer, R.L.: Elementos de muestreo. Thomson Editores Spain Paraninfo, Madrid [etc, sixth ed. edn. (2006)
40. Sloane, N.: Bell or exponential numbers: number of ways to partition a set of n labeled elements. (2014), https://oeis.org/search?q=a000110&sort=&language=&go=Search.
41. Valencia, A., Qiu, J., Chang, N.B.: Integrating sustainability indicators and governance structures via clustering analysis and multicriteria decision making for an urban agriculture network. Ecological Indicators 142, 109237 (2022), https://www.sciencedirect.com/science/article/pii/S1470160X22007099
42. Veltkamp, R.C., Latecki, L.J.: Properties and performance of shape similarity measures. In: Batagelj, V., Bock, H.H., Ferligoj, A., Žiberna, A. (eds.) Data Science and Classification. pp. 47–56. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
43. Vlachos, M.: Encyclopedia of Machine Learning and Data Mining. In: Sammut, C., Webb, G. (eds.) Similarity Measures, vol. 3, pp. 1163–1666. Springer, second edn. (2017)

Cesar Lugo Medrano is a PhD student in Business Analytics from Universidad Autonoma de Occidente, Mexico, and holds an M.S. in Engineering with a specialty in Quality and Productivity Systems from Instituto Tecnologico y de Estudios Superiores

de Monterrey, Universidad Virtual. He holds a B. S. in Industrial and Systems Engineering from Instituto Tecnologico y de Estudios Superiores de Monterrey, Campus Sinaloa, Mexico, and is professor of the undergraduate degree at Tecnologico de Monterrey, Campus Sinaloa, and collaborates as a tutor-professor for postgraduate studies in the Online Education modality in the Master's Degree in Quality and Productivity Systems at Tecnologico de Monterrey.

Diego Alonso Gastelum Chavira is a full-time Economics and Management Sciences Department professor at Universidad Autonoma de Occidente, Mexico. He holds a PhD in Management Sciences from Universidad de Occidente, Mexico, an M.S. in Applied Informatics, and a B.S. in Informatics from Universidad Autonoma de Sinaloa, Mexico. He is a member of the National Researchers System of Mexico.

Octavio Valdez Lafarga is a full-time Engineering and Technology Department professor at Universidad Autonoma de Occidente, Mexico, and coordinator for the Business Analytics Master of Science program at the same university. He holds a PhD in Business Administration (concentration in Agribusiness) from Arizona State University and holds an M.S. in Agribusiness from Arizona State University. He holds a B. S. in Industrial and System Engineering from Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Sinaloa, Mexico.

Jose Luis Velarde Cervantes is a B. S. student in Informatics from Universidad Autonoma de Sinaloa, Mexico.

Received: March 02, 2023; Accepted: March 27, 2024.

Computer Science and Information Systems 21(3):1179–1201 https://doi.org/10.2298/CSIS230223032G

Analysis of the Countries' business attraction with the ELECTRE-III method

Tanya Samantha Garcia-Gastelum[1], Pavel Anselmo Álvarez[1], Ernesto León-Castro[2], and Cristhian Ramón Uzeta-Obregon[1]

[1] Department of Economic and Management Sciences, Universidad Autómoma de Occidente, Mexico
{tanya.garcia, pavel.alvarez, cristhian.uzeta}@uadeo.mx
[2] Faculty of Economics and Administrative Sciences, Universidad Católica de la Santísima Concepción. Concepción, Chile
eleon@ucsc.cl

Abstract. Attracting foreign investment is essential for the competitiveness and prosperity of nations. When deciding where to invest, an investor may be interested in considering specific criteria for investing or doing business and preferences for those criteria. In the same way, when evaluating the situation of a country to attract investment, consider different aspects to determine its ease of doing business. The MultiCriteria Decision-Making (MCDM) methodology is suitable for evaluating nations according to their ease of doing business due to the multifactorial elements of each nation. This work applied the ELECTRE-III method; it evaluated 190 nations based on the decision-maker's preferences, giving different importance to the ten criteria considered in the World Bank's Doing Business 2020 study. The results with this methodology show better-positioned nations in the ranking compared to the report presented by the World Bank Group.

Keywords: World Bank, Doing Business, business attraction, ELECTRE-III, multicriteria ranking, MCDM

1. Introduction

When developing a project to invest in a country or city, a fundamental criterion to be considered by decision-makers is the geographical location where it will be implemented. However, even considering its importance, there are other things to consider. The decision-makers tend to value additional aspects, such as security and legal elements related to the formalization of the investment, among others.

Similarly, the countries and their governments may be interested in attracting investment, requiring knowledge of the investors consider consistent with a competitive environment such as the current context. In this sense, both government institutions and investors have the possibility of analyzing information published by international organizations that evaluate nations in various areas, such as economic development, poverty rates, and inflation, as applied in Gryshova et al. [21]. Authors assess Ukraine's competitiveness based on information generated and published by the World Bank's Doing Business, the Corruption Perceptions Index, the Human Development Index, and the Global Competitiveness Ranking.

However, various factors influence investment attraction and decision-making in investment sites. Borissova et al. [6] mention that some factors are macroeconomic stability, trade policies, and even aspects of technology and financial security, among others. In this sense, multiple decision criteria are involved when choosing a place to invest, even making a distinction by city or region, including aspects related to direct and indirect competitors, target market, and subjective interests of decision makers.

In decision-making, obtaining all the information regarding the aspect is necessary to choose the best option. In this sense, various institutions carry out studies and generate information on evaluating nations and regions in different areas; these data are quickly available through digital portals and open databases. An international institution that carries out such measurements and evaluations through different programs is the World Bank Group, allowing using the available information, whether by business people, academics, researchers, or the governments themselves, for analysis processes and decisions, such as investigations on investment sites [26], correlational studies for the investment of small and medium-sized enterprises [14] and credit risk assessment [29], to name a few examples.

Regarding the evaluation of countries according to their ease of attracting and implementing investment projects, the World Bank Group periodically prepares reports on the ranking of nations according to their ease of investing through the Doing Business report [19], which evaluates 190 nations considering ten criteria related to the ease of doing business, publishing a ranking of the nations evaluated. The criteria considered in the World Bank study are the number and cost of the procedures necessary to set up a company, legal and professional records, and the minimum amounts to open bank accounts and start operations.

Other indicators evaluated by the World Bank are ease of processing construction permits, installation of electrical infrastructure, purchase and registration of properties, obtaining bank loans for foreign entities and tax payment. On the other side, it is essential to consider the government policies for protecting foreign and minority investors, import and export costs and times, laws regarding procedures and judgments for contract compliance, and the applicable legal framework for the alienation of assets due to debts.

In attracting foreign investment problems, it seems pertinent to approach it from a methodology that considers multiple decision criteria and the decision-maker's point of view. In this sense, the MultiCriteria Decision-Making (MCDM) methodology considers a wide range of methods with different approaches to address decision problems involving multiple criteria and the decision maker's preferences, resulting in solution proposals that match the view and value's system of the decision-maker.

The paper aims to answer the research question, how to rank the countries regarding different attributes to attract investment?. This research analyzes the countries' attractiveness for investment. The information published by the World Bank on its 2020 edition Doing Business microsite is analyzed. Here 190 countries are valued with 10 decision criteria established by the World Bank Group. The problem is approached from a multicriteria decision analysis (MCDA) methodology, applying the ELECTRE-III method. One of the main differences presented by the World Bank in its Doing Business report is that the proposed method allows the possibility of including the decision-maker's preference in the decision model to rank countries according to the country's situation to attract investment.

The main contributions of the paper are listed as follows:

- Analysis of the attractiveness of countries for investing with a multicriteria decision analysis methodology.
- Application of the ELECTRE-III method from the outranking approach to model the decision-maker preferences in the decision problem.
- The prioritization of countries with the best indicators to attract investment.

The paper is organized as follows. Section 2 presents a theoretical and conceptual description of studies of multicriteria decision-making methods with data generated by international institutions. Section 3 presents the methodology implemented to analyze the country's attraction. The discussion of the results is shown in Section 4. Finally, Section 5 presents the conclusions of the research.

2. Related Work

Various entities generate a large amount of data through surveys and censuses, the information is presented freely through repositories and digital data banks and free access, which can be used to carry out studies, calculations, estimates and make decisions in different areas, such is the case of the World Bank and Global Countries Competitiveness Rating.

In particular, the World Bank Group presents the microsite Doing Business, whose reports present information resulting from studies that they summarize, with the possibility of downloading and using them by the same countries referenced in the report as part of the decision-making process in relation to their government policies and mechanisms to attract foreign investment, as well as by companies or corporations interested in investing or establishing subsidiaries in other countries according to their expansion policy [20].

Using information such as that published by the World Bank can be considered one of the aspects that experts take to carry out a decision, such as choosing an investment site; However, other factors may not be included in the studies carried out by the institutions responsible for collecting information, but which have equal or greater weight for the experts since they have to do with their reality and the achievement of their objectives, such as the collection of funds and macroeconomic criteria or other aspects that can be considered as part of the decision maker's preferences. A decision problem can be approached from different approaches, applying tools that support decision-making, which offer the possibility of considering the decision maker's preferences, such as multicriteria decision-making methods (MCDM). In this sense, various authors have analyzed decision problems applying multicriteria methods; such is the case of Álvarez et al. [4], who analyze the innovation capacity in Mexico using the ELECTRE-III multicriteria method, considering many criteria involved and also the decision-maker's preferences, presenting as results a ranking of the 32 regions of Mexico according to their innovation capacity, valuable information for a decision process in this regard.

Continuing with the relevance of applying multicriteria methods in various decision problems, Álvarez et al. [1] and Muñoz-Palma et al. [25], evaluate the competitiveness in a country through a distinction by region, differentiating regions with a higher and lower competitiveness index. In both studies, ELECTRE-III was applied in a hierarchical version and through the conformation of composite indices, respectively. The results of these

studies provide the opportunity to analyze aspects of the competitiveness in Mexico, with the purpose of a possible application of decisions regarding public policies and attraction of investment capital in regions of the country.

The ELECTRE-III method has been applied in various decision problems, where specific characteristics are identified, such as heterogeneous measurement scales in the criteria to be evaluated and taking into account imperfect aspects of the construction of the same criteria, as well as qualitative information in their performance and their translation into quantitative information [16]. In the same way, this multicriteria method is used for group decisions in both governmental and private problems [3] [22], this being relevant for its use in problems such as the selection of research sites of investment, with information from institutions such as the World Bank Group, whose construction of the criteria to be evaluated tend to have characteristics such as those previously mentioned.

Regarding the studies that use information from the World Bank with multicriteria methodologies for decision-making, Soares-Silva et al. [29] present the credit risk assessment of bank bonds applying the Rough set method. This idea is also found in a study by De Lima Silva et al. [9] when considering the linguistic assessments converted into scales for the criteria, each scale being a discrete value for their weights.

In the same way, research has been presented with algorithms and models to address decision-making problems, such as generating proposals for places to do business, mainly for small and medium-sized companies, and analyzing the information presented by Doing Business, such as the one presented by Borissova. et al. [6], where group decision-making and weighted sum and simple additive weighting models are applied, the method used is similar to that used by the World Bank.

It should be noted that the studies presented by these authors [29][9][6], coincide with the present investigation when using multicriteria decision-making methods; however, a difference What is notable with our proposal is that the decision-maker's preferences are not taken into consideration, ignoring the possible needs of the decision-maker concerning the objectives to be achieved, or simply an adaptation of the model to be used by the regions to be analyzed in the problem, giving greater weight to specific criteria.

Analyzing the information presented by Doing Business with other multicriteria analysis methods allows obtaining new rankings under specific evaluation criteria, detecting weaknesses in the report published by the World Bank, considering that each dimension evaluated has the exact weighting and that possibly the indicators evaluated by Doing Business, do not reflect the particular reality of each nation [28]. The above is intended to be addressed in this investigation with a multicriteria methodology where the decision-maker's preferences are considered holistically, correcting one of the possible weaknesses detected in the World Bank ranking.

The use of digital databases makes it possible to implement decision-making methodologies proposed in studies and research projects, as well as to generate public policy proposals in different regions of the world, based on different decision-making problems [13] [14][18], simulating scenarios in which the expert is not willing to provide any information about their holistic preferences.

For this analysis, data from the World Bank, from the Doing Business [19] repository, was used to develop a ranking of the viability of nations to do business under two different techniques to those presented by the World Bank Group, using an ELECTRE-III multi-

criteria method, which has not been used to analyze this information, in order to compare the results and analyze the similarities and differences.

3. Methodology

The current research analyzes the information on Doing Business in its 2020 edition. Doing Business publishes information in five general aspects, divided into ten criteria that classify economies according to the ease of doing business [19]. In the present investigation, an experimental methodology was implemented, which consisted mainly of applying a multicriteria method for decision-making (MCDM), considering the preferences of the decision-maker, being a method different from that used by the World Bank in its Doing Business, where there is no preference distinction towards the relative importance of the criteria to carry out the resulting ordering.

The analysis of the valuation of countries is carried out by applying the ELECTRE-III, MCDM method. The methodological process to analyze countries' situation is depicted in Figure 1. It is carried out mainly in three stages. In Stage 1, the multicriteria decision-making problem is defined. Here, the decision criteria and alternatives are defined. The valuation of alternatives by each criterion corresponds to the performance matrix of the countries. The defined elements of Stages 1 and 2 correspond to the database from Doing Business [19].

Stage 2 corresponds to the method's parameters definition. The ELECTRE-III method requires the relative importance of criteria (w), indifference (q), preference (p), and veto (v) thresholds. Each of those parameters is defined for each criterion. The application of the MCDM method is carried out at this stage. The output is a ranking of alternatives.

Stage 3 corresponds to the analysis of the countries generated in the last stage. If the decision-maker does not agree with the result, she/he can go back to previous stages to modify her/his preferences until the resulting ranking matches her/his point of view.

Table 1 shows the decision criteria to analyze economies. Here, it is considered regulations that affect companies in all their stages, the creation of the company, obtaining a location, access to financing, management of daily operations, and operation in a secure business environment. The data concerns 190 economies around the world.

Table 1. Criteria for Evaluating the Ease of Doing Business

Criteria	Description
g1	Starting a business
g2	Dealing with construction permits
g3	Getting electricity
g4	Registering property
g5	Getting credit
g6	Protecting minority investors
g7	Paying taxes
g8	Trading across borders
g9	Enforcing contracts
g10	Resolving insolvency

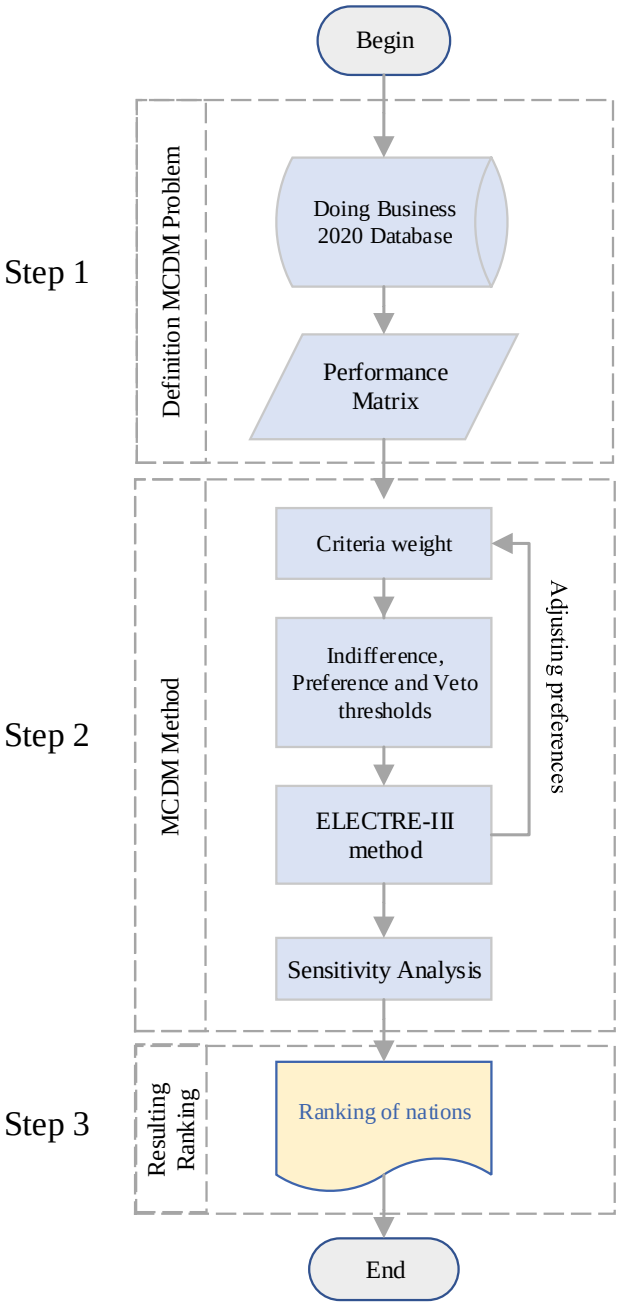


Fig. 1. Methodological process to analyze countries' situation

The ELECTRE-III method is applied to evaluate countries' performances regarding their ability to attract investment and compare each one against the others. The method is described as follows in the next section.

3.1. The ELECTRE III Method

ELECTRE III is an outranking method that compares each pair of alternatives in the $(a_i, a_l) \in A \times A$ set to assess the credibility of the assertion "action a_i is at least as good as action a_l", denoted as $a_i S a_l$. The evaluation of the assertion considers three aspects: a) the indifference and preference thresholds defined for each criterion, b) the degree or coefficients of importance attached to each criterion and c) the possible difficulties of relative comparison of two actions when one is significantly better than the other on a subset of criteria, but much worst on at least one criterion from a complementary subset.

The aggregation phase. In the aggregation phase the ELECTRE III (EIII) constructs a fuzzy outranking relation expressed as $\sigma(a_i, a_l)$ [27]. EIII constructs the comprehensive concordance index from the partial concordance index.

$C_j(a_i, a_l)$ is the partial concordance index, a fuzzy measure that evaluates if "action a_i is at least as good as action a_l" on the criterion g_j, using the DM's preference defined in the indifference q_j and preference p_j thresholds for the criterion g_j. The concordance index is one when the difference between criteria is very small, and it is cero when the difference is enough from the DM's point of view (see Equation 1).

$$C_j(a_i, a_l) = \begin{cases} 1, & \text{if } g_j(a_l) - g_j(a_i) \leq q_j, \\ \dfrac{p_j - [(g_j(a_l)) - (g_j(a_i))]}{p_j - q_j} & \text{if } q_j < g_j(a_l) - g_j(a_i) < p_j, \\ 0, & \text{if } g_j(a_l) - g_j(a_i) \geq p_j. \end{cases} \tag{1}$$

The resulted value of the partial concordance index is a fuzzy number that supports the asseveration "a_i is at least as good as a_l on criterion g_i". The values 0 and 1 are the case of no doubt about the asseveration. When the difference between alternatives increases, the asseveration loses credibility. The partial concordance index represents the credibility of the asseveration for each criterion as a fuzzy number (see Figure 2). However, no fuzzy operators are applied.

$C(a_i, a_l)$ is the comprehensive concordance index for each $(a_i, a_l) \in A \times A$, evaluating all criteria and assessing the assertion "a_i outranks a_l" (see Equation 2).

$$C(a_i, a_l) = \frac{\sum_{j=1}^{n} w_j \cdot C_j(a_i, a_l)}{\sum_{j=1}^{n} w_j} \tag{2}$$

In case of disagreement because the decision maker expresses veto values, some criterion calculates the index of discordance index $d_j(a_i, a_l)$. The discrepancy of criterion

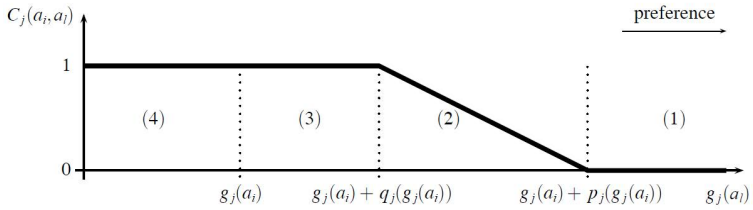


Fig. 2. Partial concordance index

g_j indicates the extent to which this criterion disagrees with the statement "a_i outranks a_l". The discordance index d_j is 0 when the difference between the two alternatives is very small for criterion g_j. As this difference increases, it is represented by a fuzzy value that increases its discordance until reaching the value of 1 when the difference between the two alternatives reaches the veto value defined for criterion g_j. Equation 3 shows the formulation of this discordance index, also represented in Figure 3.

$$d_j(a_i, a_l) = \begin{cases} 1, & \text{if } g_j(a_l) - g_j(a_i) \geq v_j, \\ \dfrac{[g_j(a_l) - g_j(a_i)] - p_j}{v_j - p_j} & \text{if } p_j < g_j(a_l) - g_j(a_i) < v_j, \\ 0, & \text{if } g_j(a_l) - g_j(a_i) \leq p_j. \end{cases} \tag{3}$$

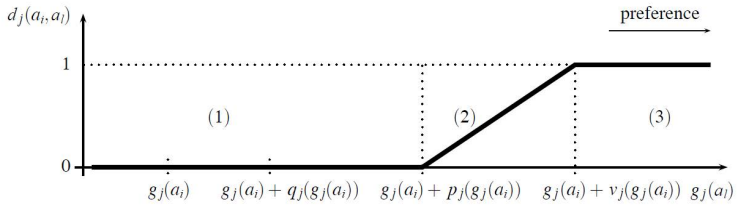


Fig. 3. Discordance index

The credibility index $\sigma(a_i, a_l)$; $(0 \leq \sigma(a_i, a_l) \leq 1)$ assesses the strength of the assertion that "a_i is at least as good as a_l", $a_i S a_l$.

$$\sigma(a_i, a_l) = \begin{cases} C(a_i, a_l), & \text{if } \overline{F}(a_i, a_l) = 0 \\ C(a_i, a_l) \times \displaystyle\prod_{j \in F(a_i, a_l)} \frac{1 - d_j(a_i, a_l)}{1 - C(a_i, a_l)} & \text{if } \overline{F}(a_i, a_l) \neq 0 \end{cases}, \tag{4}$$

where $d_j(a_i, a_l)$ is the discordant index and $F(a_i, a_l)$ is the set of pairs where $d_j(a_i, a_l) > C(a_i, a_l)$.

The exploitation phase: the distillation procedure. In the exploitation phase, the distillation procedure analyzes the fuzzy outranking relation. The distillation procedure measures the credibility of the asseveration of a_i outranks a_i with the value of $\sigma(a_i, a_l)$ and then presents a partial or complete preorder. It is generated from the descending and ascending distillation procedures, which show a complete order each.

The procedure finds $\sigma(a_i, a_l)$ corresponding with the highest value in the credibility matrix (5).

$$\lambda_0 = \max_{a_i, a_l \in A} \sigma(a_i, a_l) \tag{5}$$

Then the procedure estimates the distillation threshold function $s(\lambda_k)$ with the input parameters α and β.

$$s(\lambda_k) = \alpha\lambda_k + \beta, \tag{6}$$

where α and β are two thresholds that define the function $s(\lambda_k)$. It is a discrimination coefficient [10] for researching pairs where a_i is strictly preferred to a_l with a certain cut-off level. It is common to find $\alpha = -0.15$ and $\beta = 0.30$ as established values for the procedure.

We need to find the $\sigma(a_i, a_l)$ with the highest value in the next cut-off level $k + 1$. In each next level we shall have $\sigma(a_i, a_l) < \lambda_k - s(\lambda_k)$.

$$\lambda_1 = \max_{\sigma(a_i, a_l) < \lambda_k - s(\lambda_k)} \sigma(a_i, a_l) \tag{7}$$

In the process, the comparison of credibility degrees for pairs $\sigma(a_i, a_l)$ and $\sigma(a_l, a_i)$ is carried out. The $aS^{\lambda_1}b$ condition states a relation of power and/or weakness between alternatives. $aS^{\lambda_1}b$ if and only if $\sigma(a_i, a_l) > \lambda_1$ and $\sigma(a_i, a_l) > \sigma(a_l, a_i) + (\alpha \times \sigma(a_i, a_l) + \beta)$.

In each distillation, we shall find a reduced value of λ_k. It corresponds to a better condition to a_i is preferred than a_l for remaining subset of pairs.

Process of descending distillation

The first time the procedure begins, Step 1 and Step 2 need to performed in sequence. A iterative procedure identifies the minimum subset of alternatives that qualify for its placement in the descending complete order. The procedure finishes when all the alternatives are placed in the complete order.

Step 1: Calculate (5) as initial level to estimate the cut-off level (7) in order to identify the best remaining alternatives.

Step 2:

- **2.1** Find the highest degree of credibility with λ_k (5), estimate the next cut level λ_{k+1} (7) to find the maximum $\sigma(a_i, a_l)$ lower than $\lambda_k - s(\lambda_k)$. The subset of found alternatives with (7), are placed in the set D.
- **2.2** Calculation of power, weakness and qualification of alternatives from D. Every time a_i outranks a_l, the strength of a_i is incremented with 1 and the weakness of a_l is incremented with -1. For each alternative, the strengths and weaknesses are added together to give a final qualification score.
- **2.3** Select the alternatives with higher qualification, conforming the set D'.

The Ascending distillation performs a similar procedure in the opposite direction; the alternatives with the lowest qualification scores are assigned to the ranking's last position. The ranking is constructed from the bottom to the top.

Intersection between descending and ascending preorders

For the intersection between complete preorders, we need to find strict preferences in pairs of alternatives between preorders.

- a_i *is strictly preferred to* a_l if a_i is better positioned than a_l in at least one of the rankings, and if a_i is at least as good as a_l in the other rank.
- a_i *is indifferent to* a_l if a_i and a_l are placed in the same position (belong to same group) in the two rankings.
- a_i *is incomparable to* a_l if a_i is better positioned than a_l in one ranking and a_l is better positioned than a_i in the other ranking.

4. Results and Discussion

The Doing Business database, available on the official website of the World Bank Group, evaluates and reports 190 economies according to their ease of doing business through five general aspects, divided into ten criteria that consider all the regulations that affect investment for the implementation of a company, from its beginning, development, and end of operations [19].

The World Bank, through its Doing Business report, generates an ordering of the nations evaluated according to their global performance in the criteria determined by the same international organization, designed to evaluate their ease of investment and doing business. The ranking presented in the 2020 edition of Doing Business applies the simple average method, where it weighs all the criteria equally, giving the same weight to each of them [19].

On the other hand, in decision-making problems, such as deciding on a place to invest, the decision-maker's preferences become relevant to choosing the option to invest, turning problems of this type into decision problems with multiple criteria. In this sense, multicriteria decision-making methods (MCDM) are necessary to achieve the objectives defined for an investment site decision problem. The outranking MCDM presents heuristic approaches based on principles of agreement and disagreement, combining majority rules with respect to significant minorities and also handling uncertain and diffuse information [5].

The ELECTRE-III methodology is an MCDM outranking method, which includes the possibility of accepting qualitative as well as quantitative scales for the definition of the criteria, also allowing imperfect knowledge of the data to be modeled, an aspect that is considered by these methods with thresholds of preference and indifference required [5]. Figueira et al. [16] mention that it is appropriate to use the methods of the ELECTRE family if at least one criterion handles imperfect knowledge in its construction, which is not possible with other multicriteria methods.

Based on the above and considering the characteristics of the data that are analyzed in the present investigation, as well as the imperfect knowledge in the construction of the criteria and the definition of their importance for the expert, the ELECTRE-III outranking method was applied to determine the ranking of nations according to their ability to do business.

For applying the ELECTRE-III method, the indifference, preference, and veto thresholds were defined, as well as the weight of each of the ten criteria involved in evaluating nations. The weight of the criteria refers to the decision maker's preferences, defining whether one criterion is preferred over another.

In this research, the ELECTRE-III method was applied in two variants concerning the decision-maker's preferences, specifically regarding the relative importance of the criteria. In this sense, one instance considers the same importance for each criterion, similar to the Doing Business index. Another instance considers the different importance of the criteria regarding the expert's preference.

For this work, a local expert in finance and investment site selection was involved. The expert was assisted continually by the analyst during the definition process. The indifference (q) and preference (p) thresholds, as well as the weights (w) are the minimum parameters to be defined. This action corresponds to the elicitation process depicted in Figure 4, and it is carried out by interacting with the expert until reaching her agreement with the ranking solution. The following stages are derived from the elicitation process: definition of thresholds and weights; execution of the MCDM method, analysis of preliminary results; and adjusting the parameter if necessary.

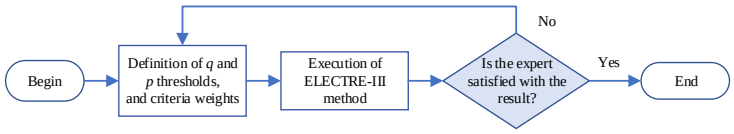


Fig. 4. Definition of parameters

The definition of the indifference (q), preference (p), and veto (v) thresholds meant a challenge for the participating expert. It was difficult to understand the meaning and express her viewpoint on this parameter. Applying the preliminary veto values resulted in solutions different from the expert's preferences. In this case, being the veto threshold (v) an optional parameter for the method used, the analyst and expert decided not to consider it in the processes presented below.

At first action, the analyst used the Simos' Revised Procedure [15]. It is a deck of cards procedure developed to support the definition of weight parameters. In this action, the expert order the criteria from the least important to the most important. The outcome is a set of weight values. The second action corresponds to explaining the meaning of indifference (q) and preference (p) thresholds. In the first instance, the expert proposed the first set of values for those parameters on each criterion. The definition by the participating expert resulted from an iterative process. Any change to the parameter values was evaluated by applying the ELECTRE-III method to generate the corresponding ranking, and the expert carefully analyzed each ranking. She worked in an interactive process, adjusting the parameter according to her ranking observations until their preferences were reflected in the solution obtained.

4.1. First instance: similar importance weights

In the first instance, the procedure simulates the point of view applied in the Doing Business index. Here the criteria are considered equally important. Table 2 list the parameters used in ELECTRE-III for the first instance. The weight value of each criterion g_i is set to 0.1. The indifference (q) and preference (p) threshold is set at 5 and 10, respectively.

Table 2. ELECTRE-III parameters regarding same importance of criteria

Parameter	g1	g2	g3	g4	g5	g6	g7	g8	g9	g10
w	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
q	5	5	5	5	5	5	5	5	5	5
p	10	10	10	10	10	10	10	10	10	10

Table 3 presents the ranking of the first ten positions considering an equal criteria weight with applying ELECTRE-III. The ten countries in the first instance ranking correspond to the same ten countries in the Doing Business index. The differences are in the position of some countries. For example, in the Doing Business report, New Zealand ranks higher than Singapore; however, the positions in our ranking are fifth and first, respectively. That is because when comparing these two nations, although the difference is minimal, Singapore is better evaluated in six of the ten criteria concerning New Zealand, which leads to deep reflection when analyzing the results when applying a different method than a simple average.

Table 3. Ten first position of the ranking regarding the same importance of criteria with ELECTRE-III

Rank	Code	Economy
1	SGP	Singapore
2	DNK	Denmark
3	KOR	Korea
4	HKG	Hong Kong
5	NZL	New Zealand
6	GBR	United Kingdom
7	NOR	Norway
8	US	United States
9	GEO	Georgia
10	SWE	Sweden

It should be noted that the results obtained with the ELECTRE-III method differ from the official ranking of the World Bank. Mainly, it is because the ELECTRE-III method is an outranking method based on a relational approach where the alternatives are compared between them, in this case the nations evaluated. The World Bank ranking uses an additive method related to the full aggregation approach in the multicriteria decision-making methodology [2]. Besides the above, the difference between results resides mainly in the

inclusion of the preference parameters necessary for applying the ELECTRE-III method (indifference, preference, and veto thresholds); that thresholds are directly involved in the final evaluation of an alternative, being determining factors in the result presented by the method, even if the criteria have the same weight (Table 2) because the performance of each nation in each criterion is compared with the values of the thresholds, having at the end a general performance of an alternative, which is compared concerning another, determining its position in the ordering.

Table 4 shows the performance in each criterion of the first five nations from Table 3. Being able to identify this information makes it possible to assess which criteria could be more relevant for a decision-maker when choosing a nation to set up a company or make an investment.

Table 4. Performance of the five first economies regarding the same importance of criteria

Economy / Criteria	g1	g2	g3	g4	g5	g6	g7	g8	g9	g10
Singapore (SGP)	98.2	87.9	91.8	83.1	75.0	86.0	91.6	89.6	84.5	74.3
Denmark (DNK)	92.7	87.9	90.2	89.9	70.0	72.0	91.1	100.0	73.9	85.1
Korea (KOR)	93.4	84.4	99.9	76.3	65.0	74.0	87.4	92.5	84.1	82.9
Hong Kong (HKG)	98.2	93.5	99.3	73.6	75.0	84.0	99.7	95.0	69.1	65.7
New Zealand (NZL)	100.0	86.5	84.0	94.6	100.0	86.0	91.0	84.6	71.5	69.5

Regarding the best-evaluated criteria for each of the nations presented in the ranking of Table 3, The first five economies positioned in this ranking are those with the best performance in the different criteria considered in the evaluation of nations (see Table 4).

The criteria performances of Singapore are one of the highest, the criteria Protecting minority investors (g6) and Enforcing contracts (g9) are the best values. It is the second best evaluated in the Starting a business (g1) criteria, Dealing with construction permits (g2), Getting credit (g5), and Paying taxes (g7). The Doing Business index ranks better New Zealand than Singapore. However, ELECTRE-III ranks better (first position of that ranking) Singapore than New Zealand (Fifth position) (see Table 3).

Continue with the five economy in ranking, Denmark ranks best in Trading across borders (g8) and Resolving insolvency (g10). Korea is the best evaluated for Getting electricity (g3) criteria. For its part, Hong Kong is best evaluated in the criteria of Dealing with construction permits (g2) and Paying taxes (g7). And, New Zealand performed better in the Starting a business (g1), Registering property (g4), Getting credit (g5), and Protecting minority investors (g6) criteria.

4.2. Second instance: different importance weights

In the second instance, a different point of view is considered regarding the relative importance of criteria. It is worthwhile it when a decision-maker is concerned with one or more criteria in the business investment. E. g. For the decision-maker, when cross-border trade (g8) is the most important criterion for making a decision, and the nations best evaluated in this criterion are not found in the top positions, their preferences are not considered in this decision model.

In this sense, considering the decision-maker's preferences, the second instance generates a ranking with different value weights of criteria. In this process, an expert participates in the definition of parameters. Table 5 list the parameters used in ELECTRE-III for the second instance. The most important criteria are Starting a business (g1), Dealing with construction permits (g2), and Getting electricity (g3). The weight value of each criterion g_i is set differently for the first three criteria. The rest is set to 0.07. The indifference (q) and preference (p) threshold is set at 5 and 10, respectively.

Table 5. ELECTRE-III parameters regarding different importance of criteria

Parameters	g1	g2	g3	g4	g5	g6	g7	g8	g9	g10
w	0.2	0.16	0.15	0.07	0.07	0.07	0.07	0.07	0.07	0.07
q	5	5	5	5	5	5	5	5	5	5
p	10	10	10	10	10	10	10	10	10	10

By defining different weights to specific criteria, this second result's ranking differs from the one previously presented in Table 3 and also differs from that presented in Doing Business 2020, with the same weighting for each of the ten criteria. The current ranking from Table 6 shows three new nations that did not appear in those first ten positions from the previous ranking: Lithuania, Taiwan, United Arab Emirates. When analyzing the first five positions of the new ranking, it is found that the first four positions remain the same countries in both rankings generated in this research with ELECTRE-III and in the one presented by the Doing Business.

Table 6. Ten first position of the ranking regarding different importance of criteria

Rank	Code	Economy
1	HKG	Hong Kong
2	SGP	Singapore
3	DNK	Denmark
4	KOR	Korea
5	GBR	United Kingdom
6	NZL	New Zealand
7	LTU	Lithuania
8	TWN	Taiwan
9	GEO	Georgia
10	ARE	United Arab Emirates

The performance of the first five places in the ranking makes it possible to observe that by giving greater weight to the g1, g2 and g3, the order of the nations is influenced by them (see Table 7). Hong Kong now appears in the first place, being the best evaluated in the Starting a business (g1) criterion, like Singapore, which is in the second position of the ranking, indicating that the g1 criterion is the one with the greatest weight according to the parameters defined.

Table 7. Performance of the five first economies regarding the different importance of criteria

Economy / Criteria	g1	g2	g3	g4	g5	g6	g7	g8	g9	g10
Hong Kong (HKG)	98.2	93.5	99.3	73.6	75.0	84.0	99.7	95.0	69.1	65.7
Singapore (SGP)	98.2	87.9	91.8	83.1	75.0	86.0	91.6	89.6	84.5	74.3
Denmark (DNK)	92.7	87.9	90.2	89.9	70.0	72.0	91.1	100.0	73.9	85.1
Korea (KOR)	93.4	84.4	99.9	76.3	65.0	74.0	87.4	92.5	84.1	82.9
United Kingdom (GBR)	94.6	80.3	96.9	75.7	75.0	84.0	86.2	93.8	68.7	80.3

As presented in Table 7, Hong Kong, in addition to being best evaluated in Starting a business (g1), is also the one with the highest score in the Dealing with construction permits (g2) and the second best evaluated in Getting electricity (g3); as well as in Getting credit (g5) and Paying taxes (g7) has the highest score.

In the case of the United Kingdom, which appears in fifth place, it is well evaluated in the Getting credit (g5). Its position is also determined by its performance in the g1, g2, and g3 criteria, which despite not having the best values, is the second or third in their performance.

It should be noted that even though each nation's performance was not modified, by applying a method that considers the preferences of the decision-maker through weight parameters of the criteria or the thresholds of indifference, preference, and veto, the result obtained may vary concerning that published by the World Bank, giving the possibility of choosing the most relevant criteria for a problem.

The decision-maker's preferences in problems of selecting places to do business can also be considered through the weighted sum method, as mentioned by Borissova et al. [6], whose methodology is similar to that used by Doing Business report; the authors assign a non-negative weighted coefficient for relative importance among the five criteria they analyze. However, the ranking with this method is generated by the general average of the performance of the alternatives, and there is no comparison between alternatives.

On the other hand, considering the modifications of the parameters by the decision maker, it can be observed that the implementation of MCDM methods, such as ELECTRE-III, allows a different weighting for each criterion in its weight of importance, as in the indifference thresholds, preference, and veto, allowing the decision maker to obtain the results according to the objectives that are intended to be achieved and selected the most appropriate alternative.

The relevance of using the ELECTRE-III method in problems where nations are compared and an exact evaluation of them is difficult because qualitative aspects are involved. It can be supported by the procedures for exploiting binary overcoming relationships [11], allowing the comparison between nations in an interval allocation format. That is similar to what was presented by Álvarez et al. [4] in its evaluation of the innovation capacity in the regions of Mexico, where the problem is assessed on qualitative scales for specific criteria, heterogeneous scales, and a certain arbitrariness in the construction of the criteria; characteristics that are also present in this investment site evaluation problem and that tend to justify the use of this family of methods in contexts with similar characteristics, where, additionally, the ELECTRE aggregation procedure does not allow compensation between

criteria performance, expressing clearly and simply the evaluation of each alternative in the different criteria, as they are valued [16][4].

As mentioned, multicriteria methods allow criteria to be considered with different weights according to their importance to the decision maker, unlike the report presented by the World Bank Group and the Doing Business 2020 report, where all criteria have the same weight, in the final ranking it is not possible to generate an in-depth analysis of the performance of the alternatives in each criterion. As well as, using a weighted sum method does not allow identifying in which criteria the alternatives differ or are similar to each other, being a relevant aspect when choosing a decision option.

Analyze different alternatives in a decision problem, specifying criteria and their relative importance for each one through weights, parameters of indifference and preference, implementing multicriteria methods for decision-making, such as the one applied in this research work with ELECTRE-III, allows to incorporate different points of view of the decision maker and even in an investment location problem, allows the option of considering ranges between the performances of the alternatives in a holistic way [8] based on their same preferences and objectives.

4.3. Sensitivity Analysis

Following the last action of Step 2 in Figure 1, it corresponds to the sensitivity analysis. The objective of the sensitivity analysis is to examine the robustness of the proposed solution, where the expert provides a range of values to modify in different parameters even considering their preferences [23] [5].

The procedures in this stage were established into scenarios that showed changes in the indifference and preference thresholds and weights values of some of the ten criteria considered for this problem. The expert, supported by the analyst, makes some modifications according to his understanding of the problem and preferences. Table 8 shows the used values on that parameter modification. The changes in the different scenarios were established as follows:

- Scenario 1. Change in a single criterion, g1, in the value of the preference threshold (p).
- Scenario 2. Change in a single criterion, g1, in the value of the indifference threshold (q).
- Scenario 3. Change in two criteria, g2, and g3, in the preference threshold values (p).
- Scenario 4. Change in two criteria, g2, and g3, in the values of the indifference threshold (q).
- Scenario 5. Change in two criteria, g2, and g3, in the weight values of the criteria (w).

In the sensitivity analysis process, five scenarios are made according to the variations of values of parameters (see Table 8). The ELECTRE-III method was applied in each scenario resulting in five rankings shown in Table 9.

The rankings of scenarios 1 and 2 are the same as the proposed ranking of the Table 3. On the other hand, in the remaining scenarios, there were minimal variations concerning the position of the nations, which are marked in gray. These variations consist mainly of a change in the position of the nations, such is the case of Scenarios 3, 4, and 5. For example, in Scenario 3, Georgia appears in eighth position and Taiwan in ninth, however,

Table 8. Parameters in sensitivity analysis scenarios

Scenario	Criteria change	Parameters		
		w	q	p
1	g1	0.2	5	20
2	g1	0.2	10	10
3	g2	0.16	5	13
	g3	0.15	5	14
4	g2	0.16	8	10
	g3	0.15	3	10
5	g2	0.18	5	10
	g3	0.13	5	10

Table 9. Ranking in sensitivity analysis scenarios

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 4
Rank	Economy	Economy	Economy	Economy	Economy
1	Hong Kong	Hong Kong	Hong Kong	Hong Kong	Hong Kong
2	Singapore	Singapore	Singapore	Singapore	Singapore
3	Denmark	Denmark	Denmark	Korea	Denmark
4	Korea	Korea	Korea	Denmark	Korea
5	United Kingdom	United Kingdom	United Kingdom	United Kingdom	New Zealand
6	New Zealand	New Zealand	New Zealand	New Zealand	United Kingdom
7	Lithuania	Lithuania	Lithuania	Taiwan	Taiwan
8	Taiwan	Taiwan	Georgia	Lithuania	Lithuania
9	Georgia	Georgia	Taiwan	United Arab Emirates	Georgia
10	United Arab Emirates	United Arab Emirates	Sweden	Georgia	Norway

in the ranking accepted by the expert (Table 3), these two nations are in reverse order, that is, Taiwan in eighth position and Georgia in ninth position. In the same way, in Scenarios 4 and 5, the nations shaded in gray present variation in position to the final ranking, as they are positioned in an inverted manner.

The nations shaded in green in Scenarios 3 and 4, Sweden and Norway, respectively, did not appear in the first ten positions in the proposal accepted by the expert; however, in that ranking, both nations were in the eleven and twelve positions and did not present significant variation concerning the preferences of the decision maker. This analysis concludes the sensitivity analysis process, showing stability in the ranking provided as the final to the expert.

4.4. Method Comparison: ELECTRE-III and PROMETHEE

In this section, a comparison between the outranking ELECTRE-III and PROMETHEE methods is performed. It aims to compare the rankings, highlighting coincidences and differences in the nations' orders.

PROMETHEE is an outranking method working with a finite set of alternatives based on often conflicting criteria [9]. It is based on six forms of preference functions (see [7]). Each criterion can apply any preference function.

- Usual Criterion
- U-shape Criterion (or quasi criterion)
- V-shape Criterion (or criterion with linear preference)
- Level criterion
- V-shape with indifference Criterion (or criterion with linear preference and indifference area)
- Gaussian Criterion

It seems the functions of preference V-Shape and V-Shape with an indifference criterion are most used [24]. In comparing methods, the selected preference function for PROMETHEE is V-Shape with an indifference criterion. The selected preference function in PROMETHEE is the same function ELECTRE-III uses to construct the fuzzy outranking relation between alternatives $\sigma(a_i, a_l)$. In the current preference function, the indifference (q) and preference (p) threshold are defined. The alternatives a_i and a_l are considered indifferent until the difference between them does not exceed q. Above this threshold, the degree of preference (p) increases linearly until a strict preference is reached [7].

In order to carry out the comparison of the results of ELECTRE-III and PROMETHEE, the same definition of indifference and preference thresholds was applied, $q = 5$ and $p = 10$, respectively. The weight considered for the criteria in this analysis considers the expert's preferences, where he considers three criteria particularly important, which can be observed in Table 5. Table 10 shows the ranking of the first ten positions of the nations according to the application of PROMETHEE and ELECTRE-III, respectively.

The resulting rankings from applying the PROMETHEE and ELECTRE-III methods (Table 10) show some differences in the position of nations. It is possible to observe that the ten nations continue in the first ten positions, with some position variations; Hong

Table 10. Ten first position of the PROMETHEE and ELECTRE-III rankings

	PROMETHEE	ELECTRE-III
Rank	Economy	Economy
1	Hong Kong	Hong Kong
2	Singapore	Singapore
3	New Zealand	Denmark
4	Denmark	Korea
5	Georgia	United Kingdom
6	Korea	New Zealand
7	United Kingdom	Lithuania
8	United Arab Emirates	Taiwan
9	Taiwan	Georgia
10	Lithuania	United Arab Emirates

Kong and Singapore persist in the same position (first and second, respectively) in both rankings.

The differences in the rankings lie from the third position, having that with the ranking obtained with PROMETHEE, New Zealand is positioned in the third place; on the other hand, this same nation, in the ranking with ELECTRE-III, is in the sixth position. Similarly, Georgia is the nation that occupies the fifth position in the ranking generated by PROMETHEE, and the ninth in the one generated by ELECTRE-III, with significant differences to be considered in the rankings.

An interesting aspect of the variation in the rankings is that ELECTRE placed Lithuania in seventh position, better than Taiwan and the United Arab Emirates. However, PROMETHEE places Lithuania in the tenth position. On the other hand, Korea and the United Kingdom present the same order between them in both rankings; however, PROMETHEE considers these nations in positions 6th and 7th, and ELECTRE-III places them in positions 4th and 5th, respectively.

With the above, despite applying the two methods considering the same weight parameters and indifference and preference thresholds, the compared methods differ mainly in the exploitation process of the valued matrix. When analyzing in detail the performance of the alternatives and their ordering, it is possible to observe that in the ranking generated by PROMETHEE, of the five best-positioned alternatives, two have the lowest performance in two of the three criteria considered most important by the expert. On the other hand, of the first five alternatives in the ranking obtained by ELECTRE-III, they generally have a better evaluation in the three criteria with the most significant weight for the decision-maker.

Analyzing the results obtained by both mentioned methods allows both the analyst and the decision maker to consider applying one method or another to the problem posed in this investigation, selecting the ELECTRE-III method as the one that provides the ranking that most reflects the user's preferences decision maker. One of the detected strengths of this method is that it allows a more intuitive differentiation for the decision maker in comparing pairs of alternatives, even when they have very similar performances in specific evaluation criteria; this is supported by the ELECTRE-III procedure by classifying nations

according to their concordance index and credibility matrix, thus evaluating the strength of the statement "a_i is at least as good as a_l".

In the current application, the ELECTRE-III is desirable when the DM is willing and knows the values of the indifference and preference situation when comparing alternatives by each criterion. It means she/he is able to define if "$g_j(a_i)$ is at least as good as $g_j(a_l)$". When the difference between alternatives is not in the indifference or preferences situation, then the fuzzy number represents the uncertainty. On the other hand, the ELECTRE-III is inadequate if the DM and analyst are unwilling to express the uncertainty with indifference and preference situation.

The binary outranking relations expressed in ELECTRE enables the formulation of recommendations in an interval format [12]. It supports the comparison of nations, particularly when it is difficult to derive exact investment assessments. The method represents the imperfect knowledge that characterizes the decision model.

The multicriteria aggregation procedure of ELECTRE is conceived such that they do not allow for compensation of performances among criteria [17]. The Countries' business attraction can be analyzed with the ELECTRE-III, in case the expert explicitly considers cases where business attraction deteriorates significantly due to poor performance on specific critical criteria.

The use of ELECTRE methods is particularly pertinent in contexts where at least one of the following features is present [17]: 1) the presence of qualitative scales for some criteria; 2) the presence of heterogeneous scales; 3) the need for avoiding systematic compensatory effects; 4) the need to take into account the imperfect knowledge of data and some arbitrariness when building criteria; and 5) the need of taking into account the reasons for and the reasons against an outranking. In the current problem, at least the features 3) and 4) are presented.

5. Conclusions

The problem of choosing an investment site is relevant for those investors who seek to select the ideal location to set up a company. In the same way, this is a relevant problem for the governmental leaders of the nations intending to be a competitive market and attract investors.

In that sense, in this paper, we have presented a ranking of nations according to their ease of doing business, applying the MCDM ELECTRE-III method, defining criteria weight parameters, indifference, and preference thresholds. This research evaluated various aspects of several nations, such as Dealing with construction permits, Registering property, and trading across borders, among others. That criteria are valuable information according to international organizations that impartially evaluate these aspects, serving as a reference for various studies.

Identifying investment locations can be challenging for business leaders and governments looking to expand their international presence, putting financial capital and global market position at risk. Thanks to the opening of markets, the need to be at the forefront of economic changes, and the accessibility of information through digital media, it is possible to have data that allows an evaluation of different locations for the installation of a company, this being an advantage to carry out a decision.

A large number of criteria are involved in the problems of choosing locations to invest; in this sense, presenting rankings obtained with a multicriteria decision-making method for investment location selection problems provides the opportunity to consider information published by international organisms, such as the World Bank Group by Doing Business study and report, with the possibility of adding other criteria that may also be of interest to decision-makers.

The multicriteria decision-making methods allow differentiating the weights of the criteria according to the objectives to be achieved and the decision-maker's preferences based on an investment decision and having that the application of ELECTRE-III for investment site selection problems, as well as similar problems, results in a ranking of the nations according to the assessment of each evaluated criterion corresponding to the weight assigned to each of them, thus causing obtaining information that allows decision-making that reflects the decision-maker's preferences and analyzes all the aspects defined as relevant for an optimal decision.

References

1. Álvarez, P., Muñoz-Palma, M., Miranda-Espinoza, L.E., Lopez-Parra, P., León-Castro, E.: Multiple criteria hierarchy approach for analyzing competitiveness of states in mexico. Inquietud Empresarial 20(2), 29–51 (2020)
2. Alvarez, P.A., Ishizaka, A., Martínez, L.: Multiple-criteria decision-making sorting methods: A survey. Expert Systems with Applications 183, 115368 (2021)
3. Alvarez, P.A., Morais, D.C., Leyva López, J.C., de Almeida, A.T.: An ELECTRE III based consensus-reaching process to improve a collective solution. International Transactions in Operational Research 29(2), 1048–1088 (2022)
4. Alvarez, P.A., Valdez, C., Dutta, B.: Analysis of the innovation capacity of mexican regions with the multiple criteria hierarchy process. Socio-Economic Planning Sciences 84, 101418 (2022)
5. Alvarez Carrillo, P.A., León Santiesteban, M., Gastelum Chavira, D.A., Vega Osuna, L.A.: An empirical analysis of competitiveness on cities of sinaloa, mexico with an outranking method. In: Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support. pp. 156–163. Atlantis Press (2013)
6. Borissova, D., Korsemov, D., Mustakerov, I.: Multi-criteria decision making problem for doing business: Comparison between approaches of individual and group decision making. In: Computer Information Systems and Industrial Management. pp. 385–396. Springer International Publishing, Cham (2019)
7. Brans, J.P., Vincke, P., Mareschal, B.: How to select and how to rank projects: The promethee method. European journal of operational research 24(2), 228–238 (1986)
8. Chejarla, K.C., Vaidya, O.S.: Ease of doing business: Performance comparison of g20 countries using gray mcdm. International Journal of Information Technology & Decision Making pp. 1–41 (2022)
9. De Lima Silva, D.F., Silva, J.C.S., Silva, L.G., Ferreira, L., De Almeida-Filho, A.T.: Sovereign credit risk assessment with multiple criteria using an outranking method. Mathematical Problems in Engineering 2018, 1–11 (2018)
10. Dias, J., Figueira, J.R., Roy, B.: The software ELECTRE III-IV: methodology and user manual (Version 3.X). Université Paris-Dauphine LAMSADE, Paris (2006)
11. Doumpos, M., Figueira, J.R.: A multicriteria outranking approach for modeling corporate credit ratings: An application of the electre tri-nc method. Omega 82, 166–180 (2019)

12. Doumpos, M., Figueira, J.R.: A multicriteria outranking approach for modeling corporate credit ratings: An application of the electre tri-nc method. Omega 82, 166–180 (2019)
13. Ekel, P., Bernardes, P., Vale, G.M.V., Libório, M.P.: South american business environment cost index: reforms for brazil. International Journal of Business Environment 13(2), 212–233 (2022)
14. Estevão, J., Lopes, J.D., Penela, D., Soares, J.M.: The doing business ranking and the GDP. a qualitative study. Journal of Business Research 115, 435–442 (2020)
15. Figueira, J., Roy, B.: Determining the weights of criteria in the ELECTRE type methods with a revised simos' procedure. European Journal of Operational Research 139(2), 317–326 (2002)
16. Figueira, J.R., Greco, S., Roy, B., Słowiński, R.: An overview of electre methods and their recent extensions. Journal of Multi-Criteria Decision Analysis 20(1-2), 61–85 (2013)
17. Figueira, J.R., Greco, S., Roy, B., Slowinski, R.: An overview of electre methods and their recent extensions. Journal of Multi-Criteria Decision Analysis 20(1-2), 61–85 (2013)
18. Garcia-Gastelum, T.S., Uzeta-Obregon, C.R., Alvarez, P.A., Espinoza-Audelo, L.F., León-Castro, E.: Building a regional employability indicator based on ordered weighted average operators. Cybernetics and Systems 53(7), 1–23 (2022)
19. Group, W.B.: Doing business: midiendo regulaciones para hacer negocios. World Bank (2020), [Online]. Available: https://espanol.doingbusiness.org/es/doingbusiness
20. Group, W.B.: Datos de libre acceso del banco mundial. World Bank (2021), [Online]. Available: https://datos.bancomundial.org/
21. Gryshova, I., I., K., Mitiai, O.: Ukraine's positions in international ratings evaluation as a factor of its competitiveness. Economic Annals-XXI 151(5–6), 24–27 (2015)
22. López, J.C.L., Carrillo, P.A.Á., Valenzuela, O.A.: A multicriteria group decision model for ranking technology packages in agriculture. In: Soft computing for sustainability science, pp. 137–161. Springer (2018)
23. Malczewski, J.: Multiple criteria decision analysis and geographic information systems. Trends in multiple criteria decision analysis pp. 369–395 (2010)
24. Martínez López, L., Ishizaka, A., Qin, J., Álvarez Carrillo, P.A.: Multi-Criteria Decision-Making Sorting Methods: Applications to Real-World Problems. Academic Press, London (2023)
25. Muñoz-Palma, M., Alvarez-Carrillo, P.A., Miranda-Espinoza, E.L., Avilés-Ochoa, E., León-Castro, E.: Multicriteria analysis model for the evaluation of the competitiveness of the states in mexico. In: Intelligent and Complex Systems in Economics and Business, pp. 1–20. Springer (2021)
26. Pinheiro-Alves, R., Zambujal-Oliveira, J.: The ease of doing business index as a tool for investment location decisions. Economics Letters 117(1), 66–70 (2012)
27. Roy, B.: The outranking approach and the foundations of ELECTRE methods. In: Bana e Costa, C. (ed.) Reading in Multiple Criteria Decision Aid, pp. 155–183. Springer-Verlag, Berlin (1990)
28. Ruiz, F., Cabello, J.M., Pérez-Gladish, B.: Building ease-of-doing-business synthetic indicators using a double reference point approach. Technological Forecasting and Social Change 131, 130–140 (2018)
29. Soares Silva, J.C., de Lima Silva, D.F., Ferrerira, L. de Almeida-Filho, A.T.: A dominance-based rough set approach applied to evaluate the credit risk of sovereign bonds. 4OR 20(1), 139–164 (2022)

Tanya Samantha Garcia-Gastelum is a PhD student in Administrative Sciences at the Autonomous University of Occident, Sinaloa, Mexico. She is a member of the State Network of Science and Technology Communicators in Sinaloa. Her research interests are in the fields of information and communication technology management, multi-criteria analysis, and decision-making support system.

Pavel Anselmo Alvarez Carrillo is a full professor at the Department of Economics and Management at the Autonomous University of Occident, Mexico. He holds a PhD in Management Science from the University of Occident, Mexico. The professor has achieved B.Sc. and M.Sc. in Applied Informatics from the Autonomous University of Sinaloa, Mexico. Alvarez's research interests include Group Decision Support Systems and Multicriteria Decision Analysis. Alvarez performed a Postdoctoral at the Federal University of Pernambuco, Brazil. The project regards Multicriteria Group Decision Making. He has been a guest editor in a special edition of the "International Journal of Computational Intelligence Systems".

Ernesto León-Castro Professor at Universidad Católica de la Santísima Concepcion (Chile) and Instituto Tecnológico de Sonora (México). PhD in Administrative Sciences from Universidad de Occidente (Mexico). Member of the National System of Researchers (Level 1). Within his research areas are aggregation operators, decision making, innovation, finance, fuzzy logic and other related topics. He has more than 50 publications in different journals indexed in the Web of Science / Scopus and has participated in different international conferences. He is part of the Ibero-American Network for Competitiveness, Innovation and Development (REDCID) and is a member of the Business Studies in Uncertain Environments research group. He has participated as thesis director/co-director at different universities in Mexico, Colombia and Chile. In addition, he has participated as a guest editor of different special editions of journals indexed in the Web of Science / Scopus and books in prestigious publishers.

Cristhian Ramón Uzeta-Obregon is a mechatronics engineer and has a master's degree in business administration. He is currently a PhD student in Administration Sciences at the Autonomous University of Occident, Sinaloa, Mexico. His research interest is in fields such as business administration, finance, multi-criteria decision making and markets.

Received: February 23, 2023; Accepted: October 23, 2023.

Computer Science and Information Systems 21(3):1203–1219 https://doi.org/10.2298/CSIS230213030B

The incidence of some critical cognitive factors that prevent change of order in university students: uncertainty analysis

José M. Brotons-Martínez[1], Rubén Chavez Rivera[2], and Jesús Ricardo Ramos-Sánchez[3]

[1] University Miguel Hernández,Elche, Spain
jm.brotons@umh.es
[2] University Michoacana de San Nicolás de Hidalgo, Morelia, Mexico
ruben.chavez@umich.mx
[3] Autonomous University of Tamaulipas, Mexico
lic.ramoss@hotmail.com

Abstract. School performance depends on cognitive impact while academic performance depends on how the qualitative and quantitative evaluation of students is carried out, which is open to a wide range of very complex criteria. Students' emotional intelligence is put to the test by the intense and constant completion of tasks to achieve the required credits. However, students do not manage to adapt to this system, which causes rejection and procrastination in completing tasks, and is evident from their really bad state of mind and stress intolerance. These factors can be so strong that they prevent adequate development; school commitments and obligations are neglected, affecting academic performance, which remains in the first order. Consequently, we aim to identify the critical cognitive factors that prevent the change of order and process them through fuzzy cognitive maps. To do so, a survey of students and teachers from Spanish and Mexican universities was carried out. Through confirmatory factor analysis, the number of items to be analysed was reduced. The variance analysis detected significant differences between students' attitude and what teachers thought, and an assignment matrix was obtained. Finally, by obtaining Hamming distance, the critical factors that prevent good academic performance could be found. The main result obtained is that the critical factors that prevent change of order in cognitive elements are mainly problem solving, stress tolerance, reality testing , empathy, self-concept and happiness.

Keywords: change of order, fuzzy, emotional intelligence, academic performance.

1. Introduction

School failure cannot be tackled through quantitative and qualitative evaluations alone. What is truly important is to define the educational objectives to be pursued by analysing the difficulties students come up against and proposing actions that enable them to continue progressing towards achieving them [33]. Academic performance from a psychological perspective comprises various cognitive factors such as intellectual level, personality, motivation, aptitudes, interest, study habits, self-esteem, and interpersonal relations ([35], [33]). However, in the different educational programs, meeting goals and objectives is represented by grades or marks awarded through quantitative and qualitative evaluations ([36], [6]). When students do not obtain the desired results, they lapse into a series of

justifications related to external factors, using them to shield themselves from any responsibility, which causes a recurrence of bad habits ([41], [42]). One of these bad habits is to put off taking action, which in turn prevents academic development ([11], [26]).

From a psychological perspective, the identification of cognitive-emotional patterns starts with the teacher's observation of their students. When they observe signs of a behaviour that prevents a student from accomplishing their learning objectives, their tutor will be informed so that they can receive early attention; since, if this is not done, a student could experience a cognitive weakness. Students can be conditioned by their own prejudices or by external events that prevent the full development of their activities, which is known as a first order state ([19]). Psychological support enables these weaknesses to be overcome by making substantial changes. If these changes manage to restructure their ways of interacting and lead them to a personal transformation, a student would then be in a second order state ([4],[46], [29], [19], [16]).

If we consider the first order, where students do not manage to overcome lagging behind in their academic and school performance, students manifest great dissatisfaction with the educational system; as a result, this prevents them from fully developing their cognitive skills ([25], [39]): According to Lamas [33], this dissatisfaction is expressed through impulsive actions that affect their learning capacity and regularly affects crystalized intelligence (teaching-learning process) but not fluid intelligence (ability to establish relationships independent of previously acquired knowledge).

Emotional intelligence can be defined as the capacity that determines learning practical skills based on self-awareness, motivation, self-control, empathy and self-regulation ([24], [20]). According to Vargas ([56]), there are some basic principles that influence people's sensitivity in their interpretation of others' emotions and feelings when socializing in groups. In this sense, Bar-On ([1]) considers that an individual relates to the people and the cultural environment around them, where intrapersonal, interpersonal, adaptability, stress management and state of mind are implicit.

Regardless of educational level, any learning process involves various sets of tasks and demands that require a large number of psychological as well as physical adaptation skills ([7], [17], [45], [13]). Generally speaking, the cognitive factors that influence the skill for understanding and showing emotions and feelings are related to emotional self-awareness, assertiveness, self-development and emotional independence ([54], [55]). One of the factors to emerge from this interpersonal dimension is self-concept, understood as the ability to understand, accept, and respect oneself. Personality and the relationship with academic, social and family self-concept is associated with self-esteem (comprising various behavioural, cognitive, emotional, and interaction variables. The patterns recognised by [22] are: withdrawal; anxiety-shyness; leadership; prosocial, anti-social, assertive, passive, or aggressive behaviours; self-assertiveness; hetero-assertiveness; and social adaptation. The interventions of these factors are behind the causes of school stress in general; a phenomenon that affects this community significantly today ([36], [8]). The theoretical structure of stress is classified as positive, normal and negative. Positive stress (eustress) permits an individual to be alert and to confront situations that they consider a threat to their performance, normal stress is necessary to confront situations that require alertness, and negative stress (distress) refers to the inability to confront threats, which can be frustrating for an individual ([2], [40], [23], [7]). Nevertheless, research into academic stress

management is still in its initial phase, and many studies have only focused on work-related stress ([3], [47])

Consequently, this study has two aims: one is to identify the differences between lagging and non-lagging students; while the other is to determine which items show a greater difference between them by using fuzzy cognitive maps.

In this article, through Bar-On surveys of non-lagging students, lagging students, and teachers, and the use of FCM, we analyse why students fall behind in their studies. Using factorial analysis, the variables included in the study are reduced and the subsequent variance analysis enables the construction of an assignment matrix in order to generate a weight matrix, which is used as the basis for applying FCM. This application will enable us to find the aspects that lead students to fall behind in their studies.

2. Methodology

The identification of the cognitive indicators of lagging students is proposed in order to identify the behaviour patterns that prevent their academic development. To do this, the following steps will be followed:

1. Identification of the cognitive-emotional factors that differentiate lagging students from non-lagging students
   (a) A Bar-On survey is carried out on lagging students (LS) so as to gather information about the following dimensions: interpersonal, intrapersonal, adaptability, stress management, and state of mind.
   (b) We perform a confirmatory factor analysis of 60 questions answered by lagging students through the Kaiser-Meyer-Olkin test (KMO) and Bartlett's test of sphericity in order to determine partial correlation between the variables and lagging students. This makes it possible to reduce the number of items; that is to say, the questions in the initial survey that refer to the cognitive-emotional factors.This reduced group of items is used to interview non-lagging students and teachers. With the partial correlation matrix, the initial situation of the case study of lagging students in contrast to non-lagging students will be obtained, which will serve as a basis for adjusting weight assignments wij in the assignment matrix.The weights are interpreted according to Table 1, oscillating between -1 for a p-value close to zero and +1 for a p-value close to one. The p-value corresponds to the cognitive-emotional factors in each group of students (lagging and non-lagging).
   (c) A variance analysis (ANOVA) is performed for each of the items selected in the previous section between the three target groups,non-lagging students, lagging students and teachers. The ANOVA analysis enables us to find out the cognitive patterns that indicate significant differences through the subsequent application of the Scheffé method (for groups of different sizes). The results of the ANOVA and the Scheffé method allow us to visualize the differences for each item between the three target groups of the analysis so as to detect the cognitive variables that present a greater impact on academic performance.
2. The search for the main differences between groups is carried out by following the steps outlined below:
   (a) To obtain the assignment matrix the following procedure is carried out:

i. The levels of significance obtained are transformed into weights according to Table 1. If the significance level of the null hypothesis is lower than 0.05 (Fisher distribution), as it is opposite to the null hypothesis, negative ωs weights will be taken, interpolating for the values not considered in the table. Conversely, if the acceptance level is greater than 0.05 the positive ωs weights will be taken in a similar way.

Table 1. Correspondence between p-value and weights (ωs)

Pvalue	0.000	0.010	0.025	0.030	0.040	0.050	0.051	0.095	0.380	0.570	0.760	0.950
ωs	-1	-0.8	-0.5	-0.4	-0.2	-0.1	0.1	0.2	0.4	0.6	0.8	1.0

ii. All variables belonging to the same dimension will have a value of ± 1 according to whether or not they are opposite to the null hypothesis.

iii. Based on the ωs vector, the assignation matrix for the cognitive variables is obtained, indicating the incidence between each pair of variables

$$\omega_{ij} = |\omega s_i + \omega s_j|/2 \tag{1}$$

Where ω_{ij} is the element ij of the assignment matrix, ωs_i is the significance value of the cognitive variable in row i and ωs_j is the significance value of the cognitive variable in column j. Two matrices will be constructed, one for non-lagging students and another for lagging students. Once all the elements of the cognitive matrix are obtained, academic performance (AP) will be evaluated, which is the sum of all the elements from row k divided between the number of elements in this row. Once all the elements of the cognitive matrix are obtained, AP will be evaluated, which is the sum of all the elements in row k divided by the number of elements in said row.

$$RA_i = \frac{1}{k}\sum_{j=1}^{k} \omega_{ij} \tag{2}$$

In addition, because there are two groups of students (lagging and non-lagging), the analysis requires two matrices to compare the factors between the group of non-lagging students and the group of lagging students. The first matrix contains the group of lagging students according to the valuation by teachers, while the second matrix comprises the group of non-lagging students according to the teachers' valuation.

(b) To identify, define and validate the cause-effect relations in the assignment matrix, fuzzy cognitive maps (FCM) are used as they enable decision-making strategies to be established ([18], [59]). To do so, the assignment matrix should be constructed with predictive and reflexive sensitivity oriented towards the actions carried out by the experts ([18], [28]). The experts adjust the weights of the rotated component matrix (ANOVA) according to their experience to form an assignment matrix. This adjustment means possible biases in the rotated component matrix

(survey carried out on students) can be avoided because students may misrepresent their cognitive-emotional weaknesses, and in this way it is closer to reality. ([19], [30], [15], [58]).

In order to do this, an analysis is made through graph analysis, using causal concepts and connections, where 0 indicates the absence of concept while the value +1 or -1 indicates the positive or negative presence between the two connected concepts ([12], [30]). Using assignments through experts gives certainty to the qualitative information for the interpretation of the causes and effects constructed in an assignment matrix ([32], [10], [44], [18]). In accordance with the iterative procedure, the resulting vector of concepts in state $t+1$, C_{t+1} is updated according to the t state that immediately precedes it (C_t). This update in turn depends on the weight matrix ω and the transfer function f, $C_{t+1} = f(C_t, \omega)$.

(c) The search for the critical factors that affect academic performance. The FCM method is used through multiplication of the initiator concept vector, which is chosen by discretion: $C^t = [C_i^t]_{i\times n}$ with $C_i^t \in [0,1]$ and matrix ω_{ij} , according to the following expression:

$$R_t = C_t\omega \tag{3}$$

The new vector C_{t+1} is obtained from the R_t result, each i element will be one for those elements of the product such that $R_{ti} \geq 0$, and zero for the elements of the product such that $R_{ti} < 0$. Thus, the new vector C_{t+1} obtained will be formed by zeroes and ones. This new vector will again multiply the matrix $[\omega_{ij}]$ and so on and so forth for k iterations until the vector of iteration $k+1$, C_{t+k+1} is equal to the previous vector C_{t+k}.

(d) The valuation of Hamming distance allows us to identify the critical variables that prevent a satisfactory academic performance. The Hamming distance of the two fuzzy subsets $\tilde{LS}$ and $\tilde{NS}$ is found in an interval $0, 1$ and is obtained with the following expression:

$$\delta\left(\tilde{LS}, \tilde{NS}\right) = \frac{1}{n}\sum_{i,j=1}^{n} |\mu_i^{LS} - \mu_i^{LS}| \tag{4}$$

Where μ_i^{LS} and μ_i^{NS} are the pertinence functions of the sets $\tilde{LS}$ and $\tilde{NS}$ once equilibrium is achieved in the cognitive maps. Each indicator represents the transfer function of academic performance for the subset $\tilde{LS}$, which represents the 63 lagging students, while $\tilde{NS}$ represents non-lagging students.

3. Results and discussion

A pilot survey was carried out on lagging students. Subsequently, the number of items was reduced using the KMO and Bartlett's test, with no loss of information or variable objectivity. This was done to make the survey available to 160 students and 22 teachers from three universities in Spain and Mexico. An analysis was then made of the relation of cognitive factors (intrapersonal, adaptability, stress management and state of mood) between lagging students and teachers, and between non-lagging students and teachers. Assignment matrices were obtained for the groups of lagging students with teachers and

the groups of non-lagging students with teachers. Afterwards, fuzzy cognitive maps were applied, assigning an initiator vector to both groups. Finally, once stability was reached for both groups Hamming distance between the groups was obtained to identify critical variables.

A Bar-On (2006) survey (Table 2)is made of 63 students who are lagging academically in the first semester of 2021/22 academic year in mathematics, algebra, calculus and physics at the Faculty of Civil Engineering of the University Michoacana de San Nicolás de Hidalgo (UMSNH).

Table 2. Survey made of lagging students

p1-I like to enjoy myself	p21-I argue with people	p41-I make friends easily
p2-I'm good at understanding others	p22-I understand difficult questions	p42-I think I'm the best at everything I do
p3-I can keep calm when I'm annoyed	p23-I like smiling	p43- I have no problem saying how I feel
p4-I'm happy	p24-I try not to hurt other people's feelings	p44-When I answer difficult questions, I try to think of many solutions
p5-I care about what happens to people	p25-When faced with a problem, I don't give up until I've solved it	p45-I feel bad when people's feelings are hurt
p6-It's difficult for me to control my anger	p26-I have a bad temper	p46-when I'm upset with someone I continue to feel up set for a long time.
p7-I find it easy to tell people how I'm feeling	p27-nothing upsets me	p47-I feel happy about the type of person I am.
p8-I like all the people I know	p28-it's difficult for me to talk about my inner feelings.	p48-I'm good at solving problems
p9-I feel confident	p29-I know things will work out well	p49-It's difficult form to wait my turn
p10-I know how people feel	p30-I can give good answers to difficult questions	p50-I enjoy the things I do
p11-I know how to keep calm	p31-I can easily describe my feelings	p51-I'm happy with my friends
p12-I try to answer difficult questions in different ways	p32-I know how to enjoy myself	p52-I don't have bad days
p13-what I do is good	p33-I must always tell the truth	p53-it is difficult for me to tell others about my feelings
p14-I'm respectful towards others	p34-I am able to answer difficult questions in many wats when I want to	p54-I'm easily bothered
p15-I get over upset for any reason	p35-I easily get upset	p55-I?m able to realise when my friend is sad
p16-It´s easy for me to understand new things	p36-It pleases me to do things for others	p56-I like my body
p17-I'm able to talk about my feelings easily	p37-I don't feel very happy	p57-even if things are difficult, I don't get up
p18-I think well of everybody	p38-I can use different ways of solving problems easily	p58-when I'm upset, I act without thinking
p19-I hope for the best	p39-It takes a long time for me to get upset	p59-I know when people are upset, even if they don't say anything
p20-having friends is important	p40-I feel good about myself	p60-I like the way I look

The number of items (number of questions in the initial survey) is reduced according to the confirmatory analysis using the KMO test (approval level of the survey equal to 0.709) and Bartlett's test of sphericity with zero value in order to eliminate the variables that correlate. Results are given in Table 3, which, as can be observed, is not an identity matrix; therefore, we can rule out that the correlations).Table 4 shows the reduction of items obtained for each of the indicators from each dimension through rotated component matrix.

Table 3. Reduced rotated component matrix of dimensions

	Component					
	1	2	3	4	5	6
p17- I'm able to talk about my feelings easily	.778	.002	.237	.087	-.078	.105
p31- I can easily describe my feelings	.737	-.064	.112	.025	.080	.250
P43-I have no problems saying what I feel	.626	.473	.116	-.007	-.194	-.180
P7- I find it easy to tell people how I'm feeling	.594	.048	.526	-.061	.016	-.140
p30- I can give good answers to difficult questions	.553	.334	-.059	.342	.156	.237
p40- I feel good about myself	.072	.857	.187	.000	-.069	.059
p47- I feel happy about the type of person I am	.060	.840	-.013	.136	-.001	.207
p25- When faced with a problem, I don't give up until I've solved it	.089	.568	.069	.466	.080	.172
p10- I know how people feel	.349	-.069	.770	.107	-.060	.090
p45- I feel bad when people's feelings are hurt	-.118	.403	.678	.088	-.008	.323
p2- I'm good at understanding others	.267	.126	.664	.232	.130	-.049
p48- I'm good at solving problems	.191	.024	.142	.792	.081	.016
p44-When I answer difficult questions, I try to think of many solutions	-.039	.146	.170	.758	-.201	.144
p3- I can keep calm when I'm annoyed	-.037	-.014	.002	-.020	-.823	.009
p21- I argue with people	.044	-.180	.153	-.291	.770	-.041
p35- I easily get upset	-.159	.147	-.124	.424	.635	-.161
p29- I know things will work out well	.162	.227	.037	.132	-.169	.854
p4- I'm happy	.465	.215	.208	.066	.029	.491

A survey of non-lagging students and teachers was made with the resulting items in Table 4 (fourth column) in order to find out if there are any significant differences in the cognitive-emotional effect on academic performance and how it is perceived in universities. Specifically, 160 non-lagging students (NS) were interviewed: 124 were from the UMSNH (Mexico), 21 were students from University Miguel Hernández (Spain), and 15 students were from the Technology University of Nuevo Ladero (Mexico). The group of 22 teachers were also from these three universities. The survey represents the five cognitive dimensions comprising 18 items (Table 4). This reduced survey adapted for

Table 4. Dimensions, indicators and relation with questions students were asked

Dimension	Indicator	Items	Reduced Items
Interpersonal (ITR)	Empathy (EM)	2, 5, 10, 45, 55, 59	2, 5
	Interpersonal Relation (IR)	14, 20, 41, 51	
	Social responsibility (SR)	36, 57	
Intrapersonal	Emotional understanding (EU)	9, 17, 28, 31, 40	17, 31, 40
	Assertiveness (AS)	7, 24, 33, 43, 53, 59	7, 43
	Self-concept (SC)	3, 6, 26	3
	Self-realization (SR)	13, 42, 57	
	Independence (IN)	32, 47	32, 47
Adaptability	Problem solving (PS)	12, 22, 25, 30, 34, 38, 44, 48, 57	25, 30, 48
	Reality testing (RT)	3, 7, 11, 12, 16, 33, 38, 44	44
	Flexibility (FL)	27, 49, 56, 60	
Stress management	Stress tolerance (ST)	3, 11, 35, 49, 54, 58	35
	Impulse control (IC)	6, 15, 21, 39, 46, 49, 54, 58	21
State of mood	Optimism (OP)	8, 18, 19, 29, 52, 60	29
	Happiness (HP)	1, 4, 23, 37, 50	4

teachers is shown in Table 5 (the non-lagging students were given the same survey as lagging students, but only with the selected items).

Table 5. Survey for teachers

Item	Questions for teachers	Indicator
2	Lagging students (LS) are good at understanding people when there are changes	EM
5	LS care about what happens to people who face the same changes as they do	EM
17	LS can talk easily about what they think	CE
31	LS can describe their feelings easily	CE
40	LS feel good about themselves	CE
7	LS find it easy to tell people what they feel about the changes	AS
43	LS have no problem syaing what they feel	AS
3	LS can keep calm when they are upset about changes	SC
32	LS know how to enjoy themselves	IN
47	LS feel happy about the type of person they are	IN
25	LS don't give up until they've solved a problem	PS
30	LS usually give good answers to difficult questions	PS
48	LS are good at solving difficult things	PS
44	When LS answer, they try to think of many possible solutions	RT
35	LS get upset easily	ST
21	LS tend to be aggressive towards people	IC
29	LS feel confident about meeting goals	OP
4	LS feel happy about the new challenges that face them at this time	FZ

In order to know about the cognitive patterns and their relation between factors, we will perform ANOVA to find out the significant differences in pairs between the two groups of students, lagging students and teachers, and between non-lagging students and teachers (Table 5) . To do so, the five dimensions are analysed in their different levels of significance.

1. Interpersonal dimension. There are only differences between the two groups of students for item number 6 (Table 3) for the indicator Empathy EM_5. However, the

teachers did show differences of opinion with respect to two items of this indicator (except for item EM_5 in the case of non-lagging students)

Table 6. ANOVA significant differences between the three pairs (the two groups of students, lagging students and teachers, and non-lagging students and teachers)

Dimension	Variable / Indicator	Item	P-value			Description of item
			Scheffé NS and LS	Scheffé T and LS	Scheffé T and NS	
Interpersonal	Empathy (EM)	2	0.056	0.000*	0.000*	They understand people when there are changes
		5	0.043*	0.000*	0.264	They care about what happens to people when facing changes
Intrapersonal	Emotional understanding (EU)	17	0.000*	0.012*	0.062	They can talk easily about what they think
		31	0.204	0.95	0.283	They can describe their feelings easily
		40	0.626	0.170	0.037*	They feel good about themselves
	Assertiveness (AS)	7	0.339	0.114	0.574	They can tell people what they feel about the changes
		43	0.288	0.506	0.073	They show people how they feel
	Self-concept (SC)	3	0.040*	0.378	0.006*	They keep calm when they are upset about changes
		47	0.111	0.727	0.780	They feel happy about the type of person they are
Adaptability	Problem solving (PS)	25	0.000*	0.000*	0.990	They don't give up until they've solved a problem
		30	0.028*	0.910	0.063	They give good answers to difficult questions
		48	0.198	0.432	0.037*	They're good at solving difficult things
	Reality testing (RT)	44	0.000*	0.008*	0.589	They try to think of many possible solutions
Stress, management	Stress tolerances (ST)	35	0.000*	0.998	0.003*	They get upset easily
	Impulse control (IC)	21	0.000*	0.005*	0.016*	They tend to be aggressive towards people
State of mood	Optimism (OP)	29	0.927	0.000*	0.000*	They feel confident about meeting goals
	Happiness (HP)	4	0.137	0.001*	0.076	They feel happy about the new challenges that face them at this time

2. Intrapersonal dimension. Both groups of students showed significant differences for items EU_{17} of Emotional Understanding and SC_3 of Self-concept. For the former, the teachers also differed from the lagging students, while for the latter they differed from both groups of students. .
3. Adaptability dimension. The items Problem Solving PS_{25} and SP_{30} and Reality Testing (RT_{44}) showed significant differences between the two groups of students. The teachers showed differences of opinion for item SP_{25} and PR_{44} with respect to lagging students.
4. Stress management dimension. The items Stress Tolerance (ST_{35}) and Impulse Control (CI_{21}) showed differences between the two groups of students. The teachers opinion differed for item CI_{21} with respect to lagging students.

5. State of mood. None of the items in this group showed significant differences between the two groups of students, while the teachers showed differences in the two items with respect to lagging students.

Matrix assignation. Two comparison matrices are considered: one between the group of lagging students and the group of teachers (Table 7), and the other between the group of non-lagging students and the group of teachers (Table 8).

Table 7. Assignation matrix ($\omega_{ijSL\times E}$)

Iteration	EM_2	EM_5	EU_{17}	EU_{31}	EU_{40}	AS_7	AS_{43}	SC_3	SC_{47}	PS_{25}	PS_{30}	PS_{48}	RT_{44}	ST_{35}	IC_{21}	OP_{29}	HP_4	AP
EM_2	-1.00	-1.00	-0.90	0.00	-0.63	-0.40	-0.24	-0.30	-0.12	-1.00	-0.02	-0.28	-1.00	0.00	-1.00	-1.00	-1.00	-0.58
EM_5	-1.00	-1.00	-0.90	0.00	-0.63	-0.40	-0.24	-0.30	-0.12	-1.00	-0.02	-0.28	-1.00	0.00	-1.00	-1.00	-1.00	-0.58
EU_{17}	-0.90	-0.90	-0.45	0.20	-0.28	-0.43	-0.35	-0.40	-0.19	-0.90	0.16	-0.27	-0.90	0.20	-0.90	-0.90	-0.90	-0.48
EU_{31}	0.00	0.00	0.20	0.84	0.63	0.61	0.77	0.70	0.89	0.00	0.98	0.73	0.00	1.00	0.00	0.00	0.00	0.43
EU_{40}	-0.38	-0.38	-0.30	0.63	0.26	0.37	0.31	0.33	0.51	-0.75	0.61	0.35	-0.75	0.63	0.75	0.75	0.75	0.22
AS_7	-0.34	-0.34	-0.30	0.61	0.23	0.22	0.37	0.31	0.49	-0.40	0.59	0.33	-0.40	0.61	-0.40	-0.40	-0.40	0.05
AS_{43}	-0.24	-0.24	-0.14	0.77	0.39	0.37	0.51	0.47	0.65	0.24	0.75	0.47	-0.24	0.77	-0.24	-0.24	-0.24	0.22
SC_3	-0.30	-0.30	-0.20	0.70	0.33	0.31	0.47	0.42	0.59	-0.30	0.68	0.43	-0.30	0.70	-0.30	-0.30	-0.30	0.14
SC_{47}	-0.12	-0.12	-0.02	0.89	0.50	0.49	0.65	0.59	0.67	-0.12	0.87	0.61	-0.12	0.89	-0.12	-0.12	-0.12	0.31
PS_{25}	-1.00	-0.02	-0.91	0.00	-0.38	-0.58	-0.52	-0.30	-0.12	-1.00	-0.02	-0.28	-1.00	0.00	-1.00	-1.00	-1.00	-0.54
PS_{30}	-0.05	-0.05	0.08	0.80	0.61	0.59	0.75	0.68	0.87	-0.02	0.81	0.71	-0.12	0.98	-0.12	-0.12	-0.12	0.37
PS_{48}	-0.28	-0.28	-0.18	0.73	0.35	0.33	0.49	0.43	0.61	-0.28	0.71	0.46	-0.28	0.73	-0.28	-0.28	-0.28	0.16
RT_{44}	-1.00	-1.00	-0.90	0.00	-0.63	-0.40	-0.24	-0.30	-0.12	-1.00	-0.02	-0.28	-0.51	0.00	-1.00	-1.00	-1.00	-0.55
ST_{35}	0.00	0.00	0.20	1.00	0.63	0.61	0.77	0.70	0.89	0.00	0.98	0.73	0.00	1.00	0.00	0.00	0.00	0.44
IC_{21}	-1.00	-1.00	-0.90	0.00	-0.63	-0.40	-0.24	-0.30	-0.12	-1.00	-0.02	-0.28	-1.00	0.00	-1.00	-1.00	-1.00	-0.58
OP_{29}	-1.00	-1.00	-0.90	0.00	-0.63	-0.40	-0.24	-0.30	-0.12	-1.00	-0.02	-0.28	-1.00	0.00	-1.00	-1.00	-1.00	-0.58
HP_4	-1.00	-1.00	-0.90	0.00	-0.63	-0.40	-0.24	-0.30	-0.12	-1.00	-0.02	-0.28	-1.00	0.00	-1.00	-1.00	-1.00	-0.58
PA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 8. Assignation matrix ($\omega_{ijSN\times E}$)

	EM_2	EM_5	EU_{17}	EU_{31}	EU_{40}	AS_7	AS_{43}	SC_3	SC_{47}	PS_{25}	PS_{30}	PS_{48}	RT_{44}	ST_{35}	IC_{21}	OP_{29}	HP_4	AP
EM_2	-1.00	-0.34	-0.41	-0.33	-0.52	-0.22	-0.41	-0.78	-0.15	-0.07	-0.41	-0.52	-0.22	-0.80	-0.69	-1.00	-0.41	-0.48
EM_5	-0.34	0.33	0.26	0.34	0.15	0.45	0.26	-0.11	0.52	0.60	0.26	0.15	0.45	-0.14	-0.03	-0.34	0.26	0.18
EU_{17}	-0.41	0.26	0.18	0.27	0.08	0.37	0.19	-0.19	0.45	0.52	0.19	0.08	0.38	-0.21	-0.10	-0.41	0.19	0.11
EU_{31}	-0.33	0.34	0.27	0.35	0.16	0.46	0.27	-0.10	0.53	0.61	0.27	0.16	0.46	-0.13	-0.02	-0.33	0.27	0.19
EU_{40}	-0.52	0.15	0.08	0.16	-0.03	0.27	0.08	-0.29	0.34	0.42	0.08	-0.03	0.27	-0.32	-0.21	-0.52	0.08	0.00
AS_7	-0.22	0.45	0.37	0.46	0.27	0.56	0.38	0.01	0.64	0.71	0.38	0.27	0.57	-0.02	0.09	-0.22	0.38	0.30
AS_{43}	-0.41	0.26	0.19	0.27	0.08	0.38	0.19	-0.18	0.45	0.53	0.19	0.08	0.38	-0.21	-0.10	-0.41	0.19	0.11
SC_3	-0.78	-0.11	-0.19	-0.10	-0.29	0.01	-0.18	-0.55	0.08	0.16	-0.18	-0.29	0.01	-0.58	-0.47	-0.78	-0.18	-0.26
SC_{47}	-0.15	0.52	0.45	0.53	0.34	0.64	0.45	0.08	0.71	0.79	0.45	0.34	0.64	0.06	0.17	-0.15	0.45	0.37
PS_{25}	-0.07	0.60	0.52	0.61	0.42	0.71	0.53	0.16	0.79	0.86	0.53	0.42	0.72	0.13	0.24	-0.07	0.53	0.45
PS_{30}	-0.41	0.26	0.19	0.27	0.08	0.38	0.19	-0.18	0.45	0.53	0.19	0.08	0.38	-0.21	-0.10	-0.41	0.19	0.11
PS_{48}	-0.52	0.15	0.08	0.16	-0.03	0.27	0.08	-0.29	0.34	0.42	0.08	-0.03	0.27	-0.32	-0.21	-0.52	0.08	0.00
RT_{44}	-0.22	0.45	0.38	0.46	0.27	0.57	0.38	0.01	0.64	0.72	0.38	0.27	0.57	-0.02	0.10	-0.22	0.38	0.30
ST_{35}	-0.80	-0.14	-0.21	-0.13	-0.32	-0.02	-0.21	-0.58	0.06	0.13	-0.21	-0.32	-0.02	-0.60	-0.49	-0.80	-0.21	-0.28
IC_{21}	-0.69	-0.03	-0.10	-0.02	-0.21	0.09	-0.10	-0.47	0.17	0.24	-0.10	-0.21	0.10	-0.49	-0.38	-0.69	-0.10	-0.17
OP_{29}	-1.00	-0.34	-0.41	-0.33	-0.52	-0.22	-0.41	-0.78	-0.15	-0.07	-0.41	-0.52	-0.22	-0.80	-0.69	-1.00	-0.41	-0.48
HP_4	-0.41	0.26	0.19	0.27	0.08	0.38	0.19	-0.18	0.45	0.53	0.19	0.08	0.38	-0.21	-0.10	-0.41	0.19	0.11
PA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

The initiator concept vector of empathy in the variables EM_2 and EM_5 was $C_{t=0} = [1100000000000000]$. Successive iterations were made with matrix $\omega_{ijSL\times E}$ as well

as matrix $\omega_{ijSN\times E}$. Table 9 shows the results of the successive iterations for lagging students versus teachers and their effect on academic outcome, while Table 10 shows the results for non-lagging students versus teachers. As can be observed in the graph in Figure 1, the items that negatively affect academic performance are: EM_2, EM_5, EU_{17}, PS_{25}, RT_{44}, IC_{21}, OP_{29} and HP_4. Once equilibrium was achieved, Hamming distance was

Table 9. Iteration results for $R_{t+1,LS\times T} = C_{t+1} * \omega_{LS\times T}$

Iteration	EM_2	EM_5	EU_{17}	EU_{31}	EU_{40}	AS_7	AS_{43}	SC_3	SC_{47}	PS_{25}	PS_{30}	PS_{48}	RT_{44}	ST_{35}	IC_{21}	OP_{29}	HP_4	AP
1	-2.00	-2.00	-1.80	0.00	-1.26	-0.80	-0.48	-0.60	-0.20	-2.00	-0.04	-0.56	-2.00	0.00	-2.00	-2.00	-2.00	-1.20
2	0.00	0.00	0.40	1.84	1.26	1.22	1.54	1.40	1.78	0.00	1.96	1.46	0.00	2.00	0.00	0.00	0.00	0.87
3	-9.60	-8.63	-7.42	7.17	0.51	0.49	2.78	2.13	5.14	-9.53	7.00	2.54	-9.62	7.51	-8.61	-8.61	-8.61	-1.10
4	-1.71	-1.71	-0.66	6.97	3.93	3.90	5.09	4.63	6.17	-1.63	6.98	4.82	-2.21	7.31	-0.71	-0.71	-0.71	2.34
5	-1.71	-1.71	-0.66	6.97	3.93	3.90	5.09	4.63	6.17	-1.63	6.98	4.82	-2.21	7.31	-0.71	-0.71	-0.71	3.34
6	-1.71	-1.71	-0.66	6.97	3.93	3.90	5.09	4.63	6.17	-1.63	6.98	4.82	-2.21	7.31	-0.71	-0.71	-0.71	3.34

Table 10. Iteration results for $R_{t+1,NS\times T} = C_{t+1} * \omega_{NS\times T}$

Iteration	EM_2	EM_5	EU_{17}	EU_{31}	EU_{40}	AS_7	AS_{43}	SC_3	SC_{47}	PS_{25}	PS_{30}	PS_{48}	RT_{44}	ST_{35}	IC_{21}	OP_{29}	HP_4	AP
1	-1,34	0,67	-0,16	0,02	-0,37	0,23	-0,15	-0,89	0,38	0,53	-0,15	-0,37	0,24	-0,94	-0,72	-1,33	-0,15	-0,26
2	-1,31	3,35	2,23	3,39	1,6	3,81	2,26	0,04	4,11	4,41	2,26	1,6	3,83	-0,11	0,55	-1,31	2,26	1,94
3	-5,43	4,55	3,65	4,67	0,39	5,93	3,71	-2,73	6,83	7,73	3,71	0,39	5,99	-2,63	-1,71	-5,43	3,71	2,95
4	-3,97	4,69	3,94	4,79	0,89	5,84	3,99	-1,27	6,59	7,34	3,99	0,89	5,89	-1,57	-0,25	-3,97	3,99	3,45
5	-3,97	4,69	3,94	4,79	0,89	5,84	3,99	-1,27	6,59	7,34	3,99	0,89	5,89	-1,57	-0,25	-3,97	3,99	3,45

used to evaluate the distance between the values of the variables corresponding to the last rows of Tables 9 and 10. The results are shown in Table 11.

Table 11. Hamming distances

Iteration	EM_2	EM_5	EU_{17}	EU_{31}	EU_{40}	AS_7	AS_{43}	SC_3	SC_{47}	PS_{25}	PS_{30}	PS_{48}	RT_{44}	ST_{35}	IC_{21}	OP_{29}	HP_4	AP
LS*T	-1.71	-1.71	-0.66	6.97	3.93	3.90	5.09	4.63	6.17	-1.63	6.98	4.82	-2.21	7.31	-0.71	-0.71	-0.71	3.34
NS*T	-3.97	4.68	3.93	4.78	0.88	5.83	3.98	-1.26	6.58	7.33	3.98	0.88	5.88	-1.56	-0.24	-3.96	3.98	3.45
Δ	2.26	6.40	4.60	2.20	3.05	1.94	1.10	5.90	0.42	8.96	3.00	3.94	8.10	8.88	0.47	3.26	4.68	0.11

In Table 11, the indicators of why lagging students have a low academic performance can be observed in the difference between non-lagging students and lagging students. The 10 items of greatest distance between the groups ordered from the greatest to the least are:

$$SP_{25} > ST_{35} > RT_{44} > EM_5 > SC_3 > HP_4 > EU_{17} > PS_{48} > OP_{29} > EU_{40} \tag{5}$$

This leads us to the following analysis of each of the five dimensions:

1. Interpersonal. The differences in the variances of the empathy factor for the group of non-lagging students and the group of lagging students reveals the influence of emotional intelligence and the differences in how they perceive, understand, and manage

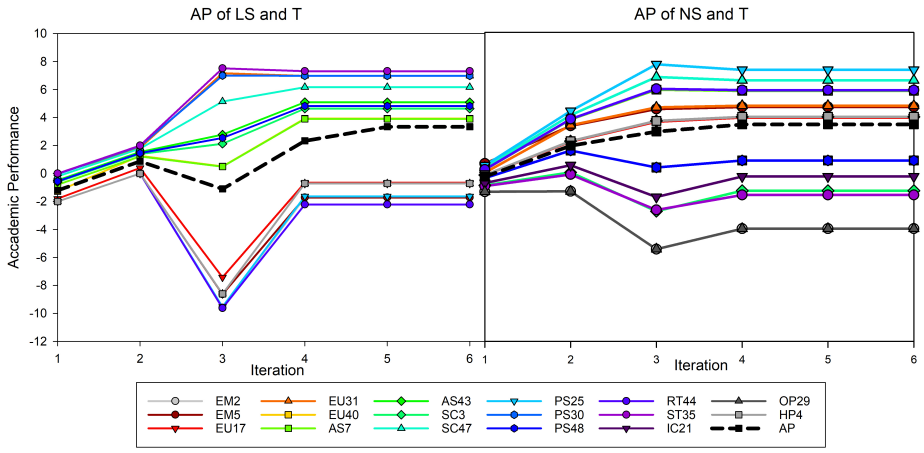


Fig. 1. Results for lagging students (LS) versus teachers (T) and non-lagging students (NS) versus teachers

their emotions. The variability between these groups implies different ways of perceiving, understanding and managing the emotions of others, represented in adaptation and social relationships ([38], [20], [37],[50]). Psychological support can influence lagging students' emotional intelligence, enabling them to cope with their feelings and emotions without it causing them any embarrassment.

2. Intrapersonal. The significant differences in self-concept between non-lagging students and lagging students is evident in item EM_5 "They care about what happens to people when facing changes". This is in line with [9] and [22], since social behaviour has an impact on self-concept and prosocial behaviour relationships in academic and family environments ([27]). High levels of self-concept ([48], [11], [26]) and social respect [60] imply defensive behaviours of a victim who has been attacked and justify procrastination in completing required activities.In this case, support for lagging students would be adequate since it would improve their abilities to perceive, understand and appreciate the feelings of others. For example, problems of empathy can be improved by forming a group of students from socially marginalized communities and supporting them in their emotional needs.
3. Adaptability. The differences between the non-lagging students and lagging students are significant and the items with the greatest Hamming differences are: PS_{25} "they don't give up until they've solved a problem"; PS_{48} "good at solving difficult things"; and RT_{44} "try to think of many possible solutions". As in [34], students from both groups do not react with the same intensity to the stimuli that university generates. Many adapt quickly, but there are others who do not, and they risk not fulfilling their academic goals. If university students adapt rapidly, this enables them to succeed academically, otherwise they are faced with overload generated by the activities themselves and demand. These potential problems trigger an inability to do the activities and tasks required, which causes stress and a negative attitude, giving rise to physical exhaustion and a low competence level ([5]). Without an appropriate adaptation their performance is compromised, resulting in abandonment, remaining longer

at university, and low completion efficiency ([53]).Support in mathematical reasoning could help any student; therefore, extracurricular courses are recommended as they will enable students to identify, define and put into practice effective solutions.

4. Stress management. The differences between non-lagging students and lagging students are expressed by the second greatest Hamming distance in the item "They easily get upset" (ST_{35}). Academic stress in university students is apparent in alarm, resistance and exhaustion ([52]). In the analysis between the groups, we can appreciate notable differences in how they value stress management. In this line, numerous studies have identified some prevailing elements, such as academic overload and evaluation, which are elements that generate stress and in the most critical cases lead to school abandonment ([45] ; [51]). The strategies oriented towards cognitive and behavioural efforts to resolve specific internal and external demands constantly change from first and second order. They are regarded as exacting or excessive for a person's resources ([16]) and are known as coping or stress management. To confirm whether a student is stressed or not, it is necessary to look into what worries them at that moment and if there are other types of symptoms. In line with [3], one indicator that a student is stressed is associated to unclear symptoms of illness: mental block, sleeping disorders, headaches, chronic fatigue, digestion problems, etc. Psychological support is fundamental for lagging students to be able to confront any adverse event in their life. Although there are a large number of stressful situations caused by school dynamics, stress management should be well oriented so that their health is not affected and strong emotions can be managed without incapacitating students, which will allow them to cope with them actively and positively.
5. State of mood. Item OP_{29} "feel confident about meeting goals" is another difference between the two groups, although according to the Hamming distance, it is weaker than the others discussed above. Despite this, it does have an impact on academic performance since negative results due to internal and external causes increase vulnerability to depression ([49], [14], [57]). Strategies directed at generating positive attitudes within school life are associated with motivation. For example, promoting cultural and sports events can generate positive optimism through the presence of other schoolmates, whereby their internal and external satisfaction is consolidated.

4. Conclusions

The analysed indicators included in the five cognitive dimensions (interpersonal, intrapersonal, adaptability, stress management and state of mind) are shown to influence students' academic outcome to a greater or lesser extent according to the rotated component matrix in the factor analysis. The expert assignment matrix is formed with the rotated component matrix, and along with the use of fuzzy cognitive maps and Hamming distance it is possible to obtain the critical variables that affect lagging students perfomance: problem solving, stress tolerance, reality testing, empathy, self-concept, happiness, emotional understanding, optimism.The values from the assignment matrices (weights) indicate the level of causality between cognitive-emotional variables. A positive value between empathy and happiness indicates that an increase in the first one causes the second to increase, and vice-versa. In this respect, different authors agree that state of mood and stress tolerance have a strong relation with academic performance ([21],[28], [31], [13], [39], [43]). Evidently, these two cognitive dimensions affect the others and academic performance.

On observing the greatest differences between lagging students and non-lagging students with respect to the cognitive elements, the greatest difference corresponds to: problem solving - "they don't give up until they've solved it; the next is associated with stress tolerance - "easily get upset"; this is followed by "try to think about many possible solutions" associated with reality testing; and "they understand people when there are changes" which is associated with empathy; then self-concept - "keep calm when they are upset about changes"; and lastly, associated with state of mind - "feel happy about the new challenges that face them at this time". The combinations of all these cognitive elements do not seem to synchronize with academic performance in lagging students, which remains in a first order state, while non-lagging students can synchronize this without any problem, thereby enabling an individual's transformation and second order change.

In summary, it can be asserted that lagging students have difficulties in problem solving; therefore, providing extracurricular courses to reinforce mathematical techniques will allow students to identify, define, and put into practice effective solutions. With respect to emotional problems, given that teachers and tutors are not trained to directly influence students' emotional intelligence, they should be dealt with through therapies which would allow students to understand their own feelings. Finally, the promotion of cultural and sports activities generate an atmosphere of positive optimism, facilitating interpersonal relations, strenghtening empathy and ultimately improving their happiness substantially.

References

1. Bar-On, R.: The bar-on model emotional-social intelligence (esi). Psicothema 18, 13–25 (2006)
2. Barraza, A., Silerio, J.: Academic stress in high school students: a comparative study. Investigacion Educativa Durangense 7, 48–67 (2007)
3. Barraza, M.: Psychometric validation of the unidimensional scale of student burnout. Intercontinental Journal of Psychology and Education 13(2), 51–74 (2011)
4. Bartunek, M., Moch, K.: First-order, second-order, and third-order change and organization development interventions: A cognitive approach. The journal of applied behavioural science 23(4), 483–500 (1987)
5. Caballero, C., Hederich, C., Palacio, J.: Academic burnout: delineation of the syndrome and factors associated with their emergence. Revista Latinoamericana de Psicología 42(1), 131–146 (2010)
6. Caballero, D., Abello, L.L., Palacios, S.: Relationship between burnout, academic performance, and satisfaction concerning study, in college students. Avances de Psic. Latinoamérica 25(2), 98–111 (2007)
7. Caldera, M., Pulido, C., Martínez, G.: Levels of stress and academic performance in psychology students of the university of los altos. Revista de Educacion y Desarrollo 7(1), 77–82 (2007)
8. Caldera, M., Reynoso, G., Gomez, C., Mora, G., Anaya, G.: Explanatory and predictive model of academic stress responses in secondary school students, ansiedad y estrés. Revista de Educacion y Desarrollo 23, 20–26 (2017)
9. Calvo, A., Gonzalez, R., Martorel, L.M.: Variables relacionadas con la conducta prosocial en la infancia y adolescencia: personalidad, autoconcepto y género, infancia y aprendizaje. Journal for the Study of Education and Development 24(1), 95–111 (2001)
10. Carlsson, C.: Knowledge formation in strategic management. Proceedings, IEEE. Computer Society Press, Los Alamitos (1994)
11. Carranza, R., Ramírez, A.: Procrastination and demographic characteristics associated with college students. apuntes universitarios. Revista de Investigacion 3(2), 95–108 (2013)

12. Carvalho, J., Tomé, J.: Rule based fuzzy cognitive maps and fuzzy cognitive maps - a comparative study. In: Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society. NAFIPS99, New York (1999)
13. Castro, M., Morales, R.: Classroom environments that promote learning from the perspective of school children. Review Electronic Educare 19(3), 1–32 (2015)
14. Cerezo, R.: Personality variables associated in bullying dynamics (aggressors vs. victims) (aggressors versus victims) in boys and girls from 10 to 15 years of age. Anales Psicología 17(1), 37–43 (2001)
15. Chen, M., Rubio, J., Ivanovic, M.: Guest editorial: Machine learning-based decision support systems in i o t systems. Comput. Sci. Inf. Syst. 20(2), i–iii (2023)
16. Chávez, R., Alcaraz, V., Ortiz, A.: Quantitative models with Fuzzy Logic in tangible and intangible assets. ECORFAN, Mexico (2018)
17. Cobo, A., Lopez, H., Herrera, V., F., H.: Scimat: A new science mapping analysis software tool. Journal of the American Society for Information Science and Technology 63(8), 1609–1630 (2012)
18. Curia, L., Lavalle, A.: Estrategias de decision en sistemas dinamicos- aplicando mapas cognitivos difusos aplicacion a un ejemplo socio-economico. JISTEM-Journal of Information Systems and Technology Management 8(3), 663–680 (2011)
19. Echeverría, R.: El observador y su mundo. Ed. Garnica, Buenos Aires (2009)
20. Extremera, N., Fernández-Berrocal, P.: La importancia de desarrollar la inteligencia emocional. Revista Iberoamericana de Educacion 19(3), 63–93 (2002)
21. Fernández-Berrocal, P., Extremera, N.: Emotional intelligence: A theoretical and empirical review of its first 15 years of history. Psicothema 18, 7–12 (2006)
22. Garaigordobi, M., Cruz, S., Pérez, J.: A correlational, cognitive and emotional factors of personality during adolescence. Studies in Psichology 24(1), 113–134 (2003)
23. Garcia, R., Pérez, G., Pérez, B.: Academic stress in first-year college students. Revista Latinoamericana de Psicología 44(2), 143–154 (2011)
24. Goleman, D.: La Inteligencia emocional. Zeta, Buenos Aires (1998)
25. González, A., Donolo, D., Rinaudo, C., Paoloni, V.: Relaciones entre motivacion, emoción y rendimiento académico en universitarios. Estudios de Psicología 32(2), 257–270 (2011)
26. González, B., Sánchez, P.: Can engagement buffer the harmful effects of academic procrastination? Acción psicológica 10(1), 117–134 (2013)
27. Gutiérrez, M., Clemente, A.: Autoconcepto y conducta prosocial en la adolescencia temprana: bases para la intervención. Revista de Psicología de la Educación 4(11), 39–48 (1993)
28. Hua, Y., Zhensheng, H., Yahui, P., Zhijian, W., Guanglong, X., Yanfang, X.: Deep semi-supervised learning with weight map for review helpfulness prediction. Comput. Sci. Inf. Syst. 18(4), 1159–1174 (2021)
29. Huff, S., Huff, O.: The change of strategy when companies change orientation. Oxford University, Mexico (2002)
30. Infante-Moro, A., Infante-Moro, J., Gallardo-Pérez, J.: Los mapas cognitivos difusos y su aplicación en la investigación de las ciencias sociales: estudio de sus principales problemáticas. Ediciones Universidad de Salamanca, Salamanca (2021)
31. Jimenez-Pena, M., Repetto, E.: Estado de la investigación en españa sobre inteligencia emocional en el ámbito educativo. Revista Electrónica de Investigación Psicoeducativa 15(6), 400–420 (2008)
32. Kosko, B.: Fuzzy cognitive maps. International Journal of Man-Machine Studies 24(1), 65–75 (1986)
33. Lamas, H.: Sobre el rendimiento escolar. Propósitos y Representaciones 3(1), 313–386 (2015)
34. Mamani-Ruiz, T.H.: Effect of adaptability in academic performance. Edu. Sup. Rev. Cient. Cepies 2(1), 38–44 (2017)
35. Martí, E.: Representar el mundo externamente. La construcción infantil de los sistemas externos de representación. Antonio Machado, Madrid (2003)

36. Martínez, E., Díaz, D.: Una aproximación psicosocial al estrés escolar. Educación y Educadores 10(2), 11–22 (2007)
37. Martínez, O.: La inteligencia afectiva. Teoría, práctica y programa. Madrid, CCS (2007)
38. Mayer, J., Salovey, P.: What is emotional intelligence? In P. Salovey and D. Sluyter (Eds). Emotional Development and Emotional Intelligence: Implications for Educators. Basic Books, New York (1997)
39. Melchor, G., López, E.: Motivación, comportamiento de los alumnos y rendimiento académico. infancia y aprendizaje. Journal for the Study of Education and Development 35(1), 61–72 (2012)
40. Naranjo, P.: Motivation: theoretical perspectives and some considerations of its importance in the field of educational. Revista Educación 33(2), 153–170 (2009)
41. Nelson, R.: Why do firms differ and how does it matter? Strategic Management Journal 12, 61–74 (1991)
42. Nelson, R., Winter, S.: An evolutionary theory of economic change. Cambridge University Press, Cambridge (1982)
43. Ordóñez, B.: Perspectives in ae competency-based education: Changing the traditional college degree power, policy, and practic. New Horizons in Adult Education and Human Resource Development 26(4), 47–53 (2014)
44. Peláez, C., Bowles, J.: Applying fuzzy cognitive maps knowledge-representation to failure modes effects analysis. In: IEEE Proceedings Annual Reliability and Maintainability Symposium. pp. 450–456. Washington DC (1995)
45. Pozos, R., Preciado, S., Plascencia, C., Acosta, F., Aguilera, V.: Academic stress and physical, psychological and behavioural factors in mexican public university students. Ansiedad y Estrés 21(1), 35–42 (2015)
46. Probst, G., R.S., Romhardt, K.: Administre el conocimiento. Pearson Educación, México (2001)
47. Reyes, G., Ibarra, Z., Torres, L., S., R.: Stress as a risk factor in health: differential analysis between public and private university teachers. Revista Digital Universitaria 13(7), 1–13 (2012)
48. Salmivalli, C., Lappalainen, M., Lagerspetz, K.: Stability and change of behavior in connection with bullying in schools: A two-year follow-up. Aggressive Behavior 24(3), 205–218 (1998)
49. Sanjuán, P., Palomares, A.: Analysis of attributional style attributional style in students with depressive mood. Estudios de Psicología: Studies in Psychology 19(61), 25–33 (1998)
50. Santander, T., Gaeta, G., Martínez, O.: Impact of emotional regulation in the classroom: a study with spanish teachers. Revista Interuniversitaria de Formación del Profesorado 95(34.2), 225–246 (2020)
51. Suárez, M., Díaz, S.: Academic stress, desertion, and retention strategies for students in higher education. Revista Salud Pública 17(2), 300–313 (2015)
52. Suárez, O., Hurtado, A., Lizarazo, O.: Underlying variables for academic stress and student motivation in higher education amid the covid-19 pandemic. Revi. Fac. Cienc. Tecnol. 51, 37–56 (2022)
53. Tortajada, J.F.: ¿hay relación entre la inteligencia, la adaptabilidad escolar y familiar y las calificaciones obtenidas? In: III Congreso Internacional Virtual de Educación. pp. 1–11 (2003)
54. Ugarriza, C., Pajares, D.: Adaptación y estandarización del inventario de inteligencia emocional de Bar-On ICE: NA, en niños adolescentes. Amigo, Lima (2001)
55. Ugarriza, C., Pajares, D.: Construct validity of the bar-on emotional quotient inventory: youth version. Persona (8), 11–58 (2005)
56. Vargas, J.: La inteligencia emocional en la educación. Groppe Libros, México (2013)
57. Velasteguí, H., Mayorga, L.: Mood, anxiety, and depression in medical college students during the covid-19 health crisis. Revista Psicología UNEMI 5(9), 10–20 (2021)
58. Wu, B., Zhang, T., Chen, Y.: Content -only attention network for social recomendation. Comput. Sci. Inf. Syst. 20(2), 609–629 (2023)

59. Yam, Z., Hongle, D., Lin, Z., Jianhua, Z.: Personalization exercise recomendation framework base on knowledge concept graph. Comput. Sci. Inf. Syst. 20(2), 857–878 (2023)
60. Yelsma, P., Yelsma, J.: Self-esteem and social respect within the high school. Journal of Social Psychology 138(4), 431–441 (1998)

José Manuel Brotons Martínez is a member of the Economic and Financial Studies Department of Miguel Hernández University in Elche (Spain). He has an 11 h-index (Web of Science), with 51 documents listed in Web of Science and 251 citations. He is expert in financial valuation and economic analysis applied to the agriculture, and company financial valuation using new methodologies such as the innovations in fuzzy logic and the analysis of the quality system in the enterprises.

Jesús Ricardo Ramos Sánchez is a research professor graduated from the Universidad Michoacana de San Nicolás de Hidalgo, currently serves as rector of the Universidad Politécnica de Altamira in Tamaulipas, his lines of research address energy efficiency, sustainability and the impact of regional development through the technological relationship in energy - human being - economy. He is currently a member of the National System of Researchers of Conahcyt and his work deals with techniques that document the intervention of energy in the evolution of society.

Rubén Chávez Rivera is a professor at the Faculty of Chemical Pharmacobiology-Biotechnology, an active external member of the postgraduate programme in Regional Development Sciences (ININEE) of the UMSNH (Mexico). He is a member of the National System of Researchers of Conahcyt. The lines of research are focused on sustainability in urban-industrial-energy processes for regional development; Analysis under uncertainty (fuzzy logic). It has 4 h-index and has 2 in i10 indexes (Web of Science) and 104 citations (School Academy), 6 books; He has participated in 6 agro-industrial projects.

Received: February 13, 2023; Accepted: September 30, 2023.