ComSIS Computer Science and Information Systems

Volume 23, Number 2, April 2026

Contents

ISSN: 2406-1018 (Online)

Computer Science and Information Systems

Vol 23, Number 2, April 2026

ComSIS

Computer Science and Information Systems

Published by ComSIS Consortium

Volume 23, Number 2
April 2026

ComSIS is an international journal published by the ComSIS Consortium

ComSIS Editorial Office:
University of Belgrade, Faculty of Organizational Sciences,
Jove Ilića 154, 11000 Belgrade, Serbia
Phone: +381 11 3950 866; +381 66 8896 155
www.comsis.org; Email: comsis@fon.bg.ac.rs

Volume 23, Number 2, 2026
Belgrade

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/

ComSIS

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 2024 two-year impact factor 1.8,
- Computer Science Bibliography, University of Trier (DBLP),
- EMBASE (Elsevier),
- Scopus (Elsevier),
- Ex Libris Knowledge Center by Clarivate,
- 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) 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 30 (using the appropriate LaTeX template).

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

Copyright and Use Agreement

Computer Science and Information Systems

Volume 23, Number 2, April 2026

CONTENTS

Computer Science and Information Systems 23(2): i–iii https://doi.org/10.2298/CSIS260200iI

Editorial

Mirjana Ivanović[1], Miloš Radovanović[1], Vladimir Kurbalija[1], and Ivan Luković[2]

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

[2]University of Belgrade, Faculty of Organizational Sciences
Belgrade, Serbia
ivan.lukovic@fon.bg.ac.rs

Volume 23, Issue 2 of the Computer Science and Information Systems journal is comprised of 10 regular articles. Once again, we express our gratitude for the hard work and enthusiasm of all authors, reviewers, and guest editors, without whom the current issue and the publication of the journal itself would not be possible.

The first article, "Joint Heart Sound Denoising Using DTCWT and Adaptive Sparsity-Assisted Signal Smoothing Algorithm," by Jianqiang Hu et al., proposes a joint heart sound signal denoising method which employs the Dual-Tree Complex Wavelet Transform (DTCWT) and Adaptive Sparsity-Assisted Signal Smoothing (ASASS) algorithms. The signal is first decomposed by DTCWT to obtain the multi-scale feature representation. Subsequently, ASASS suppresses pseudo-Gibbs artifacts around signal boundaries of DTCWT while implementing adaptive thresholding strategies to maximize the Signal-to-Noise Ratio (SNR).

In the second article, "Transformer Substation Network Disconnection Prediction via Semantic Reasoning with Causal Modeling," Jie Ren et al. propose a causal-driven disconnection prediction framework that integrates semantic reasoning with structural causal modeling, in order to address the limitations of existing correlation-based approaches for network disconnection prediction. Heterogeneous monitoring data are unified into structured events, and a causal graph is constructed to capture semantic, temporal, and topological dependencies among event sequences. A pluggable inference module is further introduced, combining path analysis, structural causal models, and counterfactual reasoning to enable root-cause identification and high-confidence risk prediction.

Fangyuan Xiong et al., in "YOLO-BDM: An Improved Ship Detection Algorithm Based on YOLOv11n," introduce YOLO-BDM, an improved ship detector from Synthetic Aperture Radar (SAR) data, based on YOLOv11. The Diverse Branch Block (DBB) is introduced into the backbone to enhance feature representation through multi-branch training and reparameterized inference. A Multi-scale Contextual Attention (MCA) mechanism is integrated into the backbone and neck to strengthen multi-scale semantic modeling and background discrimination. Additionally, a four-layer Bidirectional Feature Pyramid Network (BiFPN) is employed for efficient multi-scale feature fusion.

"Enhanced ROCKET for the Automated Detection of Epileptic Tonic-Clonic Seizures Using Accelerometer Data," by Krisztian Buza et al. develop an algorithm which may contribute to recognition of tonic-clonic epileptic seizure based on accelerometer data that can be collected from mobile and wearable devices, framing this task as a multivariate time-series classification problem. The state-of-the-art Random Convolutional Kernel Transform (ROCKET) machine-learning approach is enhanced by replacing standard con-

volution with dynamic convolution. Since dynamic convolution was originally defined for univariate time series, this work extends it to multivariate time series.

The article "An Incremental Network Information Extraction Technology for Web Topic Detection Based on Big Data," authored by Mo Chen, proposes an incremental network information extraction approach for Web topic detection, where the algorithm of theme similarity measurement for incremental network information extraction can extract Web instances related to theme, and calculate importance of Web instances related to the theme. Furthermore, the designed algorithm of incremental instance extraction for Web topic detection can analyze Pattern and BasePattern according to the extracted Web instance URL, conduct segmentation for Web instance title and text content, and extract keywords which are capable of describing the Web topic.

Lijue Liu et al., in their article "Development and Validation of a Few-Shot Rapid Screening Model for Gastrointestinal Cancers Using AGI Large Vision Models," develop a screening framework leveraging a large vision model for coarse-grained classification of gastric and colorectal tissues, to address the limitation of existing deep learning models in digital pathology which typically require extensive labeled data and show limited generalization across organs. The model was evaluated on multicenter cohorts and under limited-data conditions, and demonstrated excellent performance on both internal and external test sets.

"A Pilot Study of Multi-Method Evaluation of Machine Translation in Macedonian," by Jana Kuzmanova and Katerina Zdravkova, offers a linguistic evaluation of six machine translation systems: GPT-4o, GPT-5, Gemini 2.5 Flash, Google Translate, Microsoft Translator, and NLLB-600M applied to the translation of a short excerpt of Orwell's "1984" into Macedonian. The analysis consisted of three interconnected experiments: manual annotation of translation errors and comparison with human output, evaluation using eight popular machine translation (MT) metrics, and sentence-level similarity analysis via three different similarity/distance measures. The findings underscore the importance of combining manual and metric-based evaluation to fully understand MT quality, particularly in low-resource settings.

In "HAQCCN: A Hybrid Quantum–Classical Convolutional Network with Asymmetric Kernels for Remote Sensing Image Classification," Lianghai Chen et al. address the limitations of conventional convolutional neural networks (CNNs) concerning computational capacity and struggling to efficiently process the rapidly growing volume of remote sensing data by proposing HAQCCN (Hybrid Asymmetric Quantum–Classical Convolutional Network), a novel hybrid architecture that integrates quantum computation into the classical convolutional framework through asymmetric quantum convolutional circuits.

"FPCA: A Fully Constant-Length and Policy Updating Cross Data Domain Access Control for Cloud-Edge Collaborative Environment," by Shiwen Zhang et al., is an approach which addresses the challenges of challenges of low data access control efficiency and a lack of dynamism in such environments. This is achieved by proposing three algorithms: the Multi Data Domains Key Generation (MDKG) algorithm maintains constant secret key length and enables cross data domain access without attribute conversion, the Constant-Length Ciphertext Encryption (CLCE) algorithm maintains constant ciphertext length, reducing decryption overhead, and the Access Policy Update (APU) algorithm updates the access policy with constant-length update message and low computational complexity.

Finally, in their article entitled “A Framework for the Automated Thematic Annotation of Open Government Data,” Abdul Aziz et al. tackle the limitation of open data portals that often lack consistent and scalable mechanisms for thematic annotation, limiting dataset discoverability, by proposing a framework for automated thematic classification of open government data which integrates (1) thematic annotation quality assessment, (2) supervised machine learning models trained on annotated metadata corpora, and (3) embedding-based semantic similarity methods for theme assignment in the absence of reliable annotations.

Computer Science and Information Systems 23(2):687–706 https://doi.org/10.2298/CSIS250828017H

Joint Heart Sound Denoising Using DTCWT and Adaptive Sparsity-assisted Signal Smoothing Algorithm

Jianqiang Hu[1,2], Dafeng Shen[1], Lin Chen[1], Yu Chen[1], Shigen Shen[3], and Yan Che[2]

[1] School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, P. R. China
hujianqiang@tsinghua.org.cn
shendafeng25@gmail.com
275914270c@gmail.com
631894828@qq.com

[2] Engineering Research Center for Big Data Application in Private Health Medicine of Fujian Universities, Putian University, Putian, Fujian 351100, P. R. China
ptucy07@126.com (corresponding author)

[3] School of Information Engineering, Huzhou University, Huzhou 313000, P. R. China
shigens@zjhu.edu.cn

Abstract. Various unwanted and unavoidable noises corrupt Heart Sound Signals (HSSs). It is strongly required to suppress respiratory sound and ambient noise due to significant reduction of clarity and interpretation of HSSs. In this paper, we propose a joint heart sound denoising using Dual-Tree Complex Wavelet Transform (DTCWT) and Adaptive Sparsity-assisted Signal Smoothing (ASASS) algorithm. In this research, the signal is first decomposed by DTCWT to obtain the multi-scale feature representation of the signal. Subsequently, ASASS suppresses pseudo-Gibbs artifacts around signal boundaries of DTCWT while implementing adaptive thresholding strategies to maximize the Signal-to-Noise Ratio (SNR). Experimental validation on the PhysioNet/CinC 2016 database and Open Access Heart Sound Dataset (OAHS Dataset) demonstrates that the proposed method significantly outperforms existing techniques. Under conditions involving Gaussian white noise (GWN) SNR of 0 dB, the proposed method achieves an SNR of 9.01 dB and a Root Mean Square Error (RMSE) of 0.032, outperforming standalone DTCWT and multiple existing models.

Keywords: Heart Sound Signals, Denoising, Adaptive Sparsity-Assisted Signal Smoothing, Dual-Tree Complex Wavelet Transform.

1. Introduction

Internet of Things (IoT)-based digital stethoscopes are rapidly entering the field of family healthcare monitoring [1]. Heart sounds are collected by various digital stethoscopes, which contain a great number of biomedical signals of cardiac activity. In practice, heart sound is corrupted by unavoidable entities that easily lead to the clinical misinterpretations. Ambient noise, respiratory sound and even signals from various complex environments overlap with the heart sound signals (HSSs), making their quality degraded. In particular, the spectrum of respiratory sound overlaps perfectly with that of heart sound. Besides, electronic interferences and recording artifacts corrupt HSSs to a certain extent.

HSS is a non-linear, non-stationary weak signal with a low frequency and weak amplitude. Obviously, heavy noises bring about the poor the diagnostic interpretations. In order to obtain accurate diagnostic performance, it is necessary to effectively suppress noise contaminants from HSSs. As a result, HSS denoising has become a primary challenge in research and clinical practice.

Over the past few decades, the researchers have developed various effective denoising methods to maximize the preservation of critical components of HSSs while eliminating noises. First, multiple time-domain methods have been employed for noise removal, including conventional filters and auto-correlation techniques. For instance, Butterworth band-pass filters and Finite Impulse Response (FIR) filters can eliminate noises outside the frequency range of HSS [2]. Second, various frequency-domain methods such as Wavelet transform (WT) [3], Fourier transform [4], Empirical Wavelet Transform (EWT) [5], Singular Value Decomposition (SVD) [6] and Dual-Tree Complex Wavelet Transform (DTCWT) [7] have been utilized to for HSSs denoising. Among these, WT-based denoising relies on adaptive thresholding of wavelet coefficients, enabling simultaneous noise suppression and preservation of singularity points in HSSs. Compared to WT, DTCWT offers advantages including time-invariance, superior reconstruction capability, and absence of aliasing effects-exhibiting excellent anti-aliasing performance and near-shift invariance. However, a common challenge across all WT-based denoising methods is the introduction of pseudo-Gibbs artifacts at the singularities. This occurs primarily because the amplitude of local oscillations decreases near signal discontinuities (singularity points). Finally, machine learning and deep learning have been applied to remove noise contaminations from HSSs. Machine learning-based denoising, such as particle Swarm optimization (PSO) [8] and Twin Support Vector Machine [9], demonstrates limited generalization capability when confronted with diverse noise types. While deep learning possesses stronger feature representation capabilities than machine learning, it struggles with handling unbalanced datasets. For instance, loss functions of fully convolutional networks (FCNs) [10], denoising convolutional neural network (DnCNN) [11] and Enhancement Adversarial Convolutional Neural Network (EACNN) [12], remain challenging in practice due to the requirement for maximizing the preservation of subtle details within HSS fluctuations.

In this work, we develop a joint denoising method for HSSs based on DTCWT and Adaptive Sparsity-Assisted Signal Smoothing (ASASS) algorithm, called DTCWT+ASASS. While DTCWT exhibits excellent near-shift invariance, it may still be insufficient for effectively capturing detailed components of HSSs. By integrating ASASS with DTCWT, the proposed approach simultaneously processes low-pass recursive filters and zero-phase non-causal high-pass filters as banded matrices to denoise diverse signals. Notably, the zero-phase characteristic eliminates phase distortion caused by causal Linear Time-Invariant (LTI) filters, thereby preserving the original morphology of HSSs.

The principal contributions of this study are summarized as follows:

1. We propose a joint HSS denoising method using DTCWT and ASASS. This method prevents distortion in denoised HSS while suppressing pseudo-Gibbs artifacts at signal boundaries in DTCWT, consequently enhancing denoising adaptability.
2. We introduce an ASASS algorithm incorporating an adaptive thresholding mechanism. This algorithm smoothens HSSs and suppresses unwanted overlapping noise frequencies to recover noise-free signals. Crucially, its adaptive mechanism dynami-

cally adjusts thresholding strategies based on sub-band energy distribution and directional correlations, achieving optimal balance between noise suppression and feature preservation.

3. To validate the effectiveness of the proposed method, multiple experiments were performed on the PhysioNet/CinC 2016 Challenge database [13] and the Open Access Heart Sound Dataset (OAHS Dataset) [11]. The experiments demonstrate that our proposed method achieves significant improvements over state-of-the-art methods. Under conditions involving Gaussian white noise (GWN) Signal-to-Noise Ratio (SNR) of 0 dB, the proposed method achieves an SNR of 9.01 dB and a Root Mean Square Error (RMSE) of 0.032, outperforming standalone DTCWT and multiple existing models.

The paper is organized as follows. The next section introduces a quick survey of related work of HSS denoising. Section 3 focuses on the systematic processes of a joint HSSs denoising based on DTCWT and ASASS algorithm. The content of Section 4 is the experiment and the results are presented. Finally, a comprehensive summary of the entire text in Section 5.

2. Related work

HSS is susceptible to various types of noise during acquisition, such as Electromagnetic Interference (EMI) from the surrounding environment, power frequency interference, bioelectrical interference from the body, breath sounds, and Lung sounds (LS). Denoising methods have been explored across multiple branches of biomedical engineering, including Electrocardiography (ECG), Electroencephalography (EEG), respiratory sounds, and heart sounds, allows to extract from it the maximum amount of efficient and meaningful information.

(i) ECG Denoising. Recent research indicates that the filtering stage should employ an architecture based on Adaptive Filters (AF). Adaptive algorithms enable real-time dynamic optimization by adjusting filter parameters to adapt to input ECG signal characteristics [14]. Additionally, the conventional low-pass and high-pass filters in wavelet transforms have been replaced by fractional-order wavelets. Studies demonstrate through simulations that fractional-order wavelets outperform traditional wavelets in ECG denoising, with their efficacy dependent on wavelet decomposition level selection and coefficient thresholding strategies [15][16]. Deepak H. A. et al. proposes an adaptive thresholding technique combining Empirical Mode Decomposition (EMD) and DTCWT [17]. Traditional Discrete Wavelet Transform (DWT) is prone to induce Gibbs oscillations and frequency aliasing, whereas DTCWT significantly mitigates these issues through enhanced time-frequency decomposition properties [18]. Separate research systematically evaluates the impact of threshold selection, algorithms, and distribution functions on denoising performance within DTCWT frameworks[19]. To address the issue of multi-category noise contamination in real-time acquisition, the Adversarial Denoising Convolutional Network (ADnCNN) was proposed, which enhances model robustness through adversarial training [20]. Enhan Liu et al. introduced the method of Noise Prediction-based ECG Denoising method (NPED) [21]. This method effectively solves noise problems in grayscale images scanned from Paper-based ECGs in hospitals. Furthermore, it overcomes the limitations of traditional waveform redrawing approaches by performing direct denoising to prevent

waveform distortion. For dynamic interferences such as Baseline Wander (BW), Electrode Motion (EM), and Muscular Artifact (MA), the ECG denoising diffusion model (EDDM) was developed to establish a generative diffusion model for signal reconstruction [22]. CNN-SWT integrates the convolutional kernel constraints and architecture of Stationary Wavelet Transform (SWT) into a convolutional neural network (CNN), significantly enhancing the learning efficiency for denoising both linear and nonlinear time-frequency features in ECG signals [23].

(ii) Respiratory sound & EEG Denoising. LS signals are severely contaminated by background noise from multiple sources. Conventional denoising methods may exhibit limited effectiveness due to the non-stationary characteristics of LS and its spectral overlap with various noise sources. A joint denoising approach based on the Butterworth bandpass filter and Sparsity-Assisted Signal Smoothing (SASS) algorithm is proposed, which significantly enhances the signal-to-noise ratio of LS through frequency-domain filtering and sparse-constrained signal reconstruction [24] [25]. An adaptive denoising technique utilizing Discrete Wavelet Transform and Artificial Neural Network (DWT-ANN) has been developed. This method integrates the multi-resolution analysis capability of DWT with the nonlinear adaptive filtering properties of ANN to achieve refined purification of LS signals a noisy environment [26]. In EEG signal processing, Variational Mode Decomposition-based Blind Source Separation (VMD-BSS) and DWT-BSS effectively isolate physiological artifacts while preserving essential neural information [27]. Additionally, researchers have combined EMD with DTCWT to achieve high-fidelity EEG denoising through a two-stage processing framework [28]. Furthermore, an EMD-Hurst analysis combined with spectral subtraction has been implemented to optimize LS denoising through mode selection and energy correction [29]. Besides, a novel deep encoder-decoder-based denoising architecture (LU-Net) has been presented to suppress ambient and internal lung sound noises [11].

(iii) PCG & HSSs Denoising. Digitally acquired heart sound signals via stethoscopes are often distorted by environmental and physiological noise interference, altering their key distinctive characteristics. To address superimposed noise, Xiahou S. et al. employs variational mode decomposition (VMD) for hierarchical filtering [30] [3]. Linear filters have been to utilized to eliminate distortion and interference in fetal phonocardiogram (fPCG) signals [31]. Additionally, integrating Hilbert envelope and homomorphic envelope techniques, combined with reusing pretrained CNN filters for denoising, significantly enhances data utilization efficiency in implantable devices [32]. Incorporating DTCWT with an adaptive neuro-fuzzy inference system (ANFIS) classifier enables simultaneous optimization of signal enhancement and classification [33]. Another study establishes an objective threshold calculation method for DTCWT denoising, optimizing the noise reduction process by quantifying edge preservation and noise elimination as objective functions [34]. In deep learning-based heart sound denoising, Duggan D. et al. proposes a fully convolutional network (FCN) denoising model based on the Spleeter U-Net architecture [10]. For generative adversarial network (GAN)-based models generating normal heart sounds, incorporating EWT denoising effectively reduces GAN training cycles and computational costs [35].

3. Methodology

In this paper, a joint HSSs denoising using DTCWT and ASASS algorithm is adopted. Firstly, the original HSS is denoised by Butterworth filter, and then the denoised HSS $X(t)$ is sent to 4-layer DTCWT for decomposition. And then, sparse-assisted signal smoothing (SASS) algorithm, including sparse-assisted signal extraction layer and signal smoothing layer, is used to extract and smooth the decomposed HSS, and adaptive parameter algorithm is set to adjust the SASS parameters. Finally, it is combined with the decomposed signal and the inverse DTCWT transform is used to reconstruct the signal, so as to better realize the denoising of heart sound signal $\hat{X}(t)$. The architecture of our proposed method is shown in Fig. 1.

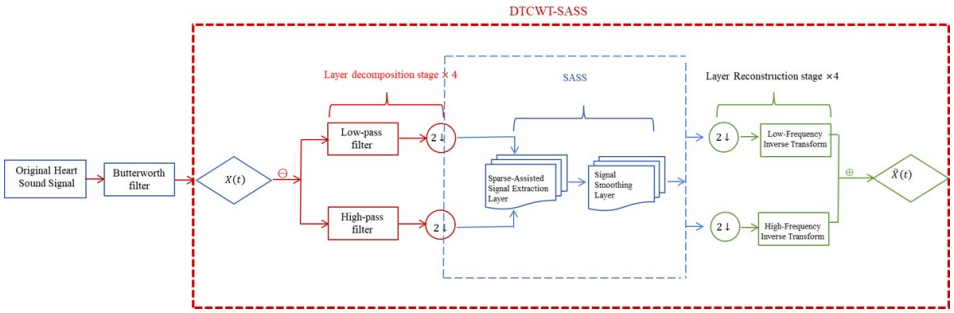


Fig. 1. The architecture of the joint HSSs denoising using DTCWT and ASASS

3.1. DTCWT

DTCWT is a wavelet transform method possessing time-frequency localization and multiscale characteristics. Compared to traditional single wavelet transforms, it can more effectively capture the time-frequency information of signals. DTCWT decomposes a signal multiscale using a set of orthogonal wavelet basis functions, where the wavelet basis functions at each scale consist of a pair of low-pass and high-pass filters. Unlike traditional wavelet transforms, DTCWT employs two parallel wavelet transform trees, processing the real and imaginary parts of the heart sound signal separately. This structure better preserves the signal phase information and provides superior time-frequency resolution. Fig. 2 illustrates the typical structure of DTCWT, encompassing both decomposition and reconstruction stages.

In the decomposition stage, the input signal is fed into two signal processing trees, referred to as the real tree and the imaginary tree. Collectively, these two trees are termed the Complex Wavelet Transform. Each tree can be viewed as a Filter Bank (FB) tree, where the upper tree (Top Tree) contains the real part's low-pass filter (h_0) and high-pass filter (h_1). The lower tree (Bottom Tree) contains the imaginary part's low-pass filter (g_0) and high-pass filter (g_1), respectively.

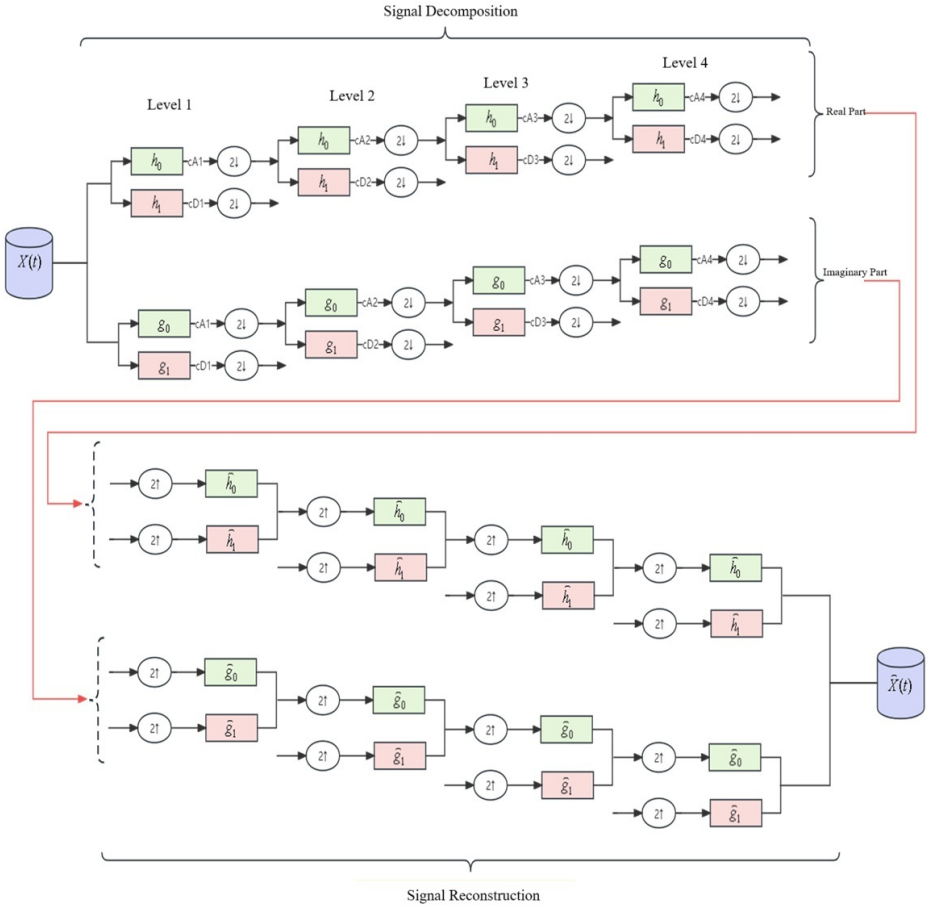


Fig. 2. The architecture of DTCWT

At the Level 1 decomposition, the real part is split into low-frequency information (h_0) and high-frequency information (h_1). The low-frequency coefficients are termed approximation coefficients (cA_1), while the high-frequency coefficients are called detail coefficients (cD_1). This process is repeated at Level 2, where the previous low-frequency component, cA_1, is sub-decomposed into the low-frequency cA_2 and the high-frequency cD_2.

The decomposition progresses iteratively, concluding at Level 4. Consequently, four approximation coefficients (cA_1–cA_4) and four detail coefficients (cD_1–cD_4) are generated for the real part's Discrete Wavelet Transform (DWT). Similarly, during the decomposition stage, this identical process is applied to the imaginary part. This yields a parallel set of four approximation coefficients (cA_1–cA_4) and four detail coefficients (cD_1–cD_4) for the imaginary component. This dual-tree structure effectively mitigates limitations inherent in the standard DWT, such as spectral aliasing and lack of shift-invariance. This

process is repeated at Level 2, where the previous low-frequency component, cA_1, is sub-decomposed into the low-frequency cA_2 and the high-frequency cD_2. The decomposition progresses iteratively, concluding at Level 4. Consequently, four approximation coefficients (cA_1–cA_4) and four detail coefficients (cD_1–cD_4) are generated for the real part's DWT.

In DTCWT, the given signal is expressed in terms of shifted and dilated forms of the mother wavelet function $\psi_{j,n}(k)$ and a scaling function $\varphi_{j,n}(k)$. cA_j denotes the approximation coefficient at level j and formally expressed as

$$cA_j(n) = \sum_k x(k)\,\psi_{j,n}(k) \tag{1}$$

$$\psi_{j,n}(k) = \frac{1}{\sqrt{2j}}\psi\left(\frac{n - k \bullet 2j}{2j}\right) \tag{2}$$

where $\psi_{j,n}(k)$ represents the mother wavelet function, k is the shift parameter, and j indicates the DTCWT decomposition level.

Similarly, detail coefficients $cD_j(n)$ capturing high-frequency components at decomposition level j and formally expressed as

$$cD_j(n) = \sum_k x(k)\,\varphi(n-k) \tag{3}$$

where $\varphi(n-k)$ represents the mother wavelet function and defined as $\varphi(n) = (-1)^n \psi(N-1-n)$, and k is the shift parameter.

3.2. Sparse-assisted Signal Extraction

The fundamental principle of the SASS algorithm leverages the concept of sparse representation, combining linear time-invariant (LTI) filtering with sparse optimization principles to model signals as the sum of a piecewise-smooth component and a low-pass component. By solving the sparse optimization problem, it identifies the sparsest vector that optimally represents the original signal, revealing its primary structure and essential characteristics. The algorithm operates similarly to wavelet denoising but performs denoising via sparse optimization, thereby avoiding the pseudo-Gibbs phenomena commonly encountered in DTCWT.

Let $y(n)$ denote the noisy HSS, which can be represented as

$$y(n) = s_1(n) + s_2(n) + \omega(n), n \in Z \tag{4}$$

where $s_1(n)$ represents a low-frequency signal, $s_2(n)$ has the sparse derivative of M-order, and $\omega(n)$ is additive white Gaussian noise (AWGN). The signal y can be written as

$$y = AQ^{-1}P^TPt + \alpha A^{-1}Q^TQ_1\mu \tag{5}$$

$$C(\mu) = \frac{1}{2}\left|\left|t - A^{-1}P^TPt - \alpha A^{-1}Q^TQ_1\mu\right|\right|_2^2 + \lambda||\mu||_1 \tag{6}$$

where μ is sparse and is determined by the cost function $C(\mu)$, P and Q are the banded Toeplitz matrices [24].

To achieve enhanced denoising performance, this paper integrates the SASS algorithm with DTCWT. After sub-band decomposition, sparse-assisted signals are extracted from the sub-bands to obtain superior signal characteristics. Through this integration, the sparse-assisted signal is ultimately extracted by solving the following optimization problem

$$s\left(t\right) = arg_{s} \min \left|\left|cD - s\right|\right|_{2}^{2} + \lambda \left|\left|s\right|\right|_{1} \tag{7}$$

Here, cD denotes the detail coefficients obtained via DTCWT decomposition, and s represents the sparse-assisted signal. The final smoothed signal $\hat{X}(t)$ is obtained by weighting and summing the sparse-assisted signal with the decomposed signal as

$$\hat{x}\left(t\right) = x\left(t\right) + \gamma s(t) \tag{8}$$

where γ is a smoothing parameter that controls the influence of the sparse-assisted signal during the smoothing process. However, as the four-level DTCWT decomposition through high-pass and low-pass filters yields sub-bands with varying scales, orientations, and threshold requirements, a fixed SASS algorithm cannot uniformly process these heterogeneous sub-bands. Consequently, this paper proposes an adaptive threshold of SASS to further process these sub-band signals. The adaptive strategies dynamically adjust the threshold based on sub-band energy distribution and directional correlations, achieving a balance between noise suppression and feature preservation.

3.3. An Adaptive Threshold of SASS Based on Sub-band Energy Distribution

First, the sub-band signals obtained from a four-level DTCWT decomposition are grouped. For each sub-band signal s_i, its energy E_i is calculated by summing the squares of its values and can be represented as

$$E_i = \sum_{n=1}^{N} \left(s_i(n)\right)^2 \tag{9}$$

where N is the length of the sub-band signal.

Subsequently, we need to set an initial threshold, adopting the standard deviation threshold method. This is a simple and intuitive approach that uses the standard deviation of the signal as the threshold. It is typically assumed that the noise in the signal follows a Gaussian distribution. Therefore, the standard deviation of the signal can be used to estimate the noise level. The threshold can be chosen as a multiple of the standard deviation with the expression as follows $T = k\sigma$, where T is the threshold, k is the multiplier coefficient, and σ is the standard deviation of the signal.

Finally, by comparing the sub-band energy E_i with the threshold T, the sub-band signals are divided into a high-energy group and a low-energy group. For each group of sub-band signals, we dynamically adjust the parameters of the SASS method based on their characteristics. As for optimal denoising performance, the SNR of the sub-band signal s_i is used as the metric to design an adaptive algorithm for dynamically tuning the SASS parameters. The SNR is expressed as

$$SNR = 10 \lg \frac{\sum_{n=1}^{N} (x(n))^2}{\sum_{n=1}^{N} (x(n) - \hat{x}(n))^2} \tag{10}$$

Here, N denotes the number of HSS samples, $x(n)$ denotes the original HSS, $\hat{x}(n)$ denotes the denoised HSS.

We continue to set the signal threshold and define the objective function F as the relative change rate of the SNR, specifically the ratio of the post-processing SNR to the pre-processing SNR. Its formula is as

$$F = \frac{SNR_{post} - SNR_{pre}}{SNR_{pre}} \tag{11}$$

where SNR_{post} is the processed SNR, SNR_{pre} is the original SNR.

For each group of sub-band signals, parameters are iteratively updated using the gradient descent method. The parameter update formula is as

$$\theta_{n+1} = \theta_n - \alpha \nabla U(\theta_n) \tag{12}$$

where θ_{n+1} denotes the updated parameter value, θ_n represents the current parameter value, and α is the gradient of the objective function and controlling the step size of each iteration. $U(\theta)$ is the objective function, characterizing its dependence on parameters θ. $\nabla U(\theta)$ is the gradient of $U(\theta)$ with respect to θ, indicating the magnitude and direction of change in the objective function at the current parameter value. We set a convergence criterion to stop the iteration when the change in the objective function is less than a predefined threshold, as illustrated by the following

$$|U(\theta_{n+1}) - U(\theta_n)| < \varepsilon \tag{13}$$

where ε denotes the set convergence decision threshold. The adjusted parameters and the processed sub-band signals are returned. Ultimately, through this adaptive algorithm, we dynamically adjust the parameters of the SASS method based on the characteristics of the sub-band signals, maximizing the relative change rate of the SNR to enhance the denoising effect.

4. Experiment

4.1. Data Resources

Publicly available heart sound datasets are used to evaluate the performance of our proposed method. (i) PhysioNet Computing in Cardiology Challenge 2016 dataset. All recorded data in this dataset are categorized as normal or pathological (abnormal) samples. We select five imbalanced recording categories (Training-A through E), comprising a total of 3,126 heart sound recordings, with uneven distribution of cardiac condition severity. Additionally, each category contains extraneous emergency noises, such as uncontrolled ambient sounds. We merge Category C and Category D data due to their limited quantity of heart sound signals and excludes significantly distorted signal recordings. 2,735

heart sound recordings are extracted, including 546 pathological recordings and 2,189 normal recordings, representing approximately 87.5% of the entire dataset. (ii) Open Access Heart Sound (OAHS) Dataset. OAHS Dataset is a publicly available cardiac audio database providing researchers with noise-free heart sound data. This dataset contains 1,000 heart sound samples, and the samples are divided into five distinct categories based on cardiac pathological characteristics: Normal (N), Aortic Stenosis (AS), Mitral Regurgitation (MR), Mitral Stenosis (MS), and Mitral Valve Prolapse (MVP). Each category comprises 200 independent representative recordings.

4.2. Performance Metrics

In order to evaluate the entire SNR performance, the global error measures include exploited E(SNR), var(SNR) and PE(SNR). E(SNR), var(SNR) and PE(SNR) are the mean value, variance and Percentage error of SNR. Besides, RMSE is another metrics that popular in HSSs analysis. RMSE measures the similarity between the denoised HSS and the original clean HSS by calculating the deviation between them. A smaller RMSE value indicates a smaller difference between the denoised signal and the original clean signal, meaning better denoising performance. RMSE is expressed as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \tag{14}$$

Here, x_i is the i^{th} sample value of the original clean HSS, $\hat{x}_i$ is the i^{th} denoised HSS, and N is the number of signal samples.

4.3. Denoising Results

To verify whether our proposed method can perform noise reduction on HSSs in real-world scenarios, we use noisy HSSs from PhysioNet Computing in Cardiology Challenge 2016 dataset, which are polluted by various unavoidable entities. To validate the stability of our proposed method, one being a normal heart sound and the other a pathological heart sound. Fig. 3 and Fig. 4 clearly demonstrates the method's exceptional denoising efficacy on HSSs. For normal heart sounds, it visually preserves the characteristic waveform of S1 and S2 while effectively eliminating interference noise during intervals. Crucially, the model retains pathological features in abnormal heart sounds during denoising, providing essential auxiliary diagnostic information. Notably, post-processing observation reveals significant noise reduction at S1 and S2 locations in original signals. This confirms the model's capability to filter out non-cardiac noise from respiration and muscle movement, resulting in enhanced signal clarity. Importantly, pathological signatures, including murmurs or accentuated heart sounds, remain well-preserved even in cardiac pathology cases. Hence, DTCWT+ASASS can denoise HSSs effectively under real-world noisy environments.

Based on the numerical analysis of E(SNR) and var(SNR) from the Table 1, DTCWT+ASASS exhibits significant superiority in signal processing. DTCWT+ASASS consistently outperforms DTCWT in output E(SNR), var(SNR) and PE(SNR). In the case of Training-A, DTCWT+ASASS provides a E(SNR)(dB) (8.1 vs. 7.9 vs. 7.1), a var(SNR)(dB)

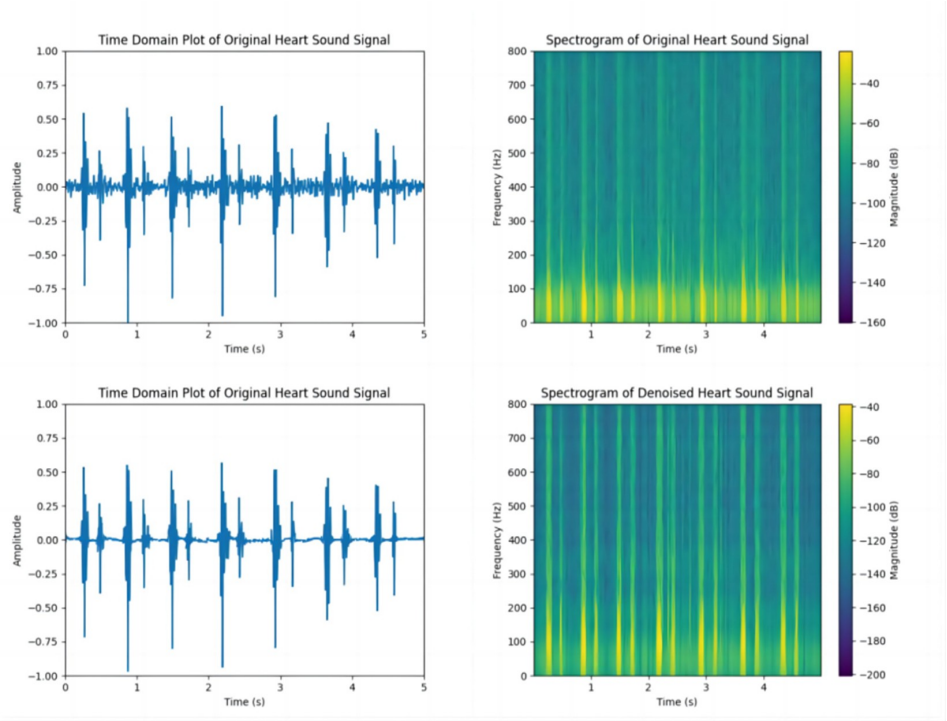


Fig. 3. Time-domain and Spectrogram Plots of Normal and Denoised Normal HSSs

(2.2 vs. 2.4 vs. 2.4) and a PE(SNR)(dB) (+0.14 vs. +0.11 vs 0). In particular, a total of 357 samples (255 normal HSSs, 102 abnormal HSSs) from 409 heart sound recordings (292 normal HSSs, 117 abnormal HSSs) were included for analysis in 121 patients with MVP. These samples included HSSs denoised by DTCWT, HSSs denoised by DTCWT+ASASS, and the original HSSs as a control. E(SNR), var(SNR), and PE(SNR) significantly increased with the increase in denoising degree ($p < 0.001$). This indicates that the observed result differences are extremely unlikely to be caused by random variation, further demonstrating the superiority and reliability of DTCWT+ASASS in HSSs denoising.

Fig. 5 and Fig.6 presents (a) Original HSS, (b) Noise-contaminated HSS, (c) HSS denoised by WT, (d) HSS denoised by DTCWT, and (e) HSS denoised by DTCWT+ASASS. WT achieves noise reduction but concurrently introduces significant signal distortion. Although DTCWT mitigates signal distortion, it demonstrates suboptimal denoising performance in overlapping noise-original signal segments and exhibits susceptibility to pseudo-Gibbs phenomena at signal boundaries. DTCWT+ASASS not only delivers superior noise suppression but also preserves clinically relevant acoustic features (e.g., S1/S2 energy retention) while eliminating perceptible distortion. For the denoised normal HSSs, the spectral energy distribution is uniform, with a clear boundary between the S1 and S2 frequency bands, and the energy is highly concentrated. For the denoised pathological HSSs, mitral stenosis leads to the disappearance of the high-frequency components in S1, with

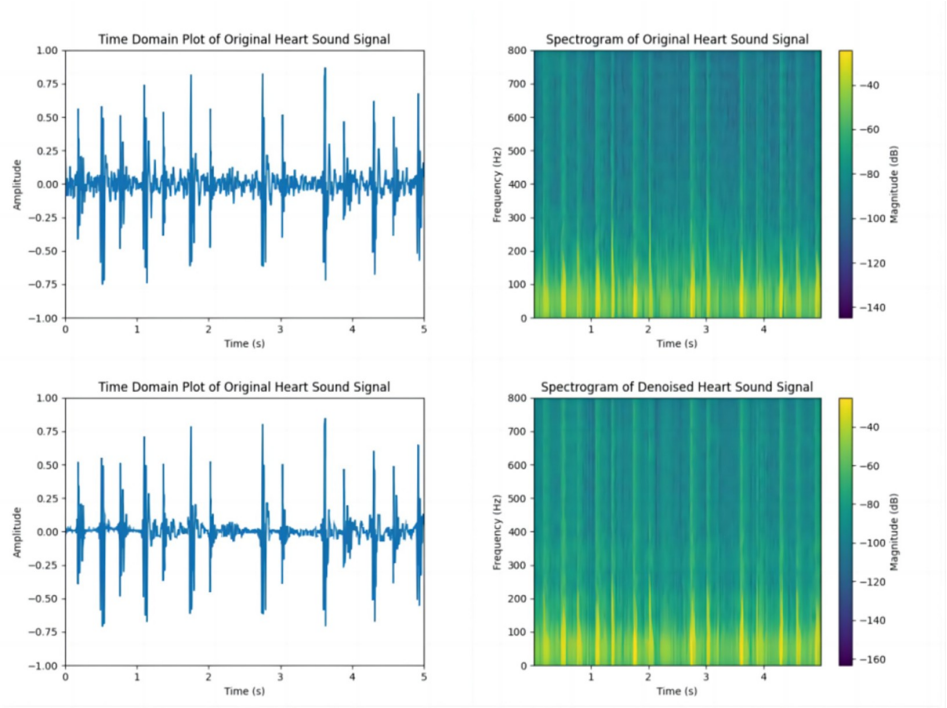


Fig. 4. Time-domain and Spectrogram Plots of Pathological and Denoised Pathological HSSs

energy shifting to the low-frequency band, while aortic insufficiency causes an increase in the low-frequency energy in S2. Compared with DTCWT, DTCWT+ASASS can effectively suppresses pseudo-Gibbs artifacts, and outperforms both benchmark methods in denoising efficacy metrics for both normal and pathological heart sounds.

Table 2 compares the denoising performance of multiple algorithms for HSSs contaminated by GWN at different SNRs of -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB. The results show that DTCWT+ASASS consistently outperforms other comparison methods in terms of SNR and RMSE metrics across all noise levels. Similarly, Table 3 compares the denoising effect of HSSs contaminated by pink noise at different noise levels (including -5 dB, 0 dB, and 5 dB). Under conditions involving GWN SNR of 0 dB, DTCWT+ASASS can reduce the RMSE 30.43% and 15.78%, compared to WT and DTCWT. Correspondingly, under conditions involving pink noise SNR of 0 dB, DTCWT+ASASS can reduce the RMSE by 9.89% and 46.42%, compared to WT and DTCWT. Unlike GWN, pink noise has more energy concentrated in the low-frequency range and gradually diminishes in the high-frequency range. In contrast, we select EMD, which is currently commonly used, as the comparison method for HSSs denoising. This table also shows that DTCWT+ASASS has greater advantages over EMD and DTCWT in terms of SNR and RMSE metrics. Overall, compared to WT, EMD, and DTCWT, DTCWT+ASASS stronger anti-noise ability and greater stability.

Table 1. Denoising performance comparison among DTCWT, DTCWT+ASASS

Recording Categories	Methods	E(SNR)	var(SNR)	PE(SNR)
Training-A	Original HSS	7.1	2.4	0
	DTCWT	7.9	2.4	+0.11
	DTCWT+ASASS	8.1	2.2	+0.14
Training-B	Original HSS	0.7	0.6	0
	DTCWT	10.8	3.5	+13.96
	DTCWT+ASASS	11.1	3.5	+14.86
Training-C+D	Original HSS	5.6	3.7	0
	DTCWT	8.9	4.0	+0.58
	DTCWT+ASASS	9.5	3.8	+0.69
Training-E	Original HSS	5.6	3.0	0
	DTCWT	9.3	4.8	+0.65
	DTCWT+ASASS	10.5	3.9	+0.88

In this way, DTCWT+ASASS has the following characteristics:

(i) The DTCWT achieves time-shift invariance through its unique dual-tree structure, enabling more effective handling of transient features in non-stationary signals while preserving time-frequency localization properties. Compared to conventional WT, the DTCWT demonstrates superior noise separation capability during HSS denoising, thereby enhancing reconstruction accuracy.

(ii) The ASASS algorithm is an advanced HSS processing technique primarily used to smooth high-frequency signals and suppress overlapping noise frequencies in selected bands, thereby recovering noise-free signals. By effectively capturing fine features within HSS fluctuations, the DTCWT+ASASS method achieves superior SNR compared to the DTCWT.

Table 2. Comparison of Denoising Performance at Gaussian White Noise Levels

Noise Level	Performance metrics	WT	DTCWT	DTCWT+ASASS
-5 dB	SNR	4.78	4.66	5.65
	RMSE	0.064	0.052	0.050
0 dB	SNR	7.53	8.62	9.01
	RMSE	0.046	0.038	0.032
5 dB	SNR	11.76	11.48	13.23
	RMSE	0.028	0.026	0.024
10 dB	SNR	16.12	18.01	18.18
	RMSE	0.017	0.016	0.013
15 dB	SNR	19.01	20.96	21.52
	RMSE	0.013	0.010	0.008
20 dB	SNR	23.88	24.37	25.68
	RMSE	0.007	0.006	0.005

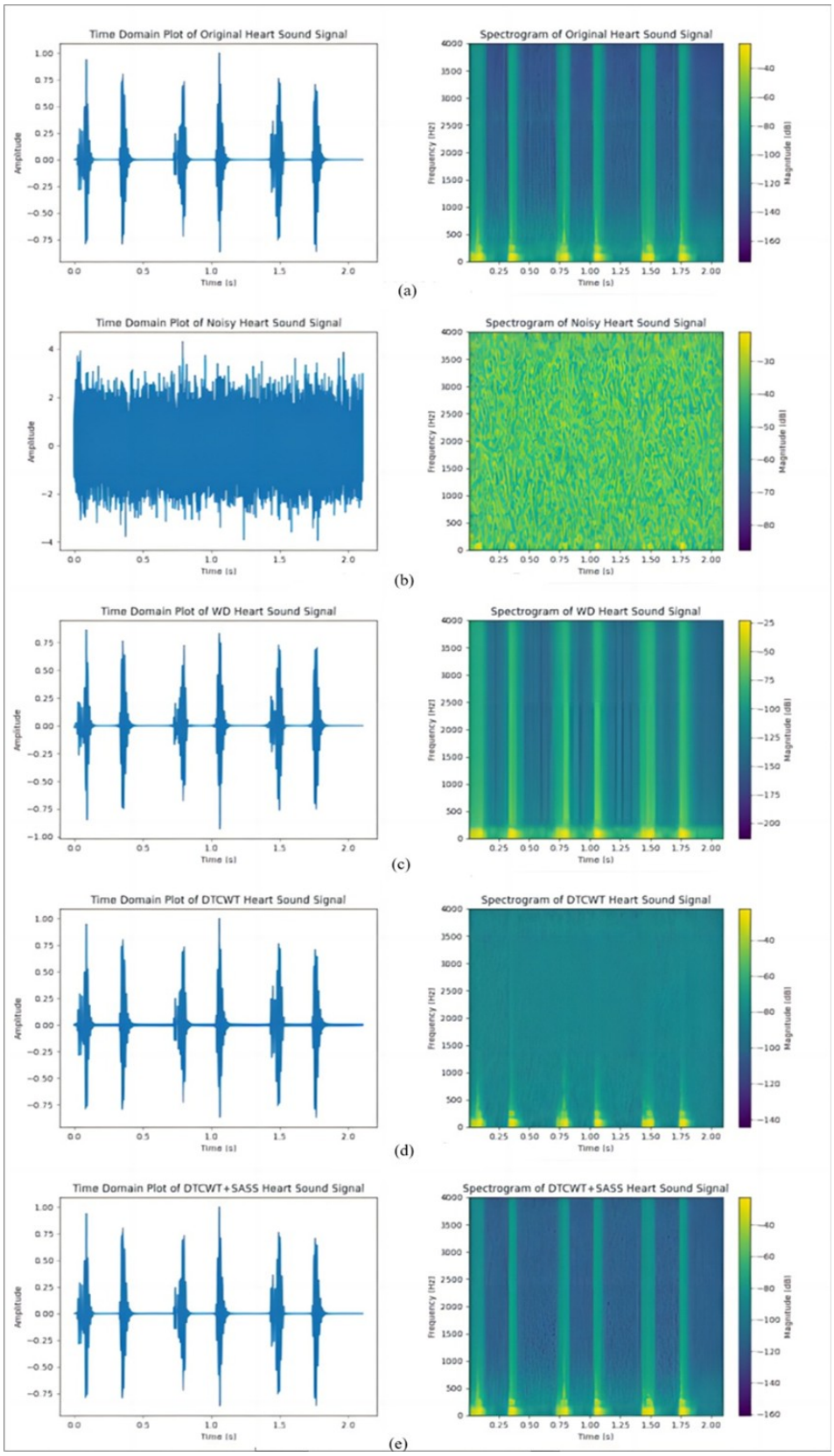


Fig. 5. Denoising Effects of Three Methods on Normal HSSs under Gaussian White Noise with SNR=0 dB: (a) Original HSS, (b) Noise-contaminated HSS, (c) HSS denoised by WT, (d) HSS denoised by DTCWT, and (e) HSS denoised by DTCWT+ASASS

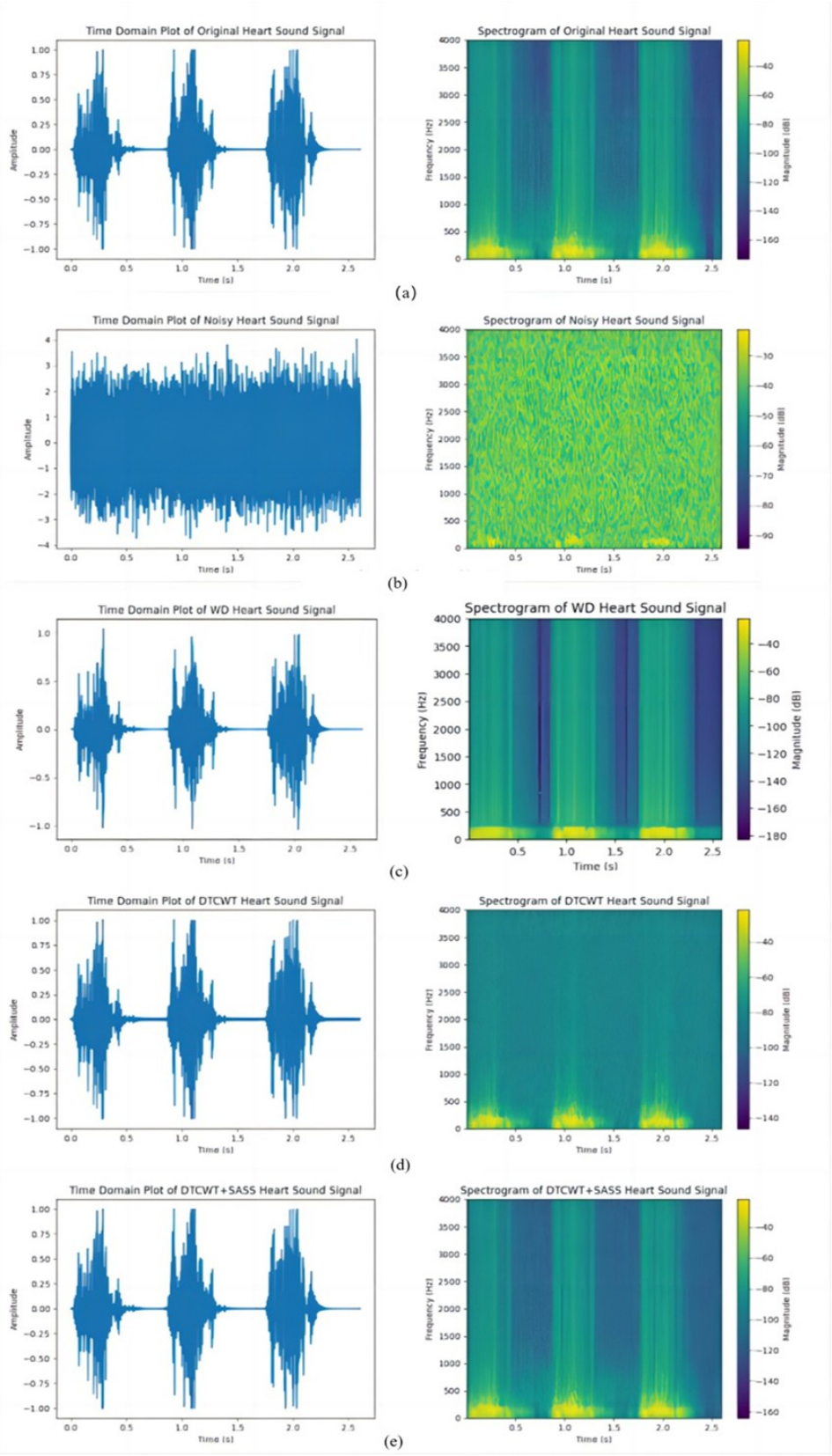


Fig. 6. Denoising Effects of Three Methods on Aortic Stenosis Murmur under Gaussian White Noise with SNR=0 dB: (a) Original HSS, (b) Noise-contaminated HSS, (c) HSS denoised by WT, (d) HSS denoised by DTCWT, and (e) HSS denoised by DTCWT+ASASS

Table 3. Comparison of Denoising Performance at Pink Noise Levels

Noise Level	Performance metrics	EMD	DTCWT	DTCWT+ASASS
-5 dB	SNR	3.25	2.82	4.43
	RMSE	0.162	0.188	0.101
0 dB	SNR	1.62	2.63	5.62
	RMSE	0.093	0.084	0.045
5 dB	SNR	6.51	7.62	11.33
	RMSE	0.053	0.043	0.023

5. Conclusion

In this work, we propose a joint HSSs denoising method using DTCWT and ASASS. First, HSS undergoes preprocessing via Butterworth filtering to reduce high-frequency noise. Subsequently, the signal is decomposed using DTCWT to obtain its multi-scale feature representation. The decomposed signal is then fed into the ASASS algorithm, which automatically adjusts its parameters through a designed adaptive mechanism. Leveraging its sparse adaptive characteristics, the DTCWT+ASASS model extracts critical feature information from the signal and reconstructs it. This approach addresses the limitations of conventional denoising methods, such as signal distortion and low SNRs, while simultaneously overcoming the inherent pseudo-Gibbs phenomena of DTCWT at signal boundaries. Besides, we have validated it on PhysioNet/CinC 2016 database and OAHS Dataset, and both healthy and pathological recordings. Both normal and aortic stenosis HSSs contaminated with AWGN or pink noise at different levels of SNR, our approach achieves significant improvements over state-of-the-art methods. Under conditions involving GWN SNR of 0 dB, the proposed method achieves an SNR of 9.01 dB and a RMSE of 0.032, outperforming standalone DTCWT and multiple existing models.

Acknowledgment. This work was supported by Natural Science Foundation of Fujian Province under Grant No. 2023J011426; Engineering Research Center for Big Data Application in Private Health Medicine of Fujian Universities, Putian University, Putian, Fujian 351100, China (MKF202408).

References

1. Hadiyoso, S., Mardiyah, D., Ramadan, D., Ibrahim, A.: Implementation of electronic stethoscope for online remote monitoring with mobile application. Bulletin of Electrical Engineering and Informatics 9(4), 1595–1603 (2020)
2. Centracchio, J., Parlato, S., Esposito, D., Andreozzi, E.: Accurate localization of first and second heart sounds via template matching in force cardiography signals. Sensors 24(5), 1525 (2024)
3. Hu, J., Hu, Q., Liang, M.: Heart sounds classification using adaptive wavelet threshold and 1d ldcnn. Computer Science and Information Systems 20(4), 1483–1501 (2023)
4. Xiao, F., Liu, H., Lu, J.: A new approach based on a 1d+2d convolutional neural network and evolving fuzzy system for diagnosis of cardiovascular disease from heart sound signals. Applied Acoustics 216, 109723 (2024)

5. Gilles, J.: Empirical wavelet transform. IEEE Transactions on Signal Processing 61(16), 3999–4010 (2013)
6. Al-Shannaq, M., Nasrawi, A., Bsoul, A., Saifan, A.: Abnormal heart sound recognition using svm and lstm models in real-time mode. Scientific Reports 15(1), 9129 (2025)
7. Nawaz, S., Li, J., Li, D., Shoukat, M., Bhatti, U., Raza, M.: Medical image zero watermarking algorithm based on dual-tree complex wavelet transform, alexnet and discrete cosine transform. Applied Soft Computing 169, 112556 (2025)
8. Nia, P., Hesar, H.: Abnormal heart sound detection using time-frequency analysis and machine learning techniques. Biomedical Signal Processing and Control 90, 105899 (2024)
9. Li, J., Ke, L., Du, Q.: Classification of heart sounds based on the wavelet fractal and twin support vector machine. Entropy 21(5), 472 (2019)
10. Duggan, D., Temko, A., Sarana, V., Factor, A., Popovici, E.: Denoising of heart sounds using lightweight fcns and spectrograms with and without context. IEEE Access (2025)
11. Ali, S., Shuvo, S., Al-Manzo, M., Hasan, A., Hasan, T.: An end-to-end deep learning framework for real-time denoising of heart sounds for cardiac disease detection in unseen noise. IEEE Access 11, 87887–87901 (2023)
12. Hu, J., Chen, L., Yang, M., Shen, S., Gao, X.: Psbd-ewt-egan: Heart sound denoising using psbd-ewt and enhancement generative adversarial network. Computer Science and Information Systems 00, 5–5 (2025)
13. Liu, C., Springer, D., Li, Q., Moody, B., Juan, R., Chorro, F., Castells, F., Roig, J., Silva, I., Johnson, A.: An open access database for the evaluation of heart sound algorithms. Physiological Measurement 37(12), 2181 (2016)
14. Mat Rozi, N., Hashim, F., Shaharuddin, S., Miskan, M., Ahmad, K., Salleh, M.: Comparison on wavelet adaptive filter performance in denoising ecg signal (2023)
15. Houamed, I., Saidi, L., Srairi, F.: Ecg signal denoising by fractional wavelet transform thresholding. Research on Biomedical Engineering 36(3), 349–360 (2020)
16. Hu, J., Chen, L., Shen, S., Wang, T.: Explainable multi-agent deep reinforcement learning for joint task offloading and resource allocation in distance- and channel-aware noma vehicular edge networks. IEEE Internet of Things Journal (2025)
17. Deepak, H., Vijayakumar, T.: Optimal threshold estimation using grey wolf optimization for emd-dtcwt based ecg denoising. International Journal of Recent Technology and Engineering (2020)
18. Wang, F., Ji, Z.: Application of the dual-tree complex wavelet transform in biomedical signal denoising. Bio-Medical Materials and Engineering 24(1), 109–115 (2014)
19. Prashar, N., Sood, M., Jain, S.: Design and implementation of a robust noise removal system in ecg signals using dual-tree complex wavelet transform. Biomedical Signal Processing and Control 63, 102212 (2021)
20. Hou, Y., Liu, R., Shu, M., Chen, C.: An ecg denoising method based on adversarial denoising convolutional neural network. Biomedical Signal Processing and Control 84, 104964 (2023)
21. Liu, E., Liu, Y., Wang, Z., Zhou, H., Wang, X.: A noise prediction method for paper-based grayscale ecg denoising. Biomedical Signal Processing and Control 111, 108408 (2026)
22. Li, Z., Tian, Y., Jin, Y., Wei, X., Wang, M., Liu, J., Liu, C.: Eddm: A novel ecg denoising method using dual-path diffusion model. IEEE Transactions on Instrumentation and Measurement (2025)
23. Peng, H., Chang, X., Yao, Z., Shi, D., Chen, Y.: A deep learning framework for ecg denoising and classification. Biomedical Signal Processing and Control 94, 106441 (2024)
24. Haider, N., Behera, A.: Respiratory sound denoising using sparsity-assisted signal smoothing algorithm. Biocybernetics and Biomedical Engineering 42(2), 481–493 (2022)
25. Haider, N., Behera, A.: Computerized respiratory sound-based diagnosis of pneumonia. Medical & Biological Engineering & Computing 62(1), 95–106 (2024)

26. Pouyani, M., Vali, M., Ghasemi, M.: Lung sound signal denoising using discrete wavelet transform and artificial neural network. Biomedical Signal Processing and Control 72, 103329 (2022)
27. Massar, H., Drissi, T., Nsiri, B., Miyara, M.: Advancements in blind source separation for eeg artifact removal. Applied Acoustics 228, 110300 (2025)
28. Anupallavi, S., Ashokkumar, S., Premkumar, M., Sangeetha, R.: Enhanced eeg signal classification through integrated dtcwt, human learning optimization, and optimized dbn-hmm models for improved performance. Biomedical Signal Processing and Control 110, 108229 (2025)
29. Haider, N.: Respiratory sound denoising using empirical mode decomposition, hurst analysis and spectral subtraction. Biomedical Signal Processing and Control 64, 102313 (2021)
30. Xiahou, S., Liang, Y., Ma, M., Du, M.: A strong anti-noise segmentation algorithm based on variational mode decomposition and multi-wavelet for wearable heart sound acquisition system. Review of Scientific Instruments 93(5) (2022)
31. Faradisa, I., Ananda, A., Sardjono, T., Purnomo, M.: Denoising of fetal phonocardiogram signal by wavelet transformation. In: E3S Web of Conferences. vol. 188, p. 00013. EDP Sciences (2020)
32. Alali, S., Kachenoura, A., Albera, L., Hernandez, A., Michel, C., Senhadji, L., Karfoul, A.: Optimized cnn-based denoising strategy for enhancing longitudinal monitoring of heart failure. Computers in Biology and Medicine 184, 109430 (2025)
33. Al-Naami, B., Fraihat, H., Al-Nabulsi, J., Gharaibeh, N., Visconti, P., Al-Hinnawi, A.: Assessment of dual-tree complex wavelet transform to improve snr in collaboration with neuro-fuzzy system for heart-sound identification. Electronics 11(6), 938 (2022)
34. Renuka, S., Edla, D.: Adaptive shrinkage on dual-tree complex wavelet transform for denoising real-time mr images. Biocybernetics and Biomedical Engineering 39(1), 133–147 (2019)
35. Narváez, P., Percybrooks, W.: Synthesis of normal heart sounds using generative adversarial networks and empirical wavelet transform. Applied Sciences 10(19), 7003 (2020)

Jianqiang Hu is an associate professor in school of computer and information engineering, Xiamen University of Technology, China. He once worked as a postdoctoral researcher at Tsinghua University. He received his Ph.D. degree in computer science and engineering from National University of Defense Technology, China, in 2005. He is the author of more than 70 articles, including respected journals such as CHINESE JOURNAL OF COMPUTERS, IEEE INTERNET OF THINGS JOURNAL and IEEE TRANSACTIONS ON BIG DATA. His current research interests include Edge Computing, Biomedical Signal Processing, and Big Data Analytics.

Dafeng Shen is a master student at school of computer and information engineering, Xiamen University of Technology, China. Her research interests include Edge Computing, Biomedical Signal Processing and Deep Reinforcement Learning.

Lin Chen is a master student at school of computer and information engineering, Xiamen University of Technology, China. His research interest include Edge Computing and Biomedical Signal Processing.

Yu Chen is a master student at school of computer and information engineering, Xiamen University of Technology, China. His research interests include Biomedical Signal Processing and Deep Reinforcement Learning.

Shigen Shen received the B.S. degree in fundamental mathematics from Zhejiang Normal University, Jinhua, China, in 1995, the M.S. degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2005, and the Ph.D. degree in pattern recognition and intelligent systems from Donghua University, Shanghai, China, in 2013. He is a Professor with the School of Information Engineering, Huzhou University, Huzhou, China. He has published more than 100 technical papers, including respected journals such as IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and IEEE INTERNET OF THINGS JOURNAL. His current research interests include Internet of Things, cyber security, edge computing, and game theory.

Yan Che is a professor at Putian University. She received her ME degree in Computer Application from Xiamen University, Xiamen, China, in 2005, and Ph.D. degree in Control Theory and Engineering from Donghua University, Shanghai, China, in 2016. She is currently working in Engineering Research Center for Big Data Application in Private Health Medicine of Fujian Universities, Putian University, Putian, Fujian, China. Her research interests include Data Processing, Data Analysis and Data Security.

Received: August 28, 2025; Accepted: January 5, 2026.

Computer Science and Information Systems 23(2):707–728 https://doi.org/10.2298/CSIS251022018R

Transformer Substation Network Disconnection Prediction via Semantic Reasoning with Causal Modeling

Jie Ren[1], Xiaojun Yao[1], and Hong Chen[1]

Suzhou Suneng Group Co., LTD
215004 Suzhou Jiangsu, China
rengj9@js.sgcc.com.cn
yaoxj1@js.sgcc.com.cn
ch_jt_sz@js.sgcc.com.cn (corresponding author)

Abstract. Reliable communication networks are indispensable for the stable operation of smart grids and substations. Currently, WAPI networks have been widely adopted in relevant scenarios. Nevertheless, WAPI networks are confronted with disconnection risks attributed to complex network topologies, dynamic traffic fluctuations, and external environmental disturbances. Most methods rely on correlation analysis and lack causal interpretability, which restricts their effectiveness in root-cause localization and preventive maintenance practices. To address the problem, we propose a disconnection prediction approach that integrates prompt-driven semantic reasoning with structured causal analysis. The approach constructs a causal event graph that models semantic, temporal, and topological dependencies across devices and alarm sequences after extracts heterogeneous information to unified event representation. Based on the established graph, an inference module combines causal path analysis, structural causal models, and counterfactual reasoning to assess the influence of events, predict emerging disconnection risks, and identify plausible root causes with coherent and interpretable justification. By tightly coupling semantic abstraction with causal reasoning, the proposed approach provides a proactive, explainable, and extensible mechanism for anticipating network disruptions and supporting informed maintenance decisions. Experiments demonstrate that the proposed approach improves prediction accuracy and interpretability, verifying its value for smart grid communication networks.

Keywords: Causal Inference, Network Disconnection Prediction, Root Cause Analysis, Incident Causality Graph, Substation, Disaster Recovery.

1. Introduction

The developing energy transition and the advancement of power system intelligence have made smart grids and substations vital infrastructures for secure and stable power system operation [23]. In these systems, communication networks enable device interaction and control signal transmission, forming the backbone of business continuity and reliability [14] . To meet the stringent security, reliability, and low-latency requirements of such environments, WAPI (Wireless Local Area Network Authentication and Privacy Infrastructure) network has been widely adopted due to its robust encryption, mutual authentication, resistance to common wireless attacks, and ability to maintain stable communication under high traffic and interference. As a result, WAPI-based devices are extensively

deployed in substations, supporting secure and dependable operations. However, due to the complexity of topology, traffic fluctuations, and external disturbances, these devices often experience disconnection events, leading to service interruptions and even cascading failures [25], which undermine both operational safety and economic efficiency.

To achieve high-assurance communication networks, traditional approaches generally rely on post-event response and static redundancy design [2]. However, these approaches are increasingly inadequate under highly dynamic and uncertain operating environments. In contrast, preventive maintenance mechanisms can proactively mitigate potential risks before failures occur, thus reducing the likelihood of network disconnections [18]. Within this framework, network disconnection prediction becomes a critical component: only when potential risks are predicted and explained in advance can operators take timely countermeasures, thereby enhancing the stability and resilience of the overall network. Existing mainstream approaches for network fault detection are predominantly based on correlation modeling, such as analyzing time-series fluctuations, delay drifts, and clustering of anomalous behaviors [21]. While effective to some extent, these approaches cannot generally capture causal mechanisms among variables, making it difficult to distinguish between "fault symptoms" and "fault root causes" [20]. This limitation directly constrains both the accuracy and interpretability of disconnection prediction.

Recently, the emergence of Large Language Models (LLMs) has opened new opportunities for knowledge modeling and semantic reasoning in complex systems [19]. LLMs can extract key semantic events from unstructured sources such as logs, configuration documents, and alarm texts [29]. Moreover, when integrated with causal discovery algorithms (e.g., the Peter-Clark algorithm, also known as PC) and prompt engineering, LLMs can help construct causal dependency graphs and uncover potential fault propagation paths [7]. Leveraging these capabilities, LLMs demonstrate strong potential in enhancing the generalization, interpretability, and timeliness of disconnection risk prediction.

To address the limitations of conventional approaches, we propose a novel network disconnection prediction approach that integrates LLMs-based semantic reasoning with causal modeling for WAPI network devices in smart grid and substation environments. The approach leverages LLMs to unify heterogeneous operational data and construct causal event graphs, while incorporating causal inference techniques such as path analysis, structural causal models, and counterfactual reasoning to identify root causes and predict potential disconnection risks. By combining the semantic strengths of LLMs with the interpretability of causal modeling, the proposed approach enhances adaptability to complex and unseen fault scenarios. Experimental results confirm its effectiveness in improving the accuracy and timeliness of disconnection prediction. Overall, we introduce a LLMs–causal modeling approach for proactive disconnection prediction, offering a practical paradigm for preventive maintenance in mission-critical communication systems of smart grids and substations. This study differs from existing correlation-based approaches by combining the semantic reasoning of LLMs with causal inference, enabling both accurate prediction and interpretable root-cause analysis. The proposed approach unifies heterogeneous data into causal structures and supports proactive disconnection prevention through scalable causal reasoning. The main contributions of our approach are summarized as follows:

(1) We present a unified event representation scheme that transforms heterogeneous network data into structured semantic events, providing a standardized basis for causal inference and prediction.
(2) We propose an LLMs-driven causal graph construction mechanism that leverages prompt engineering to extract causal dependencies among events, combining semantic reasoning with temporal and topological constraints to enhance accuracy and interpretability.
(3) We design a pluggable inference module that integrates causal path analysis, structural causal models, and counterfactual reasoning, enabling the identification of root causes and the prediction of high-confidence disconnection.
(4) We conduct experiments under representative smart grid network scenarios that demonstrate improved prediction accuracy, causal explainability, and response timeliness.

In the subsequent sections, section 2 reviews the relevant prior work, section 3 presents the methodology proposed in this study, section 4 details the experimental setup and configurations, section 5 provides a comprehensive analysis of the experimental results to demonstrate the effectiveness of the proposed approach, and section 6 concludes the study.

2. Related Work

2.1. Network Fault Diagnosis and Prediction

With the continuous growth of network scale and increasing diversity of services, traditional rule- or threshold-based network fault diagnosis struggles to handle the dynamic and complex modern networks effectively. Consequently, machine learning and deep learning have been widely applied to network disconnection prediction and root-cause diagnosis in recent years. They take advantage of network flow, delay, packet loss, and bandwidth utilization to identify potential anomalies or failures [17]. However, traditional machine learning models face challenges in high-dimensional, dynamic, and structured networks, including strong label dependence, insensitivity to topology changes, and limited ability to capture multi-failure dependencies, which restrict their practical effectiveness.

Deep learning approaches offer automatic feature extraction and complex pattern modeling. Recently, various neural networks have been applied to network disconnection prediction tasks. For instance, Alkaberi et al. used CNNs with MLPs to predict software faults [1]; Gupta et al. applied DNNs with hyper-parameter tuning to assess the impact of software faults [10]; and Cheng et al. introduced Attention-LSTM for time series prediction of intermittent faults [6]. While these approaches improve prediction accuracy, they largely remain black-box models, lacking interpretability and clear root-cause identification, and have limited ability to locate critical network links.

Recently, causal reasoning has been increasingly incorporated into network fault diagnosis to enhance interpretability and reasoning. Causal approaches model dependencies between variables and identify true trigger sources and propagation paths. For example, Ghosh et al. [9] proposed a cascading disconnection prediction approach using causal graphs for early warning in power transmission networks. Li et al. [12] designed a root-cause analysis approach for online service systems using causal Bayesian networks, which significantly improved response accuracy and speed.

These studies indicate that, although machine learning and deep learning show initial success in network fault diagnosis, they still struggle with modeling causal structures, handling multi-hop dependencies, and providing interpretable insights for practical operations.

2.2. Causal Analysis

Causal analysis provides an effective means to compensate for these deficiencies and has gradually become a significant research direction in the field of intelligent operation and maintenance. With the rapid development of general artificial intelligence tools, such as large models, causal analysis approaches are also gradually integrating with multimodal intelligence components, including natural language understanding and knowledge extraction. This integration provides a new technological path to realize network disconnection prediction systems with strong interpretability and generalization capabilities.

Causal analysis algorithms are classified into three categories: constraint-based, scoring-based, and non-Gaussian assumption-based. The PC algorithm is a typical constraint-based causal discovery approach, in which dependencies between variables are inferred through conditional independence tests. The approach consists of two stages: first, constructing an undirected causal skeleton graph, then determining the direction of some edges based on a series of rules, and finally forming a partially directed acyclic graph, which is suitable for data with both discrete and continuous variables [22][15].

In contrast, the GES algorithm (Greedy Equivalence Search) is a scoring-based search approach that does not require independence tests. It performs greedy optimization using scoring functions (e.g., Bayesian Information Criterion, also known as BIC) by performing structural modification operations (e.g., adding edges, deleting edges, reversing edges) in the equivalence class space to find the optimal Bayesian network structure. The approach is efficient but prone to fall into local optimality [11][16].

LiNGAM algorithm (Linear Non-Gaussian Acyclic Model) is a linear causal discovery approach based on structural equation modeling, which takes advantage of the non-Gaussian characteristics of the variables and recovers the directed acyclic structure among the variables by Independent Component Analysis (ICA), which breaks through the limitation that the direction of causality is not identifiable under the Gaussian assumption [24][4]. These approaches have been widely applied in the fields of biomedicine, social sciences, and economic modeling, providing important tools for understanding causal mechanisms in complex systems.

2.3. Large Language Models

In recent years, a large number of unstructured logs and heterogeneous alarms, such as syslog, SNMP Trap, NetFlow, etc., have been generated in network operation and maintenance scenarios, which are semantically diverse and confusing, and are often difficult to be uniformly abstracted and fused by traditional rules or keyword matching approaches. LLMs such as GPT, LLaMA, and Claude have powerful contextual understanding and linguistic reasoning capabilities [3], which can automatically map these unstructured texts into unified fault event labels, thus improving the quality of event normalization and cross-source fusion capabilities [5].

Through Chain-of-Thought (CoT) and Prompt Chaining techniques, LLMs can analyze temporal and semantic logical relationships in event sequences to help identify potential causal trigger paths. It has been demonstrated that the in-context learning approach using GPT-4 can significantly enhance the accuracy and readability of automated root cause analysis, eliminating the need for fine-tuning, in the context of fault analysis in cloud platforms [28]. In addition, studies have also designed reliability assessment mechanisms, such as the PACE LM framework [27], to calibrate the confidence level of LLMs through the cue-enhancement and retrieval-enhancement (RAG) strategy, which effectively reduces the risk of generating "hallucinatory" information and improves the usability and security in real system scenarios.

LLMs have significant advantages in log abstraction, causal reasoning, disconnected prediction, and natural language interpretation, making them an important technical component of network fault diagnosis and prediction systems.

In addition to the above applications, recent research has explored the integration of LLMs with causal graph construction and preventive maintenance frameworks. By combining LLMs-based semantic reasoning with structural causal models, these approaches enable proactive disconnection prediction, root-cause identification, and explainable fault propagation analysis. Such integration not only enhances the interpretability of automated diagnostics but also provides practical guidance for network operators in large-scale and complex infrastructures, highlighting the potential of LLMs as a core component in intelligent network operation and maintenance systems.

3. Methodology

This section proposes a disconnection prediction approach for network devices that integrates a large model and a causal reasoning approach, aiming to solve the problems of strong dependence on structural assumptions, weak scalability, and limited predictive foresight in traditional causal modeling. With LLMs as the core, the proposed approach automatically generates causal event maps based on multi-source alarm data and structural information. It combines with a downstream pluggable root cause analysis module to realize real-time localization and prediction of disconnected faults.

3.1. Overview

As shown in Fig. 1, our proposed disconnected prediction approach comprises three core stages: data preprocessing and unified representation, large model-driven causal mapping construction, and a pluggable root cause localization and disconnection prediction mechanism. The goal of this process is to fully utilize the advantages of LLMs in processing unstructured data and reasoning about contextual relationships, while retaining the explanatory power and decision support capabilities of causal modeling in complex systems.

In the data preprocessing stage, we unify the abstraction and standardization of heterogeneous information, including network device logs, status monitoring indicators, and topology. By uniformly modeling information such as Syslog, SNMP alarms, and link state changes in the network, we establish a structured event flow that provides a consistent basis for subsequent causal modeling.

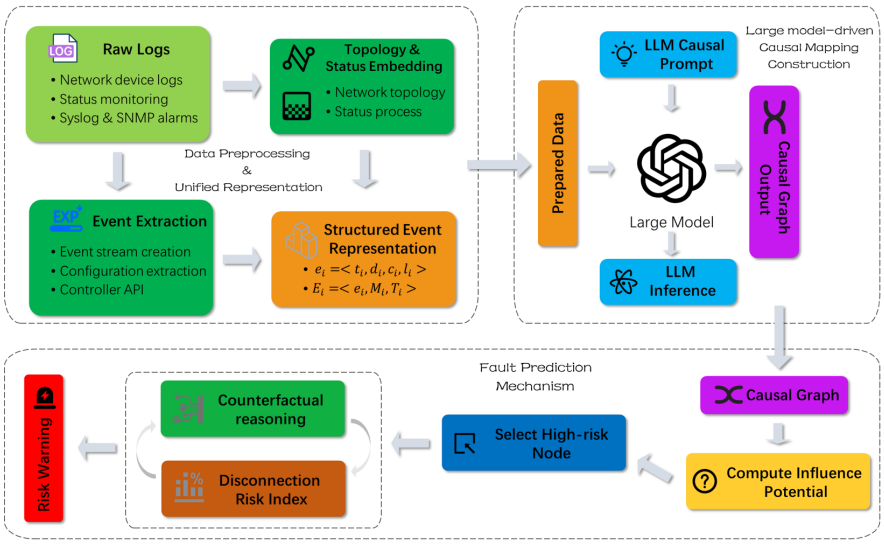


Fig. 1. The overview of our approach. The disconnected prediction approach proposed comprises three core stages: data preprocessing and unified representation, large model-driven causal mapping construction, and a pluggable root cause localization and disconnection prediction mechanism.

Subsequently, we design a prompt-based event recognition and causality extraction mechanism that leverages the semantic understanding and generation capabilities of LLMs. The large language model can generate a directed causal graph reflecting the dependencies between alarm events by receiving historical event context, a priori network structure, and system configuration information.

Finally, based on the completion of causal graph construction, we introduce a set of plugin causal inference modules, including multiple algorithms such as path analysis, structural causal modeling (also known as SCM), and counterfactual inference, among others, for identifying potential disconnection triggers and enabling high-confidence risk warnings in future windows. This multi-strategy fusion design ensures the approach's stability and migratability in different types of network environments.

Our approach constructs a network disconnection prediction framework that offers high generalization, interpretability, and predictive ability through the deep integration of large models and causal modeling. The following section introduces the design and implementation details of key modules, including causal event graph construction and the root cause inference mechanism.

3.2. LLMs-based causal event graph construction

In network operation, maintenance, and fault analysis, the first step requires accurately constructing a causal map between device anomalies. The traditional approach relies on

statistical tests (e.g., conditional independence tests) and structural learning algorithms (PC, GES). These models typically require variables to satisfy specific distributional assumptions and are poorly suited for network environments with complex structures and sparse data. To this end, we propose an automatic causal event graph construction approach centered on the LLMs, which replaces the heavy dependency test and graph structure generation process in traditional algorithms.

Structured Event Representation. In the actual network operation and maintenance environment, fault data exists in various forms, including Syslog logs, SNMP Traps, NetFlow messages, and link state monitoring. These data sources have obvious heterogeneity. These data sources are obviously heterogeneous, comprising both structured metric information and a large amount of unstructured textual content, such as alarm descriptions and log records. To ensure the accuracy of subsequent causal inference and generation structure, we will construct a unified, structured event representation mechanism to support the approach's subsequent steps effectively.

We adopt a standardization approach based on event quaternions to uniformly abstract network operation and maintenance data from different sources into structured alarm event units. Each base alarm event is represented in the following form:

$$e_i =< t_i, d_i, c_i, l_i >, \tag{1}$$

where t_i denotes the timestamp of the event, which is used to indicate the timing information of the event, d_i denotes the device number or unique identifier number that the event belongs to, i.e., device ID; c_i denotes the type of the event or the alarm category, including "link interruption", "route drift", etc.; denotes the severity level of the event, which is graded from 1 to 5, with higher levels indicating more seriousness. c_i denotes the type of event or alarm category, including "link interruption", "route drift", etc.; l_i denotes the severity level of the event, which is divided into 1 to 5 levels, with the higher level denoting the more serious.

As an example, a real-world scenario in which device R1 detects a severe link anomaly at 13:12 can be represented as:

$$e =< 2024-07-15\ 13:12:15,\ R1, linkinterruption, 5 > . \tag{2}$$

To enhance the contextual connectivity of the event, we expansively introduce two additional fields M_i and T_i, the former denoting a snapshot of the device state at the time of this base event, such as the device's bandwidth usage and CPU utilization at the time of the event, etc., and the latter denoting the network topology neighbors at the time of the event, which form an expanded event with the base event, which can be represented as follows:

$$E_i =< e_i, M_i, T_i > . \tag{3}$$

Such an event representation preserves the important contextual features of network anomalies on the one hand, and facilitates inter-event comparisons, ordering, and relationship modeling on the other. The structured representation greatly reduces the ambiguity of language model inputs and provides a standard, unified semantic foundation for subsequent causal cue design, graph structure generation, and path inference.

The following are network events during a certain period of time, the event format is: $E_i =< e_i, M_i, T_i >, e_i =< t_i, d_i, c_i, l_i >$,

{Event Symbol Explanation},

determine which events may lead to other events:

[Event 1]:<<2024-07-15 13:02:10, R1,Link Outage,5>,{CPU:0.8,Interface Utilization:0.87,Packet Loss:0.02},[R2,R3]>>

[Event 2]:<<2024-07-15 13:03:19, R2,Route Drift,3>,{CPU:0.4,Interface Utilization:0.82,Packet Loss:0.0},[R1,R4]>

...

Please output the possible causal relationships between events in the format [event i] -> [event j].

Fig. 2. Prompt template used to convert structured event sequences into natural-language queries that guide the LLMs in inferring causal relationships between events.

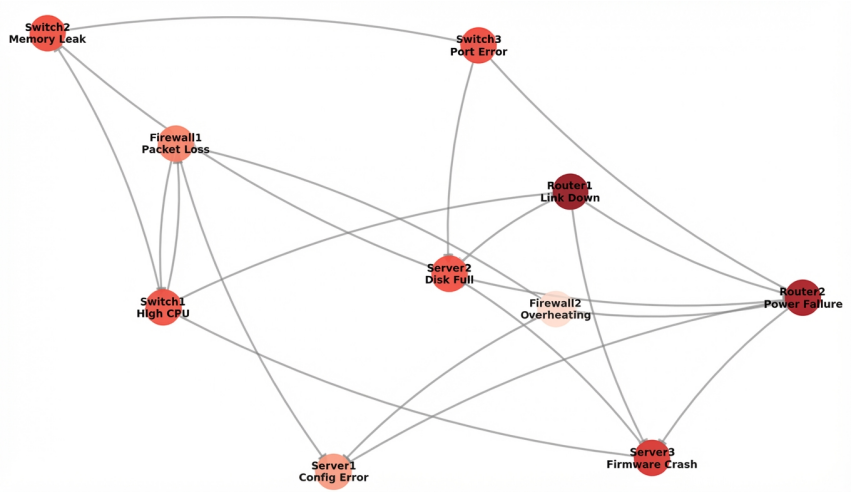


Fig. 3. Illustration of the resulting structural causal graph constructed from validated event dependencies.

Causal Edge Generation. Based on structured event representation, we further utilize LLMs to reason about potential causal relationships between events and construct a directed causal event graph accordingly. The core of the process lies in the use of prompt engineering to guide the LLMs to identify causal dependencies between events, thus replacing traditional structural learning algorithms such as the PC approach. By combining temporal, semantic, and structural information, the generated causal graph can be more closely related to the fault propagation paths in real networks.

First, we transform the event sequences collected within the time window into a unified natural language input template, which is then fed to the LLMs for causal edge inference. Each event in the template is presented in either the base event format or the extended event format, as described in the previous section, ensuring that the model has sufficient contextual information. A sample prompt template is shown in Fig.2.

After receiving the cue, LLMs will output a set of candidate causal edges $E_{causal} = \{(e_i \rightarrow e_j)\}$ based on the context. To enhance the credibility of the edges and mitigate

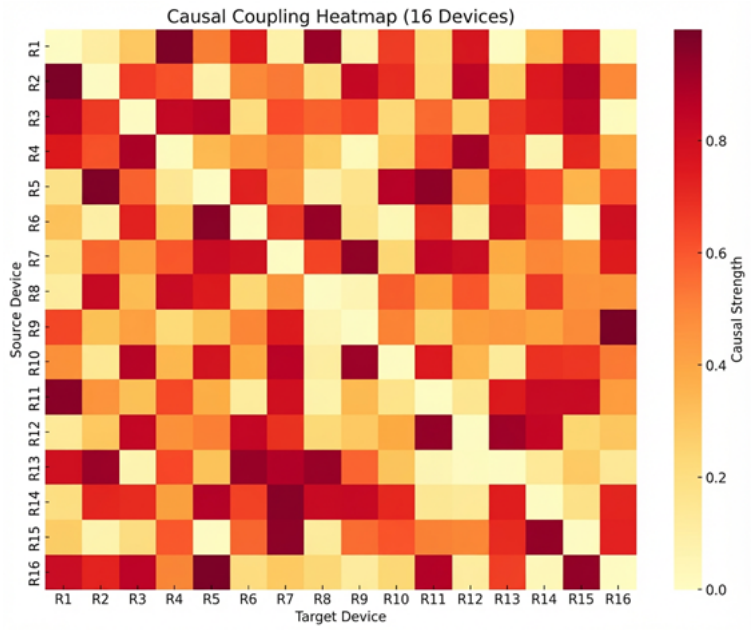


Fig. 4. Coupled heat map visualizing the pairwise causal influence scores between events in the final graph.

the noise in the graph structure, we have developed a three-stage causal edge scoring and screening mechanism.

First, for each candidate edge, we define the following causal confidence scoring function:

$$\text{score}(e_i \to e_j) = \alpha \cdot \text{sim}(c_i, c_j) \quad + \beta \cdot e^{-\lambda(t_j - t_i)} \quad + \gamma \cdot \text{adj}(d_i, d_j), \tag{4}$$

where sim (c_i, c_j)denotes the cosine similarity of two events in the semantic space, which the LLMs calculate after generating the event embedding; $e^{-\lambda(t_j - t_i)}$ is the temporal decay factor, which ensures that the causal direction adheres to the principle of temporal sequentiality; adj (d_i, d_j), on the other hand, indicates the proximity of two event-generating devices on the network topology. The weights of the three terms α, β, γ can be adjusted according to the actual application to strengthen the model perception.

Second, for causal edges with scores higher than the threshold τ, we further guide the LLMs to generate explanatory text, which is used to verify whether the model outputs have logical rationality, e.g., ask the model, "Why does event e_i cause event e_j?", and if the model cannot reasonably explain a causal relationship, e.g., the reason is ambiguous or semantic conflict, the edge will be eliminated. This step enhances the interpretability of the causal graph and reduces the inference error caused by "language illusion".

Finally, to avoid structural conflicts, we perform consistency checking in conjunction with network device topology graphs to remove unreasonable edges. This procedure consists of two constraints: loop exclusion and topology direction constraints.

Loop exclusion means that if adding the edge $(e_j \to e_i)$ would form a closed loop, it is removed to maintain the topologically orderable nature of the graph.

The topological direction constraint requires that if events e_i and e_j originate from devices d_i and d_j, respectively, then a causal edge is retained only when the inter-device hop count $dist_T(d_i, d_j) \leq 2$ and the topological direction between them is either downstream or at the sibling level. The process can be formally expressed as follows:

$$(e_i \to e_j) \in E_{\text{causal}} \Leftrightarrow \begin{cases} \text{acyclic after addition} \\ d_T(d_i, d_j) \leq 2 \\ \delta_T(d_i, d_j) \in \{\downarrow, =\} \end{cases} \tag{5}$$

Graph Construction Mechanisms. After completing causal edge extraction and multi-stage filtering, the remaining high-confidence edges are aggregated to form the final directed causal graph $G_{causal} = (V, E_{causal})$, where the node set V corresponds to all structured events and the edge set E_{causal} represents validated causal dependencies. Unlike a conventional correlation graph, the resulting structure preserves both temporal directionality and physical topology constraints, ensuring that the generated causal pathways are consistent with real-world fault propagation characteristics in substations and intelligent grid networks.

Fig. 3 shows the directed structural causal event graph of an example by the proposed construction pipeline, where nodes represent semantically encoded events and arrows denote validated causal relations. Fig. 4 provides the corresponding heat map, in which darker cells indicate more substantial causal influence from event e_i to event e_j.

Temporal Enhancement of Causal Graphs. We introduce a sliding time window mechanism to serialize the construction process of the causal event graph for modeling purposes. Given a time span T, we construct the set of events ε_i within the current window at each time step t and generate the corresponding causal graph $G_t = (V_t, E_t)$ on its basis. As the window slides, the model will obtain a series of causal event graph sequences $G_{t-k}, \ldots, G_t$, forming a type of time-evolutionary map that provides the basis for subsequent trend modeling and forecasting.

And then, considering that the causal strength of different edges may change over time, we introduce a time-sensitive dynamic weight function w_t for each causal edge $e_i \to e_j$, which is defined as follows:

$$w_t\left(e_i \to e_j\right) = \eta \cdot s_0 + (1 - \eta) \cdot \text{EMA}_t\left(s_t\right), \tag{6}$$

where s_0 denotes the initial causal confidence score, s_t denotes the frequency of triggering of this edge within the current window (i.e., the number of times that event e_i occurs that causes e_j within the window), EMA_t is an exponential moving average function reflecting the recent trend of dependence, and $\eta \in [0, 1]$ controls the proportion of fusion between the prior and the current dynamics.

3.3. Downstream Causal Reasoning and Disconnected Prediction Mechanisms

After completing the causal graph construction and temporal enhancement based on large models, the system has obtained a sequence of directed causal graphs covering major

Algorithm 1: Causal Graph-Based Disconnection Prediction

Input: $G_t = (V_t, E_t)$, w_t, θ, k
Output: Alert node list alerts
Function *ComputeInfluencePotential*(G_t, w_t)
  $\Psi \leftarrow \{\}$
  foreach *node* $e_i \in V_t$ **do**
    $P \leftarrow$ FindAllPaths(G_t, e_i) path_energies $\leftarrow [\,]$
    foreach *path* $p \in P$ **do**
      prod $\leftarrow 1.0$ **foreach** *edge* $(e_j \rightarrow e_k) \in p$ **do**
        prod $\leftarrow$ prod $\times\, w_t(e_j \rightarrow e_k)$
      path_energies $\leftarrow$ path_energies $\cup$ {prod}
    $\Psi[e_i] \leftarrow \sum$ path_energies
  return Ψ
$\Psi \leftarrow$ ComputeInfluencePotential(G_t, w_t)
sorted_nodes $\leftarrow$ SortDescending(V_t, Ψ)
high_risk $\leftarrow$ sorted_nodes$[1 : \lceil k \cdot |V_t| \rceil]$
alerts $\leftarrow [\,]$
foreach *node* $e_i \in$ *high_risk* **do**
  $G_t^{\neg e_i} \leftarrow$ RemoveNode(G_t, e_i)
  $R_{\text{orig}} \leftarrow$ FindReachableNodes(G_t, e_i)
  broken $\leftarrow 0$
  foreach *node* $e_j \in R_{orig}$ **do**
    if *not IsReachable*$(G_t^{\neg e_i}, e_i, e_j)$ **then**
      broken $\leftarrow$ broken $+ 1$
  $\mathrm{DLI}(e_i) \leftarrow$ broken$/|V_t|$
  if $\mathrm{DLI}(e_i) > \theta$ **then**
    alert $\leftarrow \begin{cases} \text{node} : e_i, \\ \text{DLI} : \mathrm{DLI}(e_i), \\ \text{broken\_nodes} : \text{broken}, \\ \Psi : \Psi[e_i] \end{cases}$
    alerts $\leftarrow$ alerts $\cup$ {alert}
return alerts

network alarm events, device states, and structural paths. To transform this causal structure into a usable prediction capability, we design a downstream disconnection prediction mechanism, the core of which consists of root cause identification, risk scoring, and disconnection warning.

The primary goal of disconnection prediction is to identify the "source events" in the system, i.e., the root cause nodes of the current historical events that are most likely to trigger a wide range of impacts. Based on the constructed causal graph $G_t = (V_t, E_t)$, we use a weighted directed path analysis strategy to calculate the causal impact potential of each node:

$$\Psi\left(e_i\right) = \sum_{p \subset \mathcal{P}(e_i)} \prod_{(e_j \rightarrow e_k) \subset p} w_t\left(e_j \rightarrow e_k\right) \tag{7}$$

where $\mathcal{P}(e_i)$ denotes the set of all causal paths reachable from node e_i; $w_t(e_j \to e_k)$ is the causal weight of the edge on the current time window t; a larger value of $\Psi(e_i)$ indicates that the e_i event has a more substantial system-level influence and is prioritized as a candidate for a disconnected risk source.

After identifying the high-risk event nodes in the current window, the system enters the prediction mode to determine whether they will cause the downstream links or devices to be disconnected. At this stage, we introduce the idea of counterfactual reasoning. It estimates the question of "whether the occurrence/non-occurrence of node e_i will cause the disconnection of e_j" through the local perturbation simulation mechanism of the causal graph.

First, select the high-risk node e_i from the current graph G_t; second, construct its "counterfactual version" $G_t^{\neg e_i}$, i.e., set the node e_i did not occur, and recalculate its impact paths; if some nodes e_j are reachable by e_i in G_t but the paths break in $G_t^{\neg e_i}$, it means that $e_i \to e_j$constitutes a potential disconnected propagation.

We define the "Disconnection Risk Index" $\mathrm{DLI}(e_i)$ accordingly:

$$\mathrm{DLI}(e_i) = \frac{\left|\{\, e_j \in V_t : \; e_i \rightsquigarrow_{G_t} e_j \wedge e_i \not\rightsquigarrow_{G_t^{\neg e_i}} e_j \,\}\right|}{|V_t|} \tag{8}$$

This indicator reflects the potential systemic threat level of event e_i. Once $\mathrm{DLI}(e_i)$ exceeds the threshold θ, the system will trigger a risk warning. The flowchart of the approach is shown in Algorithm 1.

4. Experiments Settings

4.1. Data and Preprocessing

The data used in this experiment were collected from network operation and maintenance logs obtained from a large data center and several intelligent substations within a regional power grid. These datasets comprehensively reflect the operating characteristics of both IT infrastructure and industrial communication environments. The data consist of three main parts. The first part includes device-level status information such as CPU utilization, memory usage, interface error rate, BGP neighbor status, and link Up/Down events. The second part contains multi-source alarm events generated by protocols including Syslog, SNMP Trap, NetFlow, and BFD. The third part involves structured information derived from LLDP, configuration extraction, and controller APIs, including network topology and protocol relationship counts, physical topology, and protocol dependency graphs. An example has been shown in Table 1.

All data were uniformly processed and transformed into structured event representations in the form of ternary tuples $< e_i, M_i, T_i >$, and sliding event sequences were constructed within five-minute time windows for causal inference modeling based on the large language model.

4.2. Experimental Platform and Parameters

The hardware environment of the platform used in this experiment is a NVIDIA RTX A6000 GPU, the software environment is Python 3.10, and the primary LLMs tool is

Table 1. Example of network operation and maintenance data used in .

(A) Device Status Information						
Timestamp	Device	CPU(%)	Mem(%)	Interface	Error Rate	Link
10:23:01	RT-01	38	62	ge-0/0/1	0.2‰	Up
10:23:01	SW-22	12	48	eth1/3	1.5‰	Down
(B) Multi-source Alarm Events						
Timestamp	Device	Type	Severity	Description		
10:25:12	RT-01	Syslog	Warning	CRC errors on ge-0/0/1		
10:25:14	SW-22	SNMP Trap	Major	Link eth1/3 down		
(C) Topology and Protocol Dependencies						
Local Device	Local Port	Remote Device	Remote Port	Protocol Dependency		
RT-01	ge-0/0/1	SW-22	eth1/3	BGP $\rightarrow$ {BFD, OSPF}		

OpenAI GPT-4 API for causal extraction and interpretation verification, and the other tools are NetworkX for performing graph structure construction, and Neo4j for graph storage and interactive query.

The dataset size of this experiment comprises 10,000 windows, approximately 1,400 link failure events, and 15 abstract alarm types. LLMs' prompt templates are constructed based on alarm content, time, and neighbor structure, and batch reasoning and caching mechanisms are used to improve efficiency.

4.3. Metrics

To evaluate the model performance, we design the following key metrics:

(1) **Root Cause Accuracy:** whether the earliest cause event inferred by LLMs covers the actual faulty equipment or not;
(2) **Lead Time:** the average time for the first warning to be issued before the actual occurrence of a fault event;
(3) **Edge Precision:** the proportion of the edges predicted by LLMs for which there is a real causal relationship;
(4) **Edge Recall:** the proportion of real causal relationships correctly identified by LLMs;
(5) **Causal Explainability:** the proportion of edges that are validated by natural language feedback;
(6) **False Positive Rate:** the proportion of non-root-cause devices that are misidentified as risk sources.

5. Results and Analysis

This section aims to validate the effectiveness and practicality of the proposed network device disconnection prediction approach that integrates LLMs and causal reasoning in real scenarios. We simulate typical device failure scenarios by constructing multi-source

Table 2. Experimental results demonstrating how sample size influences the accuracy, interpretability, and computational cost of the proposed causal graph generation approach, with estimated standard deviations.

Sample	Edges Extracted	Precision	Recall	Explainability	Processing Time (s)
500	84 ± 4	0.74 ± 0.03	0.69 ± 0.03	0.72 ± 0.03	3.4 ± 0.2
1,000	110 ± 4	0.81 ± 0.02	0.74 ± 0.02	0.76 ± 0.02	5.8 ± 0.3
5,000	138 ± 3	0.86 ± 0.02	0.82 ± 0.02	0.84 ± 0.02	13.5 ± 0.5
10,000	142 ± 3	0.88 ± 0.01	0.85 ± 0.01	0.87 ± 0.01	22.9 ± 0.7
50,000	146 ± 2	0.90 ± 0.01	0.87 ± 0.01	0.91 ± 0.01	88.4 ± 2.0
100,000	148 ± 1	0.91 ± 0.01	0.88 ± 0.01	0.92 ± 0.01	174.1 ± 3.0

Table 3. Comparison of root-cause localization performance across baseline approaches and the proposed LLMs-based approach, including estimated standard deviations for accuracy, prediction lead time, and false positive rate.

Approaches	Root Cause Accuracy ↑	Prediction Lead Time (min) ↑	False Positive Rate ↓
NetRCA	$75.0\% \pm 1.5\%$	3.0 ± 0.3	0.150 ± 0.010
REASON	$81.5\% \pm 1.2\%$	5.0 ± 0.4	0.120 ± 0.008
RUN	$83.0\% \pm 1.0\%$	5.5 ± 0.3	0.110 ± 0.007
Attention-LSTM	$89.5\% \pm 0.8\%$	-	-
Ours	$\mathbf{90.2\% \pm 0.7\%}$	$\mathbf{7.1 \pm 0.4}$	$\mathbf{0.074 \pm 0.005}$

alarm flow datasets in real network environments. We systematically evaluate the quality of causal graph construction, root cause localization accuracy, alerting capability, and module effectiveness.

5.1. Quality Assessment of Causal Graphs Evaluation

We input samples of different sizes into the LLMs to evaluate their performance in automatically generating causal edges.

The results in Table 2 show that larger sample sizes substantially enhance the model's ability to extract causal edges: the number of identified relations increases from 84 to 148. It begins to stabilize once the sample size exceeds 50,000. Precision and recall also improve steadily (from 0.74/0.69 to 0.91/0.88), indicating clear gains in both accuracy and coverage. Complementing these numerical results, Fig. 5 visualizes the same trends and highlights the smooth, monotonic improvement of all metrics as data scale grows. The curves gradually flatten at larger sample sizes, illustrating the diminishing incremental benefits and the model's convergence behavior.

Meanwhile, the causal interpretation rate also increases from 0.72 to 0.92, indicating that the causal edges generated by LLMs are increasingly rational and semantically consistent, providing strong support for interpretability in practical applications. Although the processing time increases significantly with the sample size, it is within the acceptable range, indicating that the approach has good scalability while maintaining the quality of graph building and is suitable for large-scale network disconnection prediction scenarios.

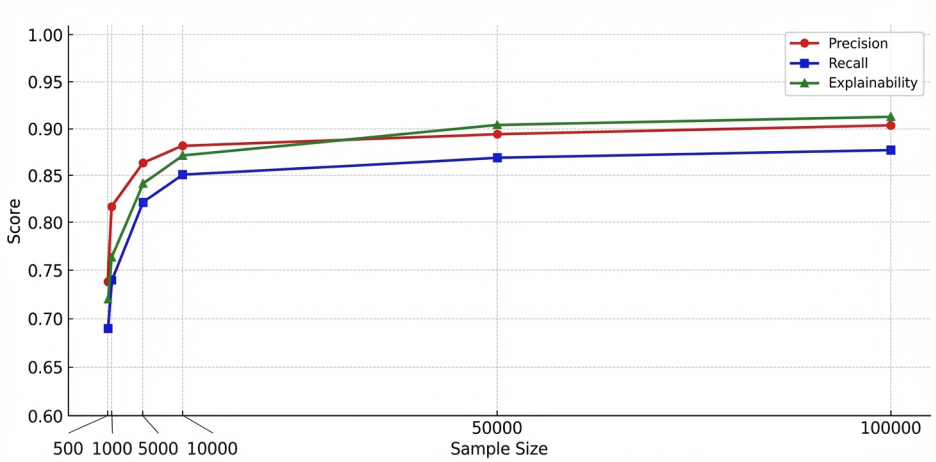


Fig. 5. Causal Graph Construction Performance Evaluation

The comparison in Fig. 6 verifies that the larger the sample size, the more stable and complete the causal structure becomes, which helps to improve the accuracy of root cause localization and the explanation of disconnected propagation paths.

5.2. Root Cause Localization Performance Evaluation

To comprehensively evaluate the performance of the proposed approach in root cause localization and disconnection risk prediction, we select four representative studies as baseline approaches, including traditional feature-based ensemble models, graph neural networks, neural Granger causal discovery, and generative adversarial frameworks:

(1) **NetRCA**[26]: A root cause localization approach that integrates multi-type feature engineering, data augmentation, and ensemble learning (including XGBoost, rule-based learning, and graph algorithms). It is primarily designed for network failure scenarios with limited samples.
(2) **REASON**[20]: A hierarchical graph neural network that constructs cross-layer causal structures and combines random walk with extreme value theory for root cause identification. It is particularly suited for analyzing multi-level fault propagation.
(3) **RUN**[13]: A approach based on neural Granger causal discovery and contrastive learning, further enhanced with PageRank to localize root causes of microservice failures. Experiments on both synthetic and real microservice datasets demonstrate superior performance over traditional approaches.
(4) **FaultGuard**[8]: A generative adversarial framework tailored for smart grids, which exhibits high robustness and ensures accurate disconnection prediction and detection even under adversarial attacks.

As shown in Table 3, the proposed approach achieves the best performance across all three key evaluation metrics. First, in terms of **root cause localization accuracy**, our approach attains 90.2%, outperforming RUN (83.0%) and REASON (81.5%), and

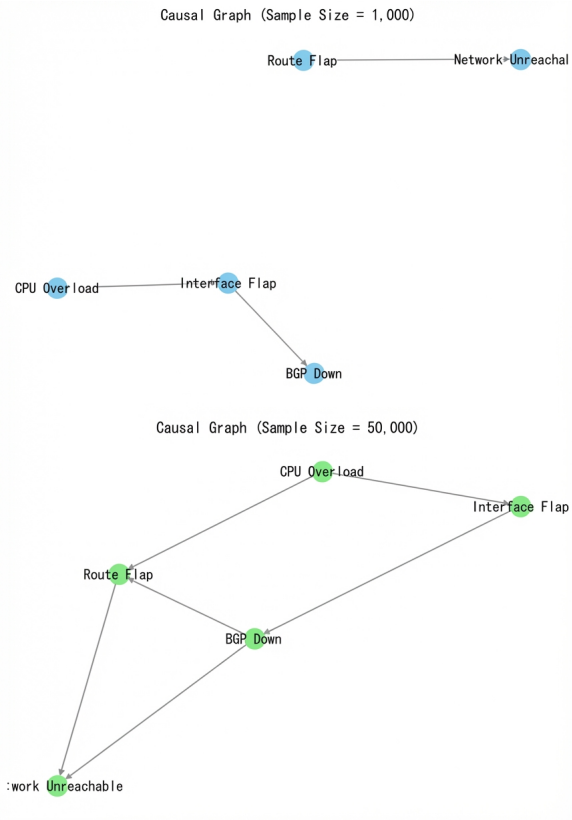


Fig. 6. Effect of sample size on the stability and completeness of the causal structure, which improves root cause localization and the explanation of disconnected propagation paths.

slightly surpassing FaultGuard (89.5%). This demonstrates the superior capability of our approach in handling complex network failure scenarios. Second, regarding **prediction timeliness**, the approach issues warnings approximately 7.1 minutes before disconnection events occur, significantly ahead of REASON and RUN (5.0–5.5 minutes), thereby providing a longer response window for preventive maintenance. Third, in terms of **false alarm control**, our approach achieves a false positive rate of only 7.4%, substantially lower than REASON (12%), RUN (11%), and NetRCA (15%). This highlights the advantage of the causal reasoning module in mitigating "language hallucinations" and reducing noise-induced misjudgments.

Overall, these results clearly demonstrate the benefits of deeply integrating LLMs with causal modeling: the proposed approach achieves significant breakthroughs in root cause localization accuracy, responsiveness, and predictive reliability, suggesting its potential to serve as a new paradigm for preventive maintenance in smart grids and critical communication systems.

Table 4. Comprehensive ablation study showing the impact of individual components and their combinations on model performance. SCC, NLE, and PO denote Structural Consistency Check, Natural Language Explanation, and Prompt Optimization, respectively. RCA, PLT, and FPR represent Root Cause Accuracy, Prediction Lead Time, and False Positive Rate, with estimated standard deviations.

Configuration	SCC	NLE	PO	RCA ↑	PLT (min) ↑	FPR ↓
Full Model	Y	Y	Y	**90.2% ± 0.7%**	**7.1 ± 0.4**	**0.074 ± 0.005**
w/o Consistency	N	Y	Y	80.1% ± 1.2%	6.6 ± 0.4	0.120 ± 0.010
w/o Explanation	Y	N	Y	83.4% ± 1.0%	6.9 ± 0.3	0.103 ± 0.008
w/o Prompt Optimization	Y	Y	N	78.0% ± 1.5%	5.3 ± 0.5	0.131 ± 0.012
w/o SCC + NLE	N	N	Y	75.8% ± 1.3%	6.2 ± 0.4	0.135 ± 0.012
w/o SCC + PO	N	Y	N	72.5% ± 1.5%	5.8 ± 0.5	0.142 ± 0.013
w/o NLE + PO	Y	N	N	77.1% ± 1.4%	5.9 ± 0.4	0.128 ± 0.011
w/o SCC + NLE + PO	N	N	N	68.0% ± 1.8%	5.0 ± 0.5	0.155 ± 0.015

These results demonstrate that the introduction of LLMs into network disconnection prediction and root cause localization tasks can significantly enhance the interpretability and foresight of the models, providing a more practical technical path for risk prevention and control in highly available network systems.

5.3. Disconnection Propagation Path Analysis

To further validate the explanatory ability and causal path modeling effect of the large model in the disconnection prediction task, we selected two typical types of network failure events from the experimental set, demonstrated the causal propagation paths automatically constructed by the model, and analyzed them in combination with the network operation data and topology.

In a particular core switching cluster, the physical layer fluctuations occur continuously on the device interfaces, and the model automatically generates the following causal chain:

```
[Interface Flap] → [OSPF Adjacency Lost] → [Route
       Withdrawal] → [Service Unreachable]
```

The explanatory text in the path states:

```
"OSPF adjacency outages are usually caused by physical link
          instability or interface restarts."
    "Route convergence failures cause service paths to
    disappear, triggering service unreachable alarms."
```

In the edge access device. The system monitors a continuous spike in resource utilization. Eventually, the user application experiences frequent timeouts. The causal path generated by LLMs is as follows:

```
[CPU Overload] → [Routing Loop Detected] → [Packet Loss] →
                  [Application Timeout]
```

The path reveals the complete mechanism that begins with the underlying device state anomaly and gradually leads to the control plane, forwarding path, and ultimately affects the application layer. The explanatory text in the path states:

```
"High CPU load may lead to delays in processing control
  packets, which generates routing convergence jitter"
"Routing loops trigger forwarding congestion and packet
 loss, which ultimately affects upper-layer application
                    response times."
```

Similar resource bottleneck paths are frequently identified in multiple experimental samples, demonstrating the stability and generality of the LLMs approach to capture the propagation of the state protocol-data-service chain.

5.4. Ablation Experiments

To further assess the role of each module in the overall approach, we designed three sets of ablation experiments, excluding key components such as large model cue optimization, structural consistency checking, and natural language causal validation, and observing their impact on root cause prediction ability.

The results presented in Table 4 provide clear evidence of each component's contribution. Removing the Structural Consistency Check (SCC) results in a substantial increase in the false positive rate, underscoring its crucial role in maintaining prediction reliability. Excluding the Natural Language Explanation (NLE) module reduces Root Cause Accuracy (RCA). It slightly shortens the Prediction Lead Time (PLT), demonstrating that semantic validation enhances both interpretability and early-warning capability. Omitting Prompt Optimization (PO) causes the largest drop in RCA, reflecting the importance of carefully designed prompts for stable LLMs' reasoning.

Furthermore, the joint ablation experiments, where two or all three components are removed simultaneously, show a compounded deterioration in performance: RCA declines sharply, PLT decreases, and FPR rises markedly. This highlights the synergistic effect among the modules, where each component not only contributes individually but also reinforces the others to improve overall model robustness and reliability.

Overall, these ablation studies confirm that each sub-module in the proposed approach is indispensable. The SCC ensures low false positives, the NLE provides causal interpretability, and PO stabilizes LLMs' outputs, together achieving enhanced accuracy, interpretability, and robustness of the disconnection prediction system.

6. Conclusion

In this paper, we propose a network device disconnection prediction approach that integrates large language models (LLMs) with causal analysis. The approach leverages structured event abstraction, combines network topology and operational state information, employs LLMs for causality extraction, and is enhanced with consistency checking and causal path interpretation, enabling accurate root-cause identification, interpretable reasoning, and early warning capabilities.

Empirical evaluation of real data center operation and maintenance logs reveals that the proposed approach outperforms traditional rule-based and statistical approaches in terms of causal graph construction quality, root-cause localization accuracy, prediction reliability, and false alarm reduction. These results demonstrate the practical utility of the framework for proactive network management and preventive maintenance in mission-critical infrastructures.

Future work will focus on adaptive prompt engineering, more efficient incremental causal graph construction, improving LLMs' generalization across diverse network environments, and enhancing model security and robustness. With ongoing technological advancements, integrating LLMs-based causal reasoning is expected to become a foundational capability for intelligent, resilient, and automated network operation and maintenance.

Acknowledgment. This work was supported by Science and Technology Project of State Grid Jiangsu Electric Power Co., Ltd., and Science and Technology Project of Suzhou Suneng Group Co., Ltd. under Grant SGSZSNJTKJJS2500967.

References

1. Alkaberi, W., Assiri, F.: Predicting the number of software faults using deep learning. Engineering, Technology & Applied Science Research 14(2), 13222–13231 (Apr 2024)
2. Alvarez-Alvarado, M., Donaldson, D., Recalde, A., Khan, Z., Velasquez, W., Rodríguez Gallegos, C.: Power system reliability and maintenance evolution: A critical review and future perspectives. IEEE Access 10, 51922–51950 (01 2022)
3. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., Xie, X.: A survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol. 15(3) (Mar 2024)
4. Chen, W., Cai, R., Zhang, K., Hao, Z.: Causal discovery in linear non-gaussian acyclic model with multiple latent confounders. IEEE Transactions on Neural Networks and Learning Systems 33, 2816–2827 (2021)
5. Chen, Y., Xie, H., Ma, M., Kang, Y., Gao, X., Shi, L., Cao, Y., Gao, X., Fan, H., Wen, M., Zeng, J., Ghosh, S., Zhang, X., Zhang, C., Lin, Q., Rajmohan, S., Zhang, D., Xu, T.: Automatic root cause analysis via large language models for cloud incidents. In: Proceedings of the Nineteenth European Conference on Computer Systems. p. 674–688. EuroSys '24, Association for Computing Machinery, New York, NY, USA (2024)
6. Cheng, X., Lv, K., Zhang, Y., Wang, L., Zhao, W., Liu, G., Qiu, J.: Rul prediction method for electrical connectors with intermittent faults based on an attention-lstm model. IEEE Transactions on Components, Packaging and Manufacturing Technology 13, 628–637 (2023)
7. Cohrs, K.H., Diaz, E., Sitokonstantinou, V., Varando, G., Camps-Valls, G.: Large language models for constrained-based causal discovery. In: AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?". pp. 1–9 (2023)
8. Efatinasab, E., Marchiori, F., Brighente, A., Rampazzo, M., Conti, M.: Faultguard: A generative approach to resilient fault prediction in smart electrical grids. In: Detection of Intrusions and Malware, and Vulnerability Assessment: 21st International Conference, DIMVA 2024, Lausanne, Switzerland, July 17–19, 2024, Proceedings. p. 503–524. Springer-Verlag, Berlin, Heidelberg (2024)
9. Ghosh, S.S., Dwivedi, A., Tajer, A., Yeo, K., Gifford, W.M.: Cascading failure prediction via causal inference. IEEE Transactions on Power Systems 40(4), 3361–3373 (2025)

10. Gupta, M., Rajnish, K., Bhattacharjee, V.: Impact of parameter tuning for optimizing deep neural network models for predicting software faults. Sci. Program. 2021, 6662932:1–6662932:17 (2021)
11. Laborda, J.D., Torrijos, P., Puerta, J.M., Gámez, J.A.: Parallel structural learning of bayesian networks: Iterative divide and conquer algorithm based on structural fusion. Knowledge-Based Systems 296, 111840–111858 (2024), https://www.sciencedirect.com/science/article/pii/S095070512400474X
12. Li, M., Li, Z., Yin, K., Nie, X., Zhang, W., Sui, K., Pei, D.: Causal inference-based root cause analysis for online service systems with intervention recognition. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 3230–3240. KDD '22, Association for Computing Machinery, New York, NY, USA (2022)
13. Lin, C.M., Chang, C., Wang, W.Y., Wang, K.D., Peng, W.C.: Root cause analysis in microservice using neural granger causal discovery. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. pp. 206–213. AAAI'24/IAAI'24/EAAI'24, AAAI Press (2024)
14. Liu, X., Chen, B., Chen, C., Jin, D.: Electric power grid resilience with interdependencies between power and communication networks – a review. IET Smart Grid 3(2), 182–193 (2020)
15. Lv, F., Si, S., Xiao, X., Ren, W.: Modified local granger causality analysis based on peter-clark algorithm for multivariate time series prediction on iot data. Computational Intelligence 40, 1–20 (2024)
16. Nazaret, A., Blei, D.: Extremely greedy equivalence search. In: Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence. pp. 2716–2745. UAI '24, JMLR.org (2024)
17. Raja, E.J., Hossen, J., Ervina, E.M.N., Tawsif, K., Jesmeen, M.Z.H.: Broadband network fault prediction using complex event processing and predictive analytics techniques. Journal of Engineering Science and Technology 15(4), 2289–2300 (2020)
18. Rana, S.: Ai-driven fault detection and predictive maintenance in electrical power systems: A systematic review of data- driven approaches, digital twins, and self-healing grids. American Journal of Advanced Technology and Engineering Solutions 01, 258–289 (02 2025)
19. Wan, G., Lu, Y., Wu, Y., Hu, M., Li, S.: Large language models for causal discovery: Current landscape and future directions. In: Kwok, J. (ed.) Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. pp. 10687–10695. International Joint Conferences on Artificial Intelligence Organization (8 2025), https://doi.org/10.24963/ijcai.2025/1186, survey Track
20. Wang, D., Chen, Z., Ni, J., Tong, L., Wang, Z., Fu, Y., Chen, H.: Interdependent causal networks for root cause localization. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 5051–5060. KDD '23, Association for Computing Machinery, New York, NY, USA (2023)
21. Wang, F., Jiang, Y., Zhang, R., Wei, A., Xie, J., Pang, X.: A survey of deep anomaly detection in multivariate time series: Taxonomy, applications, and directions. Sensors 25(1), 1–27 (2025)
22. Wang, X., Jiang, S., Li, X., Wang, M.: Causal discovery and fault diagnosis based on mixed data types for system reliability modeling. Complex & Intelligent Systems 11, 1–16 (2025)
23. Yang, Q., Hao, W., Ge, L., Ruan, W., Chi, F.: Farima model-based communication traffic anomaly detection in intelligent electric power substations. IET Cyper-Phys. Syst.: Theory & Appl. 4, 22–29 (2018)
24. Yang, T.L., Lee, K.Y., Zhang, K., Suzuki, J.: Functional linear non-gaussian acyclic model for causal discovery. Behaviormetrika 51(2), 567–588 (2024)
25. Yang, Y., Nishikawa, T., Motter, A.E.: Small vulnerable sets determine large network cascades in power grids. Science 358(6365), 1–8 (2017)

26. Zhang, C., Zhou, Z., Zhang, Y., Yang, L., He, K., Wen, Q., Sun, L.: Netrca: An effective network fault cause localization algorithm. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 9316–9320 (2022)
27. Zhang, D., Zhang, X., Bansal, C., Las-Casas, P., Fonseca, R., Rajmohan, S.: Lm-pace: Confidence estimation by large language models for effective root causing of cloud incidents. In: Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering. p. 388–398. FSE 2024, Association for Computing Machinery, New York, NY, USA (2024), https://doi.org/10.1145/3663529.3663858
28. Zhang, X., Ghosh, S., Bansal, C., Wang, R., Ma, M., Kang, Y., Rajmohan, S.: Automated root causing of cloud incidents using in-context learning with gpt-4. In: Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering. p. 266–277. FSE 2024, Association for Computing Machinery, New York, NY, USA (2024)
29. Zhang, Z., Li, S., Zhang, L., Ye, J., Hu, C., Yan, L.: Llm-lade: Large language model-based log anomaly detection with explanation. Knowledge-Based Systems 326, 114064–114074 (2025), https://www.sciencedirect.com/science/article/pii/S0950705125011098

Jie Ren is a Senior Engineer at Suzhou Suneng Group Co., LTD of China. He has rich experience on power grid technology and application. His research interests include flow control and optimization, electric power information network, Quality of Service, and artificial intelligence.

Xiaojun Yao is a Senior Engineer at State Grid Suzhou Power Supply Company of China and State Grid Suzhou Power Supply Company of China. He is majored in power grid technology and application. His research interests include computer network, electric power information network Quality of Service, intelligent operation and maintenance, artificial intelligence, and machine learning.

Hong Chen is an Engineer Suzhou Suneng Group Co., LTD of China. He is experienced in power grid intelligent operation and maintenance. His research interests include substation network, flow control and optimization, and artificial intelligence.

Received: October 22, 2025; Accepted: January 5, 2026.

Computer Science and Information Systems 23(2):729–756 https://doi.org/10.2298/CSIS251018019X

YOLO-BDM: An Improved Ship Detection Algorithm Based on YOLOv11n

Fangyuan Xiong[1], Dezhi Han[2], Xiang Shen[3], and Manlin Zhu[1]

[1] School Of Information Engineering, Shanghai Maritime University,1550 Haigang Avenue, Shanghai, 201306, Shanghai, China
xiongfangyuan@stu.shmtu.edu.cn (corresponding author)
[2] Shanghai Ship and Shipping Research Institute Co., Ltd., Shanghai, 200135, China
dzhan@shmtu.edu.cn
[3] School of Computer Science, The University of Sydney, Sydney, New South Wales, Australia
shenxiang1107@163.com

Abstract. Synthetic Aperture Radar (SAR) ship detection is crucial for maritime traffic management, search and rescue, and environmental monitoring but remains challenging due to small targets, blurred contours, and complex ocean backgrounds. To address these issues, this paper proposes YOLO-BDM, an improved detector based on YOLOv11. The Diverse Branch Block (DBB) is introduced into the backbone to enhance feature representation through multi-branch training and reparameterized inference. A Multi-scale Contextual Attention (MCA) mechanism is integrated into the backbone and neck to strengthen multi-scale semantic modeling and background discrimination. Additionally, a four-layer Bidirectional Feature Pyramid Network (BiFPN) is employed for efficient multi-scale feature fusion. Experiments on the SAR-Ship dataset show YOLO-BDM achieves 97.27% mAP, 94.11% Precision, and 93.07% Recall, surpassing the baseline and validating its effectiveness.

Keywords: Ship detection; BiFPN module; YOLOv11n; DBB module; MCA attention mechanism

1. Introduction

Object detection, as a core problem in computer vision, has garnered extensive attention in recent years across diverse application scenarios including intelligent transportation [38], medical imaging [17], industrial automation [10], and remote sensing image processing [35]. Synthetic Aperture Radar (SAR), with its all-weather, all-time imaging capabilities and robustness in complex environments, has emerged as a vital tool for ocean monitoring and vessel detection [37]. However, ship targets in SAR images often present challenges such as small size, blurred contours, complex backgrounds, and speckle noise interference [20], making high-precision detection difficult. Achieving robust and efficient SAR ship detection in complex marine environments has thus become a critical research topic in intelligent remote sensing perception, holding significant theoretical and practical value.

Early SAR ship detection methods primarily relied on manually designed features and shallow classifiers. However, constrained by noise interference and single-scale feature modeling, their robustness and generalization capabilities were limited [31]. To overcome this bottleneck, researchers gradually shifted toward detection frameworks driven

by convolutional neural networks (CNNs) [46]. Regarding two-stage detectors, the R-CNN series proposed by Girshick et al. laid the foundation for end-to-end detection. Ren et al. introduced the Region Proposal Network (RPN) through Faster R-CNN, significantly enhancing candidate box generation and detection efficiency. Subsequently, Lin et al. proposed the Feature Pyramid Network (FPN) [24], strengthening multi-scale feature representation; Cai and Vasconcelos proposed Cascade R-CNN [1], improving detection accuracy at high IoU thresholds through stepwise optimization; Pang et al. introduced Libra R-CNN [27], achieving improvements in sample allocation and feature utilization. While these methods excel in detection accuracy and semantic modeling, their slow inference speeds and high computational complexity hinder real-time detection requirements.

Against this backdrop, single-stage detectors have gradually emerged as a research hotspot. Frameworks such as YOLO [28] and SSD [25] have achieved significant improvements in inference speed while maintaining reasonable detection accuracy. However, early single-stage methods exhibited limitations in modeling complex textures and representing multi-scale features, particularly in low signal-to-noise ratio conditions where small objects were prone to detection failures. To address this, Dai et al. introduced Deformable Convolutions (DCN) [9], enhancing the adaptability of convolutions to geometric deformations and intricate textures. Hu et al. proposed Channel Attention Mechanisms [18], improving model discrimination in complex backgrounds through feature weighting. Driven by these advancements, SAR vessel detection has seen further development. For instance, Ma et al. proposed a free-bounding box detection method based on keypoint estimation and attention mechanisms, effectively suppressing false alarms in complex backgrounds [26]. However, it still suffers from insufficient discrimination between adjacent targets in high-density scenes due to ambiguous keypoint matching. Zhao et al. introduced the CRAS-YOLO model [43], achieving high-precision detection and classification of multi-category vessels, though its category coverage remains limited. Zhou et al. proposed the FGNet model [44], integrating a global context module and multi-scale feature enhancement module to improve target discrimination and multi-scale feature representation in complex scenes. Nevertheless, false negatives and false positives may still occur in strongly cluttered coastal environments. Overall, although notable progress has been achieved in SAR ship detection, existing methods still encounter intrinsic limitations when applied to complex marine environments [30], In particular, most YOLO-based detectors adopt conventional multi-scale feature fusion strategies that inadequately address semantic misalignment across feature levels, leading to suboptimal performance in small and densely distributed ship detection tasks. This issue is especially pronounced in SAR imagery, where speckle noise and weak target boundaries further degrade the reliability of shallow feature representations.

In summary, although existing methods have made some progress in SAR ship detection tasks, they still face numerous challenges in multi-scale modeling, feature focusing, and model deployment. This limitation is particularly pronounced in SAR ship detection scenarios involving small targets, where speckle noise and weak object boundaries further impair the reliability of shallow feature representations. Fundamentally, this problem stems from the insufficient correction of semantic bias during multi-scale feature fusion and the lack of effective contextual modeling to align features across different scales [5]. Furthermore, existing attention-enhanced YOLO-style architectures often focus on either channel-wise or spatial-wise feature modulation in isolation, lacking the capacity for joint

contextual modeling across multiple dimensions [4]. Finally, in pursuit of accuracy, certain detection frameworks incorporate extensive convolutional stacking and redundant modules, resulting in structurally complex architectures with substantial parameter scales. This hinders efficient deployment on edge computing or resource-constrained platforms. Consequently, achieving context-enhanced, lightweight modelling while maintaining detection precision has become an urgent research priority.More fundamentally, these limitations reflect a common design paradigm in existing YOLO-based SAR ship detectors, in which multi-scale feature fusion, attention-driven feature focusing, and structural efficiency are typically optimized independently rather than within a unified framework. Consequently, achieving a balanced integration of contextual representation enhancement, precise feature discrimination, and deployment-friendly efficiency remains an open challenge.

To address these challenges, this paper proposes the improved YOLO-BDM model based on the YOLOv11n framework. It aims to achieve precise feature extraction of small targets in complex SAR scenes, efficient fusion of multi-scale information, and suppression of background interference. Specific contributions include:

(1) The Diverse Branch Block (DBB) is introduced into the C3K2 architecture of the backbone network. During training, this multi-branch convolutional structure enriches feature representations. At inference time, it is equivalently transformed into a single convolutional layer through structural reparameterization. This approach balances model expressiveness with inference efficiency while mitigating the issue of edge weakening in small targets within SAR images.

(2) Embedding a Multi-scale Contextual Attention (MCA) mechanism at key nodes of the backbone and neck structures. This combines global average pooling with standard deviation pooling to extract multi-scale contextual information. Channel-wise dynamic weighting enhances responses in critical regions, effectively suppressing background noise interference and improving discrimination capabilities in complex environments;

(3) Introducing the BiFPN_Concat module into the neck network. It enhances interaction between shallow-layer details and deep-layer semantics through bidirectional feature propagation and learnable weight mechanisms. Simultaneously, feature concatenation preserves richer original information, improving flexibility in multi-scale modeling. This approach is particularly effective for detecting ship targets with significant scale variations.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed model and the techniques employed. Section 4 provides a detailed analysis of the experimental results. Section 5 concludes the paper and discusses future research directions.

2. Related Work

2.1. Traditional Ship Detection Algorithms

In SAR or optical imagery, traditional ship detection methods primarily rely on manually designed features and classical algorithms, typically including threshold segmentation, texture features, or keypoint descriptors. These methods are simple to implement,

computationally efficient, and can effectively detect targets in straightforward scenarios. However, their performance is limited in complex backgrounds, when dealing with small, multi-scale targets, or under low signal-to-noise ratio conditions [8]. To address these limitations, Constant False Alarm Rate (CFAR)-based detection methods have gained widespread application in complex maritime clutter environments. By adaptively estimating background noise, dynamically adjusting detection thresholds, suppressing false alarms, and maintaining high detection rates, CFAR-based approaches have become a key direction for improving traditional ship detection research [40].

However, traditional CFAR methods still suffer from limitations such as high computational complexity and low efficiency when processing high-resolution SAR images or large-scale scenes. Wang et al. [36] proposed a fast CFAR algorithm based on density screening (DC). By rapidly eliminating high-density background clutter using superpixel density features, it performs local detection only on a small number of candidate targets. This approach reduces computational complexity by 75%–96% while maintaining or even improving detection accuracy, effectively suppressing false alarms. It represents a significant breakthrough in balancing CFAR detection efficiency and accuracy. Furthermore, for complex coastal environments, Li et al. [23] proposed a hierarchical detection scheme for airborne single-channel SAR. This approach employs a K-log-normal mixture distribution model with adaptive background windows for CFAR prescreening, enhancing the resolution between sea clutter and targets. Subsequently, it introduces fine-grained discrimination based on micro-Doppler motion characteristics, further suppressing false alarms through radial velocity and image entropy analysis. This work demonstrates that traditional CFAR methods can achieve high-precision, low-false-alarm real-time detection on lightweight platforms by integrating statistical modeling with motion features. Despite CFAR's excellence in adaptive background estimation and false alarm suppression, issues of fitting inaccuracy and efficiency limitations persist in heterogeneous backgrounds and complex nearshore scenarios. To further address these challenges, Chen et al. [7] proposed a multi-modal saliency-based (MMS) vessel detection method. By integrating enhanced CFAR, superpixel (MSER), local stability analysis, and sea-land segmentation, they constructed four complementary saliency maps, effectively resolving fitting inaccuracies and oversegmentation in heterogeneous backgrounds.

Overall, the limitations of traditional ship detection methods—such as high false alarm rates, inaccurate fitting in complex backgrounds, and oversegmentation—have been significantly mitigated through continuous improvements by previous researchers. However, even enhanced traditional ship detection algorithms still struggle with challenges like detecting extremely small targets, handling multi-scale ships, and processing large-scale high-resolution images [16]. This indicates that addressing the practical challenges of SAR vessel detection requires not only innovation at the algorithmic level, but also consideration of reliability, manageability, and overall system performance at the architectural level [15].Furthermore, the advancement of deep learning techniques, such as convolutional neural networks (CNNs), has provided effective means for automatic feature extraction, multi-scale modeling, and end-to-end training. These developments have collectively propelled the evolution of deep learning-based ship detection algorithms, gradually establishing them as the mainstream approach in this field.

2.2. Deep Learning-Based Ship Detection Algorithms

Deep learning-based ship detection algorithms leverage convolutional neural networks (CNNs) to automatically extract hierarchical features, enabling end-to-end target localization and classification. Compared with traditional methods, such approaches significantly improve detection performance in complex backgrounds, for small-scale and multi-scale targets, and in high-resolution SAR imagery.

Existing deep learning-based ship detection studies can be broadly analyzed from different technical perspectives, including multi-scale feature modeling and localization accuracy, contextual attention and semantic discrimination, as well as lightweight and efficient architectural design. From an implementation standpoint, these methods are commonly realized through two-stage or single-stage detection frameworks, each emphasizing different trade-offs between accuracy and efficiency. Together, these advances have laid a solid foundation for continuous performance improvements in SAR ship detection [11].

Multi-scale Modeling and Localization-Oriented Methods Accurate localization of vessels across varying scales is a fundamental challenge in SAR ship detection, particularly in dense maritime scenes and complex cluttered backgrounds. To address this issue, many studies emphasize multi-scale feature modeling and precise localization, with two-stage detection frameworks serving as a representative implementation due to their explicit region proposal and refinement mechanisms.

Zhou et al. proposed UltraHi-PrNet [45], which improves feature alignment across scales via scale transfer and expansion layers but suffers from the computational inefficiency and generalization limits of its Faster R-CNN backbone. Tang et al.'s PEGNet [32], enhances Faster R-CNN with modules for better multi-scale fusion and noise suppression, yet its horizontal anchor design limits effectiveness for rotated targets in dense scenes. To address rotation issues, Zhang et al. proposed ORPSD [41]. using an outer rectangular projection scheme, though it retains the high computational cost typical of two-stage detectors.

In summary, multi-scale modeling and localization-oriented methods achieve high detection accuracy by explicitly refining candidate regions and integrating scale-aware features. However, the computational inefficiency and limited real-time performance of two-stage frameworks remain unresolved challenges, particularly for large-scale or resource-constrained SAR applications [6].

Contextual Attention and Semantic Discrimination Methods To address the limitations of pure multi-scale modeling in capturing long-range dependencies and discriminative features, researchers have developed methods that explicitly incorporate contextual attention and semantic enhancement mechanisms. A fundamental challenge for single-stage detectors has been to balance fine-grained feature interaction with high speed, a trade-off inherently linked to model stability and flexibility [29]. The evolution of frameworks like the YOLO series reflects this ongoing pursuit.

Representative advances in this direction include YOLOv4 [42],which introduced a CSPDarknet backbone to reduce computational redundancy and enhance feature diversity

through a split-and-fusion strategy. However, its reliance on anchor boxes limits generalization in scenes with significant scale and aspect ratio variations. To overcome such limitations, Hu et al. proposed BANet [19], an anchor-free design that integrates Local and Non-Local Attention Modules. This architecture improves fine-grained modeling of multi-angle vessels and contextual reasoning in complex backgrounds, though preserving fine texture details for small targets remains challenging.

In summary, methods focusing on contextual attention and semantic discrimination effectively address the limitations of conventional multi-scale approaches by emphasizing spatial context and fine-grained feature interactions. These techniques enhance detection robustness in complex maritime scenes, particularly for vessels with diverse orientations and subtle features, laying the groundwork for subsequent improvements in lightweight and efficient network architectures.

Lightweight and Efficient Single-Stage Detection Methods To address the challenges of deploying high-performance SAR ship detection models under resource constraints, research has focused on developing lightweight and efficient single-stage network architectures. Representative methods include: SSD-YOLO [12], an anchor-free framework enhanced with a multidimensional feature module to sharpen small target boundaries while maintaining real-time performance, though it struggles with complex contexts and fine details for very small or clustered vessels. FD-Net [13], incorporates deformable convolutions across the network, combined with an Enhanced Feature Pyramid and adaptive fusion module, to better represent vessels of varying scales and shapes, but its semantic integration and real-time efficiency are limited. LKE-Det [11]employs a decomposable large kernel to capture long-range dependencies and embeds edge gradient features to improve contour delineation and suppress clutter, yet efficient multi-scale fusion for subtle targets remains a challenge. RepGFPN [2], introduces efficient cross-layer connections and bidirectional fusion to aggregate shallow and deep features directly, enhancing detection of small and coastal targets, though preventing feature degradation during deep propagation is still a core issue [14].

In summary, existing lightweight and efficient architecture methods have achieved notable improvements in balancing computational cost, model complexity, and detection accuracy. Nevertheless, current approaches still face challenges in fully integrating multi-scale features, preserving fine-grained vessel details, and capturing rich contextual information under complex maritime conditions [3]. To address these limitations, we propose the YOLO-BDM vessel detection model, which combines enhanced multi-scale feature fusion, semantic discrimination, and lightweight design to achieve robust, real-time performance for SAR ship detection.

3. Methods

3.1. Baseline Model YOLOv11n

YOLOv11n is a lightweight variant of the YOLO series proposed in recent years, designed to improve object detection accuracy and stability while reducing computational costs. Compared to its predecessors, YOLOv11n systematically optimizes both the backbone and neck networks, notably incorporating modules such as C3K2 and C2PSA to enhance

feature extraction and fusion capabilities. As shown in Figure 1, its overall architecture primarily consists of a CSP-based backbone network, the C3K2 module, and the C2PSA attention mechanism, achieving a superior balance between detection performance and inference efficiency.

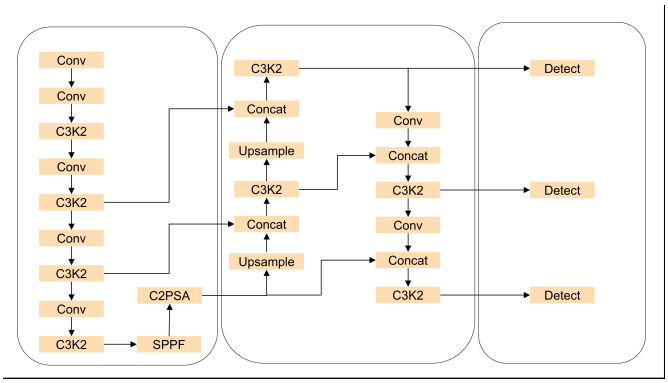


Fig. 1. YOLOv11n network architecture

In the backbone network, the C3K2 module proposed in YOLOv11n enhances the feature extraction architecture through two key innovations: dual-path residual connections and a dynamic receptive field mechanism. The dual-path design preserves lightweight characteristics while introducing a deformable convolution branch. The primary path retains standard convolutions to ensure computational efficiency, whereas the novel secondary path equips the model with stronger adaptability to multi-scale objects. The dynamic receptive field mechanism further enhances feature representation by leveraging multi-scale deformable convolutions. Moreover, unlike the channel compression strategies commonly employed in mainstream lightweight solutions, C3K2 adopts feature reorganization techniques to better preserve critical spatial information. The C3K2 module is illustrated in Figure 2.

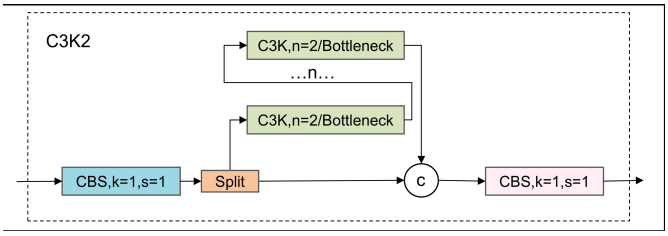


Fig. 2. C3K2 Module Architecture

In the Neck section, YOLOv11n employs the C2PSA module to enhance spatial modeling capabilities during multi-scale feature fusion. This module combines the CSP ar-

chitecture with the PSA mechanism, preserving efficient gradient flow pathways while strengthening the model's perception of target spatial locations. C2PSA first splits the input features into two branches: one follows the conventional convolutional path to preserve the original spatial structure, while the other introduces a parallel attention mechanism to model contextual regions at different scales. PSA captures semantic information within distinct receptive fields by constructing multiple sub-branches in parallel. These are then compressed and aggregated to guide the model's focus toward key areas, thereby enhancing robustness for objects with varying scales and complex backgrounds. The fused features from both branches ultimately generate representations with higher discriminative power. This module enhances spatial dependency modeling while mitigating the lack of global perception in shallow features. The C2PSA architecture is illustrated in Figure 3.

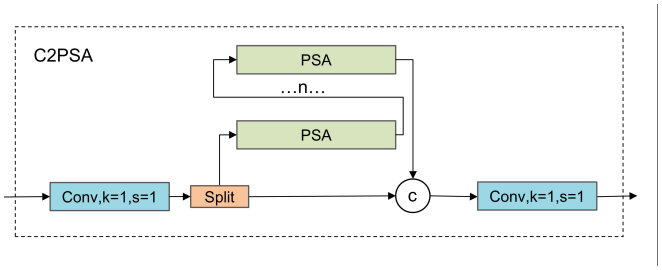


Fig. 3. C2PSA Module Architecture

During the feature fusion stage, YOLOv11n retains the dual-path architecture of FPN and PAN while adjusting network width and depth to meet lightweight deployment requirements. By introducing optimized activation functions (such as SiLU) and normalization operations in certain connection pathways, unnecessary computational redundancy is reduced. This strategy not only accelerates inference speed but also maintains a favorable balance between feature representation capability and model convergence performance, providing more stable high-level semantic support for subsequent modules.

With its compact structure, fast inference speed, and high detection accuracy, YOLOv11n demonstrates excellent adaptability in practical applications. Its optimizations in feature extraction capabilities, multi-scale modeling effects, and edge deployment efficiency make it particularly suitable for tasks sensitive to real-time performance and computational resources.

3.2. Improved Model YOLO-BDM

The overall architecture of the YOLO-BDM model is shown in Figure 4. Building upon YOLOv11n, this model introduces three key modules-DBB, MCA, and BiFPN-aimed at enhancing feature extraction and fusion capabilities to improve the accuracy and adaptability of vessel detection.

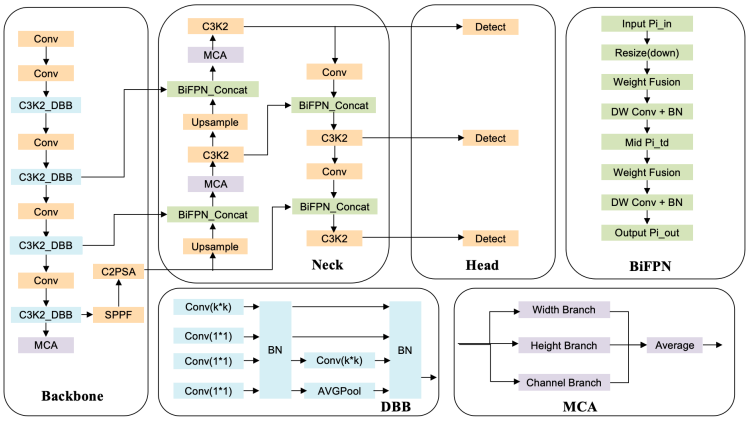


Fig. 4. YOLO-BDM network architecture

Within the YOLO-BDM model, systematic improvements to the backbone, neck, and feature fusion structures achieve unified optimization of multi-scale feature modeling and context-guided processing. In the backbone network, the Diversified Branch Block (DBB) replaces the original Bottleneck convolution, enabling multi-branch convolution fusion. This enhances the representation of objects across different scales and shapes, particularly boosting detection performance for small targets. Multiscale Contextual Attention Mechanism (MCA) is embedded at critical nodes in the backbone and neck layers. By dynamically adjusting attention distribution across channels, MCA effectively amplifies responses in target regions while suppressing background interference. The neck layer employs the BiFPN_Concat feature fusion module, which preserves multiscale information through concatenation operations, improving detection capabilities for objects with significant scale variations. Through the synergistic integration of DBB, MCA, and BiFPN, YOLO-BDM achieves significant improvements in ship detection accuracy and robustness while maintaining lightweight architecture. It is particularly well-suited for remote sensing scenarios involving small targets, ambiguous contours, and complex backgrounds.

3.3. Bidirectional Feature Pyramid Network (BiFPN)

Small vessels in remote sensing imagery typically exhibit characteristics such as compact dimensions, blurred edges, and indistinct textural features. These traits make it challenging for traditional object detection models to accurately locate and identify such targets within complex backgrounds, significantly compromising overall detection accuracy. To enhance the model's adaptability across multi-scale scenarios, this paper introduces the Bidirectional Feature Pyramid Network (BiFPN). This approach strengthens effective interactions between features at different levels and further optimizes the fusion strategy for multi-scale features.

Compared to conventional feature fusion methods like Feature Pyramid Network (FPN) and Path Aggregation Network (PANet), BiFPN introduces both top-down and bottom-up

information propagation pathways in its structural design, enabling efficient coupling of multi-level features. Furthermore, this architecture enhances cross-scale information aggregation through repeated stacking and incorporates a learnable weighted fusion mechanism. This allows the model to dynamically adjust the contribution ratios of different input features during the fusion process, thereby improving the selectivity and robustness of overall feature representation.

Given these advantages, this paper introduces the BiFPN module into the Neck section of YOLOv11n, replacing the original PANet structure to achieve more precise and stable multi-scale feature representation. The overall structure of this module is shown in Figure 5.

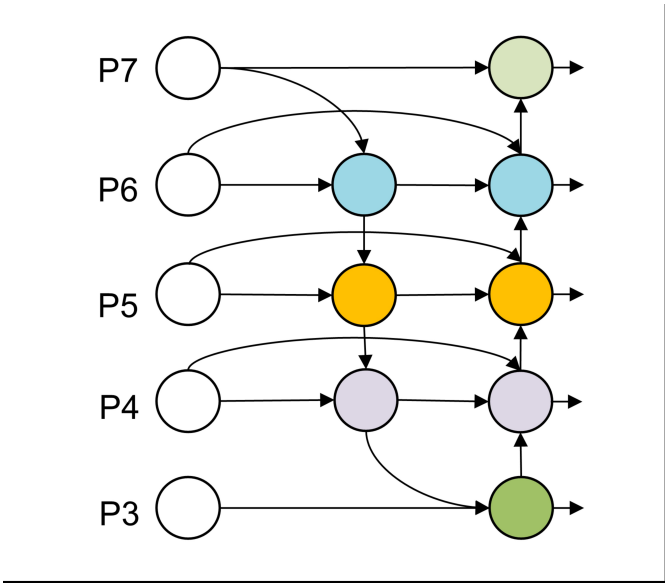


Fig. 5. BiFPN Architecture

Traditional feature pyramid structures typically employ uniform or static weighting when fusing multi-scale features, failing to dynamically adjust based on the varying contributions of different features in object detection tasks. BiFPN addresses this issue by introducing a learnable weighting fusion mechanism. This enables the network to adaptively allocate fusion weights for feature maps at different scales during training, thereby more effectively integrating shallow-layer detail information with deep-layer semantic representations. This enhances the network's modeling capability for objects at varying scales. During feature fusion, BiFPN assigns distinct trainable weights to multiple input feature maps. The fusion process can be represented as:

$$\boldsymbol{y} = \frac{w_1 \boldsymbol{x}_1 + w_2 \boldsymbol{x}_2}{w_1 + w_2 + \epsilon}. \tag{1}$$

In the formula, x_1 and x_2 represent input feature maps from different scales, w_1 and w_2 are their corresponding non-negative learnable weight parameters, and ϵ is a minimal constant introduced to prevent the denominator from becoming zero. This normalization mechanism not only ensures numerical stability but also enables the model to automatically learn the relative importance of different features during the fusion process. For scenarios with multiple input features, this formula can be naturally extended to:

$$\mathbf{y} = \frac{\sum_i w_i \cdot \mathbf{x}_i}{\sum_i w_i + \epsilon}. \tag{2}$$

This method avoids the shortcoming of simple weighted averaging, which treats all features equally, thereby endowing the network with stronger scale adaptability.

Structurally, FPN employs a top-down, unidirectional feature propagation approach, while PANet introduces bottom-up pathways to enhance the supplementation of deep-level semantics by shallow features. BiFPN further optimizes this framework by designing bidirectional pathways as independent fusion modules and processing feature maps through repeated multi-level stacking, thereby achieving efficient cross-level information exchange. Additionally, this architecture incorporates lateral cross-scale connections, enabling direct information exchange between feature maps at the same layer during fusion to enhance semantic consistency. Typically, BiFPN fuses feature maps from layers P3 to P7 to cover full-scale feature information spanning low-level details to high-level semantics.

To further enhance fusion efficiency and reduce redundant computations, BiFPN introduces a structural pruning strategy during the network construction phase. Specifically, for fusion nodes receiving input from only a single path (i.e., connected by only one upstream edge), the system determines that they do not constitute effective information exchange. Consequently, these nodes are directly skipped during fusion layer construction, thereby avoiding the introduction of ineffective fusion operations. This strategy can be formally described as follows:

$$\text{if } \text{fan}(n) = 1 \Rightarrow \text{prune}(n). \tag{3}$$

In the formula, fan(n) denotes the number of n inputs to node . When only a single input path exists, the system skips constructing this fusion node. This mechanism effectively simplifies the network architecture, reducing computational complexity and parameter size. It is particularly suitable for embedding BiFPN structures into lightweight object detection models, such as the YOLOv11n framework adopted in this paper.

Furthermore, the pruning operation does not affect feature flow along critical paths, as it only removes redundant "pseudo-fusion" nodes. Consequently, it significantly enhances computational efficiency while preserving detection performance.

As an enhanced feature fusion network, BiFPN plays a pivotal role in the YOLO-BDM model. Through strategies including weighted feature fusion, top-down and bottom-up bidirectional feature propagation, cross-scale connections, and pruning optimization, it enhances the model's detection capabilities for multi-scale objects and strengthens feature representation. This results in significantly improved detection accuracy in complex scenes.

3.4. Diverse Branch Block (DBB)

In object detection tasks, models need to simultaneously capture fine details of small objects and global semantics of large objects. Traditional convolutional layers, however, have limited feature representation due to their monolithic structure, making it difficult to effectively model multi-scale and diverse features in complex scenes. Multi-branch architectures, such as the Inception series, can enrich the feature space through branches of varying scales and complexities, but they incur significant inference overhead, limiting practical deployment. The Diverse Branch Block (DBB) addresses this issue by introducing diverse branches during training to enhance representation, which are equivalently merged into a single convolution at inference via structural re-parameterization, achieving "enhanced training representation with zero additional inference cost."

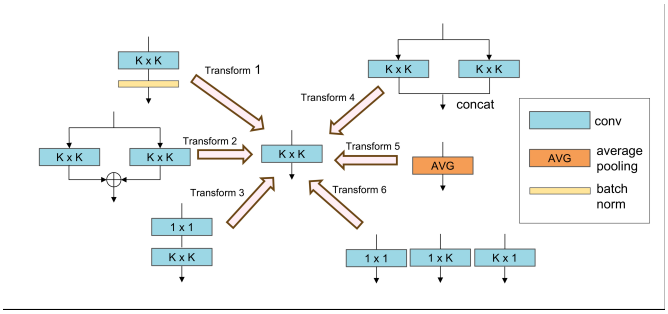


Fig. 6. DBB Architecture

The structure of the DBB module is illustrated in Figure 6. This module consists of multiple complementary branches, including a standard $k \times k$ convolution branch, a sequentially stacked 1×1–$k \times k$ convolution branch, a standalone convolution branch, and an average pooling (AVG Pooling) branch. Additionally, the module allows the incorporation of non-square convolutions (e.g.$1 \times k$, $k \times 1$,) to further expand the receptive field. Each branch is equipped with a batch normalization (BN) layer, introducing nonlinearity during training and significantly enhancing representational capability. Unlike Inception, DBB can fold all branches into a single convolution layer at inference through strict numerical equivalence transformations, avoiding the extra computational and memory overhead of multi-branch structures. This characteristic enables DBB to improve feature representation while maintaining efficient inference speed.

In mathematical modeling, the core of DBB lies in leveraging the linearity of convolutions and structural re-parameterization to equivalently transform the multi-branch structure during training into a single convolution at inference. Let the input feature be $I \in \mathbb{R}^{C\times H\times W}$, the convolution kernel $F \in \mathbb{R}^{D\times C\times K\times K}$, and the bias $b \in \mathbb{R}^{D}$. The basic convolution operation can be expressed as:

$$O = I * F + \mathrm{REP}(b) \tag{4}$$

where $\mathrm{REP}(b)$ denotes bias expansion. The key principle of DBB is based on the homogeneity and additivity of convolutions:

$$I * (pF) = p(I * F), \quad \forall p \in \mathbb{R}, \tag{5}$$

$$I * F^{(1)} + I * F^{(2)} = I * \left(F^{(1)} + F^{(2)}\right). \tag{6}$$

Leveraging this property, DBB designs six types of equivalent transformations (Transform I–VI) to progressively fold multi-branch operations into a single convolution. First, Transform I illustrates the fusion of convolution and batch normalization (BN). By absorbing the scaling and shifting parameters of BN, the convolution kernel and bias can be redefined as:

$$F'_j = \frac{\gamma_j}{\sigma_j} F_j, \quad b'_j = -\frac{\mu_j \gamma_j}{\sigma_j} + \beta_j, \tag{7}$$

thus eliminating the need for an additional BN layer during inference. Building on this, Transform II utilizes the additivity of convolutions: if multiple convolution branches share the same configuration, their weights and biases can be directly summed:

$$F' = F^{(1)} + F^{(2)}, \quad b' = b^{(1)} + b^{(2)}, \tag{8}$$

For more complex cases, Transform III addresses sequentially stacked 1×1 and $k \times k$ convolutions. Since the former only performs channel mixing, it can be merged with the latter into a single equivalent convolution:

$$F' = F^{(2)} * TRANS\left(F^{(1)}\right), \quad b' = \widehat{b} + b^{(2)}, \tag{9}$$

Transform IV corresponds to the channel concatenation commonly seen in Inception structures. Essentially, it concatenates the convolution kernels and biases of multiple branches along the output channel dimension, equivalent to a wider convolutional layer:

$$F' = \mathrm{CONCAT}\left(F^{(1)}, F^{(2)}\right), \quad b' = \mathrm{CONCAT}\left(b^{(1)}, b^{(2)}\right), \tag{10}$$

Additionally, DBB supports mapping non-convolution operations into convolutions. Transform V shows that average pooling can be regarded as a convolution with fixed weights:

$$F'_{d,c,:,:} = \begin{cases} \dfrac{1}{K^2}, & d = c, \\ 0, & d \neq c. \end{cases} \tag{11}$$

Finally, Transform VI demonstrates how non-square convolutions (e.g.$1 \times k$, $k \times 1$,) can be expanded via zero-padding into standard $k \times k$ convolutions, ensuring that all branches can be merged into a uniform form.

In summary, DBB achieves a strict mapping from "multi-branch during training" to "single convolution during inference" through these six transformations. This allows the model to exploit diverse paths for enhanced representation during training while maintaining the same computational cost as standard convolutions during inference, balancing performance and efficiency.

3.5. Multi-Scale Contextual Attention (MCA)

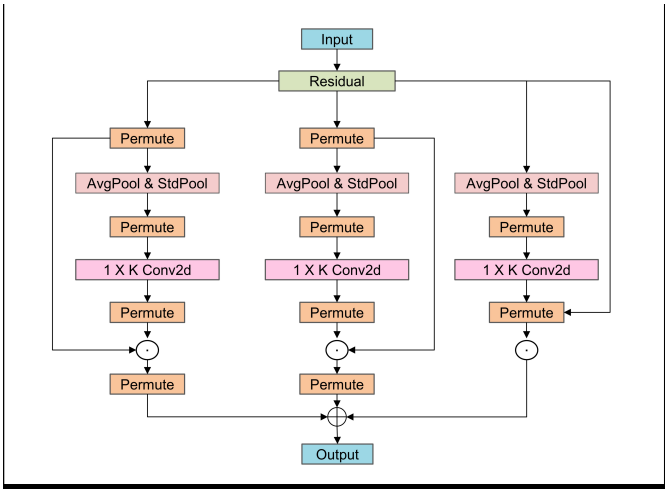


Fig. 7. MCA Architecture

The Multi-scale Contextual Attention (MCA) module is a lightweight architecture designed for visual perception tasks, widely applied in image segmentation, object detection, and similar applications. Its core objective is to enhance the model's ability to understand multi-scale semantic information and local details within images. By jointly modeling multi-scale contextual information and incorporating an attention mechanism for adaptive feature enhancement, the MCA module enables the network to focus more precisely on critical regions within images. This significantly improves the robustness and accuracy of object detection.

Traditional multi-scale feature fusion methods (such as FPN and PANet) integrate feature information across different levels to some extent. However, they often overlook deep semantic correlations between contextual information during feature fusion, leading to potential information redundancy or loss. Furthermore, these methods typically employ static or equal-weight fusion strategies, lacking the ability to dynamically model the contribution of feature maps at different scales. This makes it difficult to flexibly adjust based on target size or image complexity, thereby impacting detection performance.

To overcome the aforementioned issues, the MCA module introduces a multi-scale contextual modeling mechanism during the feature modeling stage, employing two statistical pooling operations: Global Average Pooling (AvgPool) and Global Standard Deviation Pooling (StdPool). The former models the overall background semantic information of the image, while the latter enhances sensitivity to edge textures and local structures.Given an input feature map $F \in \mathbb{R}^{C\times H\times W}$, the computation of the two contextual descriptor vectors is expressed as:

$$F_{avg} = \mathrm{AvgPool}(F), \quad F_{std} = \mathrm{StdPool}(F), \tag{12}$$

where $F_{avg}, F_{std} \in \mathbb{R}^C$ denote the channel-wise average and standard deviation features, respectively. This dual-branch structure provides complementary information at both global and local scales, establishing a context-aware basis for subsequent feature weighting.

After obtaining the contextual features, the MCA module introduces a channel attention mechanism to dynamically adjust the response strength of different features. Specifically, the two contextual vectors are passed through a shared-parameter fully connected subnetwork to generate channel weights, which are then normalized using the Sigmoid activation function:

$$\alpha_{avg} = \sigma(\mathrm{FC}(F_{avg})), \quad \alpha_{std} = \sigma(\mathrm{FC}(F_{std})), \tag{13}$$

Here, $\alpha_{avg}, \alpha_{std} \in \mathbb{R}^C$ are the attention weight vectors derived from the average and standard deviation branches, respectively, and $\sigma(\cdot)$ denotes the Sigmoid function, constraining the attention weights within the range [0, 1]. In this way, MCA effectively emphasizes feature channels that contribute to target regions while suppressing redundant information.

Finally, MCA fuses the attention-weighted signals from the two contextual branches with the original feature map to produce the context-enhanced output feature map:

$$F_{MCA} = (\alpha_{avg} + \alpha_{std}) \otimes F \tag{14}$$

where $\otimes$ denotes element-wise multiplication along the channel dimension. The fused feature map not only preserves the original semantic representation but also introduces prominent responses to critical regions along the channel dimension, enabling the network to focus more effectively on target areas.

In terms of fusion strategy, MCA does not treat all scale features equally but instead adapts feature map importance modeling based on contextual information: for small object detection tasks, the module prioritizes enhancing information from high-resolution feature layers to preserve more detailed textures; whereas in large object recognition scenarios, MCA emphasizes deeper, lower-resolution features with stronger semantic expressiveness. This mechanism significantly enhances the model's flexibility in feature modeling across different-scale objects, contributing to improved overall detection performance.

In summary, by introducing multi-scale context awareness and attention regulation mechanisms, the MCA module enhances the network's ability to synergistically understand both the spatial structure and semantic content of images. This compact and versatile module is particularly well-suited for challenging scenarios in remote sensing imagery, such as dense small objects and complex backgrounds. Integrating MCA into the high-level semantic layer of the backbone network effectively improves the model's object discrimination capability and localization accuracy in complex backgrounds.

4. Experimental Results and Analysis

4.1. Dataset

This experiment utilizes the SAR image vessel detection dataset constructed by the Chinese Academy of Sciences' Institute of Remote Sensing and Digital Earth, along with the

SeaShip dataset, as experimental data sources. SAR-ShipData primarily employs domestically produced GF-3 SAR data and Sentinel-1 SAR data as its main sources. It features diverse vessel slice types and varied backgrounds, making it suitable for multiple SAR image application scenarios. The SeaShip dataset, constructed from multi-source optical remote sensing imagery, encompasses diverse vessel types including container ships, oil tankers, passenger ferries, and fishing vessels. It features rich scale variations and complex background interferences, enabling research on model transferability and generalization performance between optical and SAR scenarios. Regarding data partitioning strategies, both datasets were randomly divided. The SAR-Ship dataset was split into training, validation, and test sets at an 8:1:1 ratio, while the SeaShip dataset was partitioned at a 7:2:1 ratio for model training, parameter tuning, and performance evaluation, as shown in Table 1.

Table 1. Distribution of dataset quantities

	SAR-Ship	SeaShip
Number of train set images	31783	4900
Number of test set images	3974	1400
Number of val set images	3972	700

4.2. Experimental Setup and Parameter Configuration

We conducted experiments on a system running Ubuntu 18.04.5, utilizing PyCharm as the software environment. Table 2 details the configuration information of the experimental platform used for training.

Table 2. Experimental platform configuration information

Item	Value
CPU	Intel(R) Xeon(R) Platinum 8362 @ 2.80GHz
GPU	NVIDIA GeForce RTX 3090
CUDA Version	11.1
Data processing	Python 3.8.10
Deep learning framework	PyTorch 1.8.1

The hyperparameters used during training are as follows: The input image size for all experiments was 640 $\times$ 640 pixels. All other parameters were set to the default values of the YOLOv11n model. To accelerate model convergence, mosaic data augmentation was disabled during the final 10 epochs of training. Detailed parameters of the trained model are shown in Table 3.

Table 3. Detailed parameters of the trained model

Parameters	Value
Epochs	200
Batch size	16
Input image size	640×640
Learning rate	0.01
Momentum	0.937
Weight decay	0.0005

4.3. Evaluation Metrics

To evaluate model performance, the following metrics are used: Mean Average Precision (mAP), target recognition accuracy per category, number of model parameters (Parameters), model floating-point operations (GFLOP), and model size. These metrics assess both model accuracy and efficiency. Here, TP denotes correctly predicted positive samples (correctly classified vessels), while FP denotes correctly predicted negative samples (incorrectly classified vessels). FN denotes correctly predicted false ship samples, while TN denotes incorrectly predicted false ship samples. AP represents the accuracy of target detection. Additionally, the average precision across all classes is denoted as mAP. mAP50 indicates the average of AP50 at an IoU threshold of 50%, where N represents the number of classes. The mAP@50-95 metric calculates the average AP values computed across IoU thresholds from 0.50 to 0.95 (in 0.05 increments). The number of model parameters, the model's floating-point operations, and its size reflect the computational complexity and resource requirements of the model. The formula is as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

$$mAP = \frac{1}{k} \sum_{i=1}^{k} AP_i \tag{18}$$

$$mAP@0.5 = \frac{1}{N} \sum_{t=1}^{N} AP_t, \quad IoU = 0.5 \tag{19}$$

4.4. Training Results and Analysis

To validate the advantages of YOLO-BDM in object detection, we compared the proposed model with YOLOv11n under identical experimental conditions on the SAR-Ship dataset. The results are shown in Table 4.

Table 4. Performance Comparison of YOLO-BDM and YOLOv11n on SAR-Ship

	YOLOv11n	YOLO-BDM
Params	2.60MB	2.90MB
FLOPs	6.4G	7.2G
Precision	93.47%	94.11%
Recall	91.81%	93.07%
F1	92%	94%
mAP@0.5	96.53%	97.27%

From Table 4, it can be seen that YOLO-BDM has a parameter size of 2.90 MB, slightly higher than YOLOv11n's 2.60 MB, indicating that our model incorporates additional parameters to enhance feature extraction capabilities. The computational cost of YOLO-BDM is slightly greater than that of YOLOv11n, but still remains within an acceptable range. In addition, Precision, Recall, and F1 increased by 0.64%, 1.26%, and 2%, respectively, demonstrating that the improved model can reduce false positives and false negatives, achieving a better balance between precision and recall and improving detection accuracy. The mAP@0.5 increased by 0.74%, indicating an overall enhancement in YOLO-BDM's detection performance. To systematically evaluate the vessel recognition capability of YOLO-BDM, we compared the two models using Precision, Recall, and mAP@0.5 metrics on the SAR-Ship dataset through bar charts, and plotted their P-R curves, as shown in Figures 8 and 9.

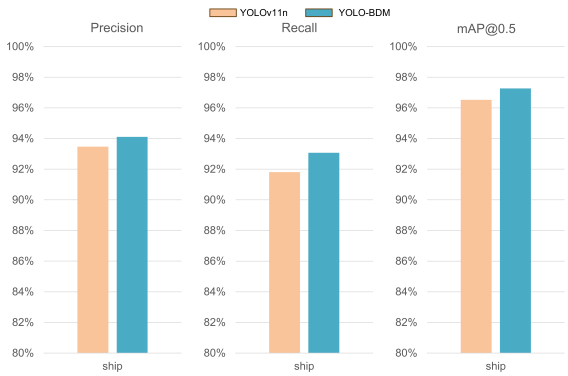
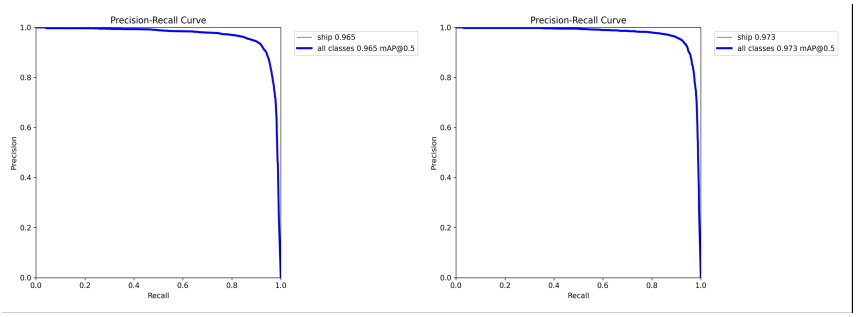


Fig. 8. Test Results for Different Types of Vessels

As shown in Figures 8, YOLO-BDM outperforms the baseline model YOLOv11n across all three metrics, indicating that our model effectively enhances vessel detection performance. The improvement in Precision demonstrates that YOLO-BDM excels at reducing false positives. The increase in Recall indicates that YOLO-BDM can detect

more targets and reduce missed detections, particularly under complex backgrounds or partial occlusions, reflecting stronger model robustness. The improvement in mAP@0.5 signifies that YOLO-BDM achieves superior overall detection performance, indicating a better balance across multiple evaluation metrics.

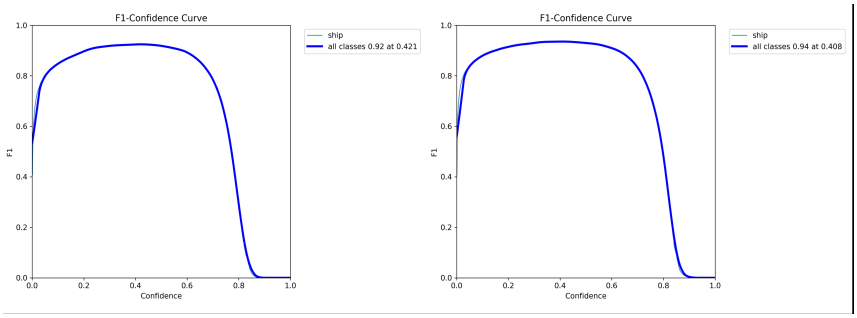


(a) P-R Curve of YOLOv11n Model (b) P-R Curve of YOLO-BDM Model

Fig. 9. Comparison of P-R Curves of YOLOv11n and YOLO-BDM on the SAR-Ship Dataset

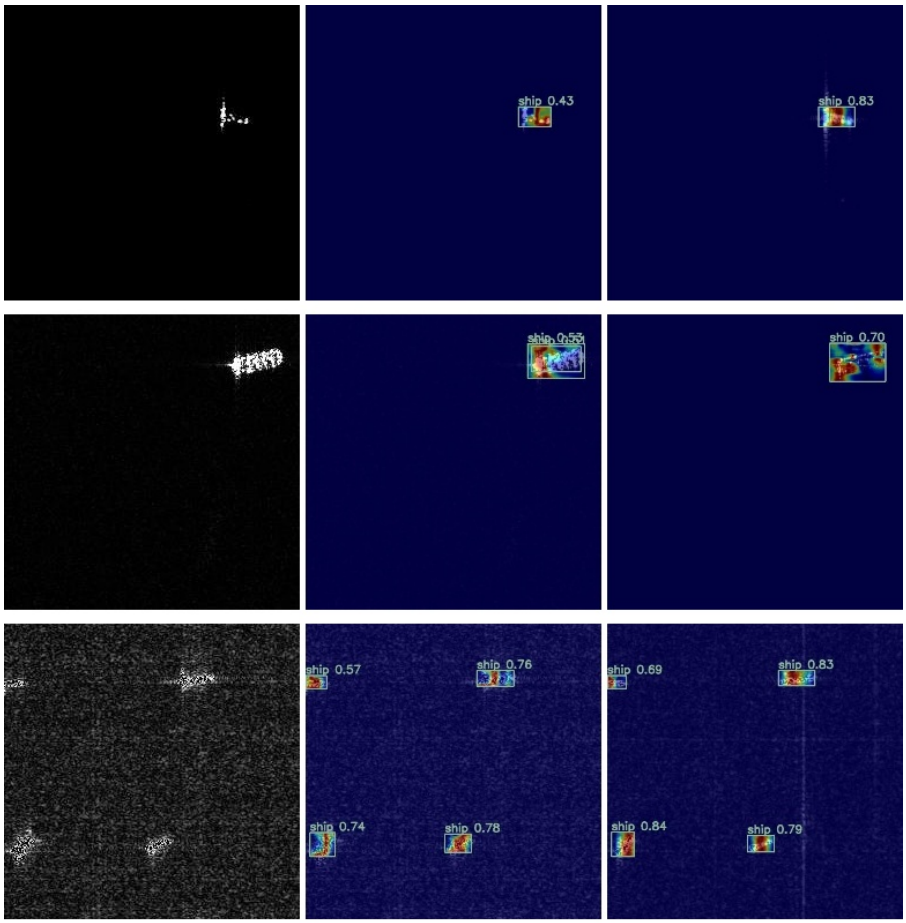
Figures 9 shows the precision-recall curves for YOLOv11n and the improved model YOLO-BDM, used to evaluate the performance of both models in the ship detection task. The closer the precision-recall curve is to the (1,1) corner, the stronger the model's detection capability. YOLOv11n's P-R curve exhibits a noticeable decline in the high Recall region, indicating that the baseline model introduces a significant number of false detections when identifying targets. In contrast, YOLO-BDM's P-R curve approaches the upper-right corner more closely, demonstrating that the YOLO-BDM model maintains high Recall while preserving high Precision, resulting in more stable performance. To compare the comprehensive performance metrics of the models, we plotted the corresponding F1 curves, as shown in Figures 10.

Analysis of the figure above reveals that YOLO-BDM achieves a 2% improvement in F1 score (0.92 $\rightarrow$ 0.94) compared to YOLOv11n. This indicates that the enhanced model strikes a better balance between Precision and Recall, resulting in superior overall detection performance. The YOLO-BDM model achieves its highest F1 score at a lower confidence threshold of 0.408, meaning it maintains good detection performance at lower confidence levels and is more sensitive to target vessels. Furthermore, to visually compare the difference in feature attention regions between the two models, this paper presents heatmap visualizations on test samples from the SAR-Ship dataset, as shown in Figures 11. The visualization reveals that YOLOv11n exhibits scattered or misfocused responses in certain complex backgrounds, whereas YOLO-BDM demonstrates more concentrated high-response zones within the ship target areas. This indicates that YOLO-BDM more effectively captures ship feature regions under complex background conditions, reducing interference from irrelevant backgrounds and thereby enhancing detection accuracy and robustness.



(a) F1 Curve of YOLOv11n Model (b) F1 Curve of YOLO-BDM Model

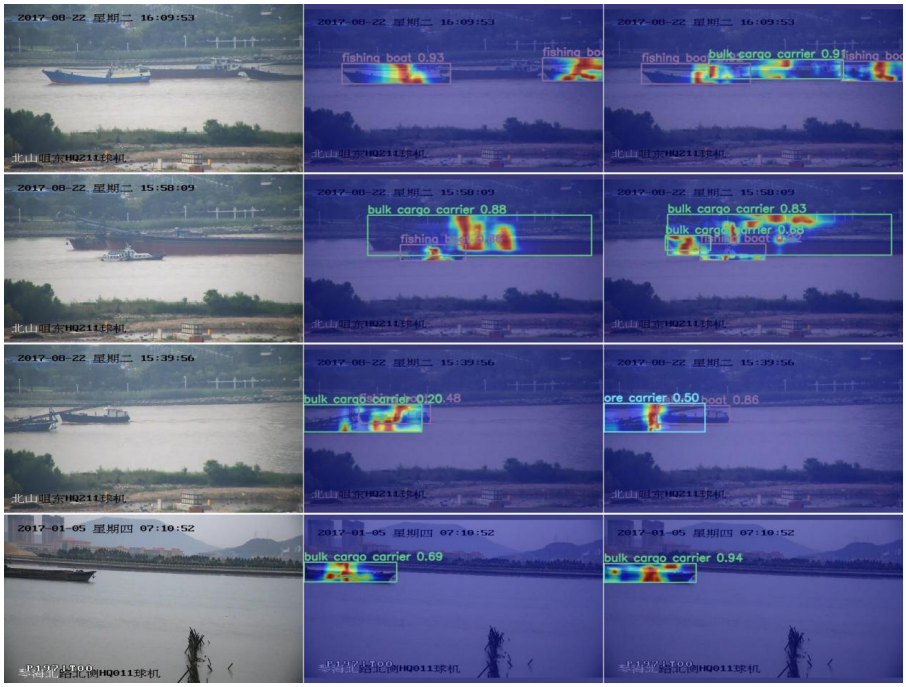
Fig. 10. Comparison of F1 Curves of YOLOv11n and YOLO-BDM on the SAR-Ship Dataset



(a) original image (b) YOLOv11n (c) YOLO-BDM

Fig. 11. Heatmaps of YOLOv11n and YOLO-BDM on the SAR-Ship Dataset

After completing training and performance analysis based on the SAR-Ship dataset, supplementary experiments were conducted using the optical remote sensing ship dataset SeaShip to further validate the generalization capability and robustness of the YOLO-BDM model in cross-modal ship detection tasks. The SeaShip dataset exhibits significant differences from SAR-Ship in imaging mechanisms, spectral characteristics, and background textures, while also encompassing a broader range of vessel types and more complex background interferences.



(a) original image (b) YOLOv11n (c) YOLO-BDM

Fig. 12. Heatmaps of YOLOv11n and YOLO-BDM on the SeaShip Dataset

Figures 12 presents a heatmap comparison between YOLOv11n and YOLO-BDM on the SeaShip dataset. Visual analysis reveals that compared to the baseline model YOLOv11n, the proposed YOLO-BDM model demonstrates improvements in several aspects. First, it exhibits enhanced localization accuracy for small object detection, capturing feature responses of minute targets with greater precision. Second, it demonstrates stronger interference resistance in complex backgrounds, effectively suppressing false detections. Third, it exhibits superior discrimination capabilities for detecting densely arranged targets.

4.5. Ablation Studies

To validate the effectiveness of each module within YOLO-BDM, we conducted a series of ablation experiments. The specific experimental results are shown in Table 5.

Table 5. Comparison of the effectiveness of different modules for vessel detection

Number	BiFPN	DBB	MCA	Precision (%)	Recall (%)	mAP@0.5 (%)	Params (M)	FLOPs (G)
1	×	×	×	93.47	91.81	96.53	2.59	6.4
2	✓	×	×	93.03	92.97	96.70	2.60	6.5
3	×	✓	×	93.74	92.95	97.08	2.62	6.6
4	×	×	✓	93.61	92.74	97.09	2.85	7.1
5	✓	✓	×	94.02	92.83	97.12	2.63	6.6
6	✓	×	✓	93.80	93.00	97.07	2.87	7.1
7	×	✓	✓	93.99	93.01	97.22	2.88	7.2
8	✓	✓	✓	94.11	93.07	97.27	2.90	7.2

In Table 5, the table presents the impact of different improvement modules (BiFPN, DBB, MCA) on vessel detection performance. Model 1 is the baseline model YOLOv11n, where "✓" indicates the module being incorporated.

Model 2 only introduces the DBB module; the Recall increases by 1.16%, but Precision slightly decreases, indicating that while more vessels are detected, the number of false positives also rises. This shows that DBB mainly improves feature extraction capability, helping detect more vessels.Model 3 only incorporates the MCA module, resulting in Precision rising to 93.74%, Recall increasing by 1.14%, and mAP@0.5 improving by 0.55%. These data demonstrate that MCA effectively enhances fine-grained feature representation and contextual modeling, improving vessel detection under complex backgrounds.Model 4 only introduces BiFPN, primarily strengthening multi-scale feature fusion and enhancing target discrimination, but it may cause some difficult-to-detect targets to be missed, so the Recall shows little change. This indicates that BiFPN mainly contributes to multi-scale feature aggregation, improving accuracy for detectable targets.

Models 5, 6, and 7 correspond to the pairwise combinations of the three modules. The combination of DBB and BiFPN increases both parameter size and computational cost, but the performance improvement is less than expected, indicating potential redundancy in the fused features. The combination of MCA and BiFPN not only enhances multi-scale feature fusion but also improves detection accuracy. The combination of DBB and MCA results in a significant increase in Precision, but Recall decreases, possibly due to stricter target selection caused by feature enhancement.

Model 8, which integrates all three modules, achieves the best results across all metrics, demonstrating that their combination can effectively enhance the accuracy and robustness of vessel detection. Among them, the MCA mechanism is the key factor in improving detection performance, particularly evident in the increases in Recall and overall mAP. The BiFPN module primarily strengthens multi-scale feature fusion, while the DBB module mainly enhances the model's feature extraction capability for vessels.

4.6. Comparative Experiments

To comprehensively evaluate the performance of YOLO-BDM, we conducted comparative experiments with several lightweight YOLO models (YOLOv4-tiny [21], YOLOv5s, YOLOv7-tiny [34], YOLOv8n [39], YOLOv10n [33], and YOLOv11n [22]). Analysis was conducted across four metrics: Precision, Recall, mAP@0.5, and Parameters. The results are as follows:

Table 6. Performance and Parameters of Different Models on SAR-SHIP Dataset

Model	**Precision (%)**	**Recall (%)**	**mAP@0.5 (%)**	**Params** ($\times 10^6$)
YOLOv4-tiny	87.41	90.28	86.01	5.92
YOLOv5s	91.12	88.15	88.77	7.19
YOLOv7-tiny	91.61	91.97	92.44	6.30
YOLOv8n	93.74	93.75	96.23	3.01
YOLOv10n	93.12	93.42	96.46	2.71
YOLOv11n	93.47	91.81	96.53	2.59
YOLO-BDM	94.11	93.07	97.27	2.90

As shown in Table 6 on the SAR-SHIP dataset, YOLO-BDM demonstrates superior performance across all metrics, particularly achieving the highest mAP@0.5 among all models. Compared with YOLOv4-tiny, YOLO-BDM improves precision by 6.7%, recall by 2.79%, and mAP@0.5 by 11.26%, indicating significant advantages in detection accuracy and stability. Compared with YOLOv8n and YOLOv10n, YOLO-BDM achieves improvements of 1.04% and 0.81% in mAP@0.5, respectively, demonstrating stronger overall detection capability.

YOLO-BDM has a parameter size of 2.90M, substantially smaller than YOLOv4-tiny, YOLOv5s, and YOLOv7-tiny, indicating a more lightweight model suitable for resource-constrained scenarios. Although its parameter count is slightly higher than YOLOv8n, YOLOv10n, and YOLOv11n, the gains in mAP@0.5 and recall justify the additional computational cost.

Table 7. Performance and Parameters of Different Models on SAR-SHIP Dataset

Model	**Precision (%)**	**Recall (%)**	**mAP@0.5 (%)**	**Params** ($\times 10^6$)
YOLOv4-tiny	94.36	91.61	95.87	5.92
YOLOv5s	97.69	97.73	97.53	7.19
YOLOv7-tiny	98.12	97.46	98.23	6.30
YOLOv8n	98.21	97.29	98.76	3.01
YOLOv10n	98.37	95.59	98.59	2.71
YOLOv11n	98.70	97.94	98.81	2.59
YOLO-BDM	99.39	98.06	98.97	2.90

To further validate the generalization capability of the model across different scenarios, this study compared the detection performance of several mainstream models on the SeaShip dataset (see Table 7). The results indicate that YOLO-BDM demonstrates excellent performance across all metrics. Its Precision reaches 99.39%, representing an increase of 5.03% over YOLOv4-tiny and improvements of 1.18% and 0.69% compared to YOLOv8n and YOLOv11n, respectively. In terms of mAP@0.5, YOLO-BDM achieves 98.97%, which is 3.10% higher than YOLOv4-tiny and shows gains of 0.21% and 0.38% over YOLOv8n and YOLOv10n, respectively. Furthermore, YOLO-BDM attains a Recall of 98.06%, exceeding YOLOv10n by 2.47%, reflecting stronger object detection capability. These results indicate that YOLO-BDM not only maintains leading overall accuracy but also exhibits clear advantages in detection stability and recall.

Overall, compared to YOLOv11n, YOLO-BDM achieves superior detection accuracy with only a slight increase in parameter count, demonstrating an effective balance between lightweight design and high-performance detection. In SAR ship detection tasks, it offers higher Precision, Recall, and mAP@0.5 while maintaining relatively low computational overhead.

5. Conclusion

This paper proposes YOLO-BDM, an enhanced SAR vessel detection model based on the YOLOv11 framework. To address typical challenges such as small target detection, blurred target contours, and complex ocean background interference, three key improvement modules are designed: DBB (Diversified Branch Block) enriches feature extraction and multi-scale modeling, MCA (Multi-scale Contextual Attention) enhances context-guided target attention capabilities, and BiFPN_Concat optimizes multi-scale feature fusion. Experimental results demonstrate that YOLO-BDM significantly outperforms existing lightweight YOLO models in core metrics including Precision, Recall, and mAP@0.5, while maintaining a low parameter count. This achieves a favorable balance between detection accuracy and computational efficiency, exhibiting strong robustness and adaptability particularly in detecting small, blurry-contoured vessels within complex backgrounds.

Although YOLO-BDM demonstrates strong performance in SAR vessel detection tasks, several areas warrant further investigation. Future work can be pursued in the fol-

lowing areas: Further optimizing the model architecture to reduce computational complexity, thereby adapting to resource-constrained devices such as drones and satellites for real-time detection requirements. Expanding the dataset scale by incorporating samples under more complex sea conditions (e.g., wave and wind effects, occlusions, camouflaged targets) to enhance the model's generalization capabilities.

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
2. Cao, R., Sui, J.: A dynamic multi-scale feature fusion network for enhanced sar ship detection. Sensors 25(16), 5194 (2025)
3. Chen, C., Han, D., Chang, C.C.: Caan: Context-aware attention network for visual question answering. Pattern Recognition 132, 108980 (2022)
4. Chen, C., Han, D., Chang, C.C.: Mpcct: Multimodal vision-language learning paradigm with context-based compact transformer. Pattern recognition 147, 110084 (2024)
5. Chen, C., Han, D., Guo, Z., Chang, C.C.: Towards bias-aware visual question answering: Rectifying and mitigating comprehension biases. Expert Systems with Applications 264, 125817 (2025)
6. Chen, C., Han, D., Shen, X.: Clvin: Complete language-vision interaction network for visual question answering. Knowledge-Based Systems 275, 110706 (2023)
7. Chen, Z., Ding, Z., Zhang, X., Wang, X., Zhou, Y.: Inshore ship detection based on multimodality saliency for synthetic aperture radar images. Remote Sensing 15(15), 3868 (2023)
8. Cui, J., Jia, H., Wang, H., Xu, F.: A fast threshold neural network for ship detection in large-scene sar images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15, 6016–6032 (2022)
9. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
10. Duan, R., Deng, H., Tian, M., Deng, Y., Lin, J.: Soda: A large-scale open site object detection dataset for deep learning in construction. Automation in Construction 142, 104499 (2022)
11. Feng, Y., Zhang, Y., Zhang, X., Wang, Y., Mei, S.: Large convolution kernel network with edge self-attention for oriented sar ship deteciton. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024)
12. Fu, X., Zhou, Z., Meng, H., Li, S.: A synthetic aperture radar small ship detector based on transformers and multi-dimensional parallel feature extraction. Engineering Applications of Artificial Intelligence 137, 109049 (2024)
13. Guo, H., Bai, H., Yuan, Y., Qin, W.: Fully deformable convolutional network for ship detection in remote sensing imagery. Remote Sensing 14(8), 1850 (2022)
14. Han, D., Shi, J., Zhao, J., Wu, H., Zhou, Y., Li, L.H., Khan, M.K., Li, K.C.: Lrcn: Layer-residual co-attention networks for visual question answering. Expert Systems with Applications 263, 125658 (2025)
15. Han, D., Zhu, Y., Li, D., Liang, W., Souri, A., Li, K.C.: 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)
16. He, J., Su, N., Xu, C., Liao, Y., Yan, Y., Zhao, C., Hou, W., Feng, S.: A cross-modality feature transfer method for target detection in sar images. IEEE Transactions on Geoscience and Remote Sensing 61, 1–15 (2023)

17. Hu, B., Liu, Y., Chu, P., Tong, M., Kong, Q.: Small object detection via pixel level balancing with applications to blood cell detection. Frontiers in Physiology 13, 911297 (2022)
18. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
19. Hu, Q., Hu, S., Liu, S.: Banet: A balance attention network for anchor-free ship detection in sar images. IEEE Transactions on Geoscience and Remote Sensing 60, 1–12 (2022)
20. Ji, P., Xing, S., Dai, D., Pang, B.: Deceptive targets generation simulation against multichannel sar. Electronics 9(4), 597 (2020)
21. Jiang, Z., Zhao, L., Li, S., Jia, Y.: Real-time object detection method based on improved yolov4-tiny. arXiv preprint arXiv:2011.04244 (2020)
22. Khanam, R., Hussain, M.: Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725 (2024)
23. Li, Z., Chen, J., Xiong, Y., Yu, H., Zhang, H., Gao, B.: A ship detection and imagery scheme for airborne single-channel sar in coastal regions. Remote Sensing 14(18), 4670 (2022)
24. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
25. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
26. Ma, X., Hou, S., Wang, Y., Wang, J., Wang, H.: Multiscale and dense ship detection in sar images based on key-point estimation and attention mechanism. IEEE Transactions on Geoscience and Remote Sensing 60, 1–11 (2022)
27. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra r-cnn: Towards balanced learning for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 821–830 (2019)
28. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
29. Shen, X., Chen, C., Han, D., Xu, Y., Wang, X., Zhou, H.: A triple-branch hybrid dynamic-static alignment strategy for vision-language tasks. Neural Networks p. 107871 (2025)
30. Shen, X., Han, D., Chang, C.C., Oad, A., Wu, H.: Gfsnet: Gaussian fourier with sparse attention network for visual question answering. Artificial Intelligence Review 58(6), 1–30 (2025)
31. Shen, X., Han, D., Chang, C.C., Xu, Y., Chen, C.: Multimodal context-aware consistency alignment for vision-language tasks. Expert Systems with Applications 295, 128857 (2026)
32. Tang, X., Zhang, J., Xia, Y., Cao, K., Zhang, C.: Pegnet: An enhanced ship detection model for dense scenes and multi-scale targets. IEEE Geoscience and Remote Sensing Letters (2025)
33. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al.: Yolov10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems 37, 107984–108011 (2024)
34. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7464–7475 (2023)
35. Wang, X., Wang, A., Yi, J., Song, Y., Chehri, A.: Small object detection based on deep learning for remote sensing: A comprehensive review. Remote Sensing 15(13), 3265 (2023)
36. Wang, X., Li, G., Zhang, X.P., He, Y.: A fast cfar algorithm based on density-censoring operation for ship detection in sar images. IEEE Signal Processing Letters 28, 1085–1089 (2021)
37. Xu, H.Y., Xu, F., Jin, Y.Q.: Optimal sensing principle of synthetic aperture radar. IEEE Transactions on Geoscience and Remote Sensing 62, 1–14 (2023)
38. Xu, X., Zhao, J., Li, Y., Gao, H., Wang, X.: Banet: A balanced atrous net improved from ssd for autonomous driving in smart transportation. IEEE Sensors Journal 21(22), 25018–25026 (2020)

39. Yaseen, M.: What is yolov8: An in-depth exploration of the internal features of the next-generation object detector (2024)
40. Zeng, T., Zhang, T., Shao, Z., Xu, X., Zhang, W., Shi, J., Wei, S., Zhang, X.: Cfar-dp-fw: A cfar-guided dual-polarization fusion framework for large-scene sar ship detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 17, 7242–7259 (2024)
41. Zhang, M., Ouyang, Y., Yang, M., Guo, J., Li, Y.: Orpsd: Outer rectangular projection-based representation for oriented ship detection in sar images. Remote Sensing 17(9), 1511 (2025)
42. Zhang, X., Yan, M., Zhu, D., Guan, Y.: Marine ship detection and classification based on yolov5 model. In: Journal of Physics: Conference Series. vol. 2181, p. 012025. IOP Publishing (2022)
43. Zhao, W., Syafrudin, M., Fitriyani, N.L.: Cras-yolo: A novel multi-category vessel detection and classification model based on yolov5s algorithm. IEEE Access 11, 11463–11478 (2023)
44. Zhou, S., Zhang, M., Wu, L., Yu, D., Li, J., Fan, F., Liu, Y., Zhang, L.: Sar ship detection network based on global context and multi-scale feature enhancement. Signal, Image and Video Processing 18(3), 2951–2964 (2024)
45. Zhou, Z., Cui, Z., Zang, Z., Meng, X., Cao, Z., Yang, J.: Ultrahi-prnet: An ultra-high precision deep learning network for dense multi-scale target detection in sar images. Remote Sensing 14(21), 5596 (2022)
46. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. Proceedings of the IEEE 111(3), 257–276 (2023)

Fangyuan Xiong is currently pursuing her postgraduate degree at the College of Computer Science and Information Engineering, Shanghai Maritime University. Her research focuses on computer vision and intelligent perception, with particular emphasis on lightweight object detection algorithms, ship detection and tracking, and the application of deep learning in marine monitoring. Her recent work primarily centers on improving the YOLO series of models.

Dezhi Han (Senior Member, IEEE) received the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China,in 2005.He is currently a Professor of Computer Science and Engineering with Shanghai Mar itime University, Shanghai, China. His research interests include cloud computing, mobile net working, wireless communication, and cloud security.

Xiang Shen is currently pursuing his Ph.D. at Shanghai Maritime University and is also a joint Ph.D. student at The University of Sydney, supported by the China Scholarship Council (CSC). His research focuses on multimodal learning, with particular emphasis on visual question answering, multimodal fusion strategies, and the development of large-scale multimodal models. His current work also explores the integration of AI agents for adaptive perception and reasoning in complex environments, aiming to bridge the gap between visual understanding and natural language processing.

Manlin Zhu is currently pursuing his master's degree at Shanghai Maritime University. His research focuses on computer vision and ship detection, with particular emphasis on image processing, object detection, and perception techniques in marine environments. He is committed to applying deep learning to marine monitoring scenarios.

Received: October 18, 2025; Accepted: January 5, 2026.

Computer Science and Information Systems 23(2):757–774 https://doi.org/10.2298/CSIS250913020B

Enhanced ROCKET for the Automated Detection of Epileptic Tonic-Clonic Seizures Using Accelerometer Data*

Krisztian Buza[1,2], Alexandros Nanopoulos[3], and Noémi Ágnes Varga[4]

[1] Budapest University of Economics and Business
Budapest, Hungary
buza.krisztian@uni-bge.hu (corresponding author)
[2] BioIntelligence Group, Department of Mathematics-Informatics
Faculty of Technical and Human Sciences, Sapientia Hungarian University of Transylvania
Targu Mures, Romania
buza@biointelligence.hu
[3] Business Informatics, Baden-Wuerttemberg Cooperative State University
Mosbach, Germany
alexandros.nanopoulos@gmail.com
[4] Institute of Genomic Medicine and Rare Disorders, Semmelweis University
Budapest, Hungary
noemiagnesvarga@gmail.com

Abstract. The detection of epileptic tonic-clonic seizures during everyday life based on accelerometric data from wearable devices would enhance the diagnostic and the follow-up of the epileptic patients. We develop an algorithm which may contribute to recognition of tonic-clonic epileptic seizure based on accelerometer data that can be collected from mobile and wearable devices. We consider this task to be a multivariate time-series classification problem. State-of-the art solutions to this problem are based on machine learning techniques, such as Random Convolutional Kernel Transform (ROCKET). We enhance ROCKET by replacing standard convolution with dynamic convolution. Dynamic convolution was originally defined for univariate time series, therefore, we extend it to multivariate time series. We perform experiments on two publicly available real-world datasets related to tonic-clonic seizures. The experimental results show that the proposed enhancements of the ROCKET algorithm significantly reduce the average classification error. Moreover, our approach outperforms other time series classifiers, including several types of deep neural networks that are commonly used in the domain of time-series classification. An enhanced version of the ROCKET algorithm is proposed for the automated detection of epileptic tonic-clonic seizures using accelerometer data. To assist reproducibility and follow-up works, we made our implementation publicly available at https://github.com/kr7/seizure .

Keywords: Epilepsy, Multivariate Time Series Classification, Human Activity Recognition, Random Convolutional Kernel Transform, ROCKET, Dynamic Convolution

* This is an extended version of the conference paper [9].

1. Introduction

WHO reports that around 50 million people are diagnosed with epilepsy based on recurrent seizures worldwide.[5] A single unprovoked seizure is a common phenomenon in the general population, with at least 10% of the population experiencing a seizure during their lifetime [7]. The rate of seizure recurrence can vary widely, but it is estimated that up to 70% of people living with epilepsy could live seizure-free if properly diagnosed and treated.

The detection and monitoring of epileptic seizures require specialized devices such as electroencephalography (EEG) and well-trained experts to interpret the results. However, in the majority of cases, the first seizures happen outside a healthcare unit. Since the patients are unconscious or partially conscious during the seizure, the description of the seizure originates from family members or stranger eyewitnesses. Patient-reported outcomes (PROs) in the form of seizure diaries are commonly used methods for monitoring the disease activity, but diaries are limited due to variable adherence and periictal amnesia [21].

With the advent of clinically relevant data that can be collected by mobile and wearable devices, such as accelerometer data, the wish for reliable control during continuous monitoring, appropriate identification of seizures, and timely alerting became more pertinent and feasible [42].

Due to its evident importance, in this manuscript, we focus on the research topic of recognizing tonic-clonic epileptic seizures[6] based on accelerometer data. Our goal is a self-detection mechanism that is both: i) more accessible, due to the use of technology that is commonly available in mobile and wearable devices; ii) more acceptable, due to more accurate recognition with a limited amount of false alarms, which can be a considerable burden for both patients and medical staff [39], as it has been identified in other kinds of self-detection mechanisms [18], too. Both these factors can increase the detection coverage in broader sets of the population. Moreover, such a self-detection mechanism can become complementary to more precise EEG-based detection, effectively acting as a feasible form of pre-screening.

Within this context, we consider the automated recognition of epileptic seizure as a time-series classification task, for which state-of-the-art solutions are based on machine learning. The underlying time series are multivariate as the acceleration is usually measured along several axes. In particular, in our case, the acceleration is measured along three axes, therefore, we work with 3-dimensional time series. Our solution is based on "Random Convolutional Kernel Transform" or ROCKET for short [15]. ROCKET is a recent time-series classification algorithm that was shown to outperform other time-series classifiers. We extend ROCKET by replacing conventional convolution with *dynamic convolution* [10]. Dynamic convolution replace standard dot-product calculations of ROCKET by dynamic time warping (DTW) calculations. The resulting classifier is thereby more robust w.r.t. local shifts and elongations that are present in time-series data. Dynamic convolution was originally defined for univariate time series, nevertheless accelerometer data corresponds to multivariate time series. Thus, we propose a new variant of dynamic convolution that works with multivariate time series.

[5] https://www.who.int/news-room/fact-sheets/detail/epilepsy

[6] Tonic-clonic epileptic seizures involve both tonic (stiffening) and clonic (twitching) phases of muscle activity.

In our experimental evaluation, we used two publicly available real-world datasets. We performed experiments according to the 10×10-fold cross-validation protocol. The experimental results show that the proposed enhancements of the ROCKET algorithm significantly reduce the average error rate (misclassification ratio) of the detection of epileptic tonic-clonic seizures.

We compared our approach to other time-series classifiers too, including several prominent deep neural networks and we observed that our approach outperforms them as well. More importantly, the amount of false alarms is also reduced drastically, which, as mentioned above, is a critical requirement for the acceptance of the envisioned self-detection mechanism.

The remainder of the paper is organized as follows: Section 2 provides on overview of related works. Section 3 introduces the background that is necessary to understand our work. Section 4 presents our approach in detail which is followed by the experimental evaluation (Section 5) and conclusions (Section 6).

2. Related Work

Machine learning techniques have been widely applied to detection and classification tasks in the biomedical domain, including emotion recognition [44], the assessment of schizophrenia [30], and the detection as well as early prediction of epileptic seizures [1, 3, 12, 22, 26, 33, 50]. A comprehensive survey of epileptic seizure detection methods is provided in [19]. Although electroencephalography (EEG) has demonstrated strong performance in controlled laboratory environments, its applicability in everyday settings remains limited. In particular, Baumgartner and Koren [3] note that, in outpatient contexts, scalp-EEG–based seizure detection is constrained by poor patient acceptance, as individuals are unlikely to tolerate wearing electrode arrays during daily activities.

The increasing availability of low-cost sensors and the widespread adoption of smart devices, such as smartphones and smartwatches, have stimulated growing interest in the automated detection of tonic-clonic seizures using accelerometer data. Several studies have explored this direction [8, 39, 6, 35, 34]. Bruno et al. [8] evaluated both medically certified and non-certified devices, with particular emphasis on the requirements of patients and caregivers. Regalia et al. [39] introduced what they described as the first commercially available multimodal wrist-worn devices designed to capture the physiological signatures of ongoing generalized tonic-clonic seizures. Beniczky et al. [6] investigated the clinical reliability of a wireless, wrist-worn accelerometer for detecting generalized tonic-clonic seizures, while Onorati et al. [35] examined classifiers based on accelerometer signals. More recently, Lupión et al. [34] analyzed seizure detection using low-cost IoT devices in combination with federated machine learning techniques.

The majority of these studies emphasize practical considerations related to seizure detection in daily life, such as device usability, patient comfort, and user experience. While these aspects are undoubtedly important, the present work focuses on the underlying machine learning methodologies used for classification. In prior research, classification models were often treated as black boxes, with existing algorithms applied without detailed justification, or classifiers constructed using manually engineered features [35], frequently without in-depth discussion of the employed learning techniques. In contrast,

we investigate advanced classification approaches, including deep learning models and state-of-the-art methods such as ROCKET.

As we formulate the automated recognition of epileptic tonic-clonic seizures as a time series classification problem, we briefly review relevant work in this area. Time series classification constitutes a unifying framework [23] for a wide range of detection and recognition tasks and has given rise to numerous methodological approaches. These include neural network–based models [20, 25, 48], Bayesian networks [36], hidden Markov models [17], decision trees [27], and methods based on frequent pattern discovery, commonly referred to as motifs or shapelets [47, 24, 32]. An additional line of research includes hubness-aware classifiers [38].

An influential study by Xi et al. [46] demonstrated that a simple k-nearest neighbor classifier combined with dynamic time warping (DTW) distance was highly competitive with, and in some cases superior to, many alternative classifiers proposed prior to that work. The effectiveness of DTW arises from its elastic alignment mechanism, which accommodates temporal shifts and local variations in length between time series.

Subsequent advances in time series classification increasingly leveraged deep learning techniques [48, 14, 49]. Among these, fully convolutional networks (FCNs) have been identified as a particularly strong baseline [45]. Wang et al. [45] reported that FCNs outperformed all other evaluated time series classifiers in terms of overall accuracy, with residual networks (ResNets) achieving comparable performance.

Despite their widespread adoption, convolutional neural networks (CNNs) exhibit inherent limitations in time series analysis. As discussed in [10], CNNs are primarily suited to rigid pattern matching. Even when convolutional layers are combined with pooling operations, the resulting architectures mainly handle translational invariance and fail to adequately accommodate temporal elongations of local patterns. Furthermore, their ability to address translations remains limited and inconsistent. These shortcomings motivated the development of dynamic convolution [10], which seeks to overcome such constraints.

More recently, the Random Convolutional Kernel Transform (ROCKET) was proposed as a highly effective approach to time series classification [15]. ROCKET and its deterministic variant, MiniROCKET [16], have demonstrated performance that surpasses many deep learning–based models, including FCNs. Nevertheless, both methods rely on standard convolution operations, and therefore inherit their associated limitations. In this work, we address these constraints and propose an approach designed to overcome them.

Preliminary investigations of the integration of dynamic convolution with ROCKET were presented in [11]. However, that work focused exclusively on univariate time series classification. In contrast, the present study addresses the detection of epileptic tonic-clonic seizures using multivariate accelerometer data. For completeness, we note that an earlier version of this work was reported in [9], outlining the core idea. The current paper substantially extends this preliminary study by broadening the experimental evaluation to additional datasets, comparing the proposed approach with a wider range of state-of-the-art methods, and providing a detailed analysis of false alarm rates.

3. Background

Next, we present a brief overview of the key concepts and techniques essential for understanding our work. Specifically, we begin with a formal definition of the time series

classification task, followed by a review of related works (Section 2), dynamic time warping (Section 3.2), dynamic convolution (Section 3.3), and ROCKET (Section 3.4). Finally, we present our approach in Section 4 and the datasets used in the experimental evaluation (Section 5.1).

3.1. Problem Formulation (Multivariate Time Series Classification)

Given a set $\mathcal{C}$ of class labels, and a set $\mathcal{D}$ (called training set) of time series together with their class labels

$$\mathcal{D} = \left\{ \left(x^{(i)}, y^{(i)} \right)_{i=1}^{n} \right\}, \;\; y^{(i)} \in \mathcal{C} \tag{1}$$

where $x^{(i)}$ is a multivariate time series. As in our case, the acceleration is measured along three axis, therefore

$$x^{(i)} = \Big((x_{1,1}^{(i)}, x_{1,2}^{(i)}, x_{1,3}^{(i)}), \ldots, (x_{l,1}^{(i)}, x_{l,2}^{(i)}, x_{l,3}^{(i)}) \Big),$$

where each $x_{j,k}^{(i)}$ is a real number, $j \in \{1, 2, ..., l\}$ and $k \in \{1, 2, 3\}$. We aim at finding a model $\mathcal{M}$ that is able to determine the class label $y' \in \mathcal{C}$ of any new (test) time series x'.

In this study, in line with the above problem formulation, we assume that tonic-clonic seizures are detected based on an acceleration time series of length l. In the case of a real-world application, this could be achieved by considering both the current observation and the previous $l - 1$ observations as the input of $\mathcal{M}$.

3.2. Dynamic Time Warping

Dynamic Time Warping (DTW) is an elastic distance measure for time series, built on the principles of dynamic programming [41]. It enables shifts and elongations when aligning two time series, allowing for a more flexible comparison.

The DTW distance of two time series, $x = (x_1, \ldots, x_l)$ and $x' = (x'_1, \ldots, x'_{l'})$, is computed by populating an $l \times l'$ matrix D. Each entry $d_{i,j} \in D$ represents the distance between two prefixes—one from x and the other from x'. Specifically, $d_{i,j}$ is the distance between the prefixes $(x_1, \ldots, x_i)$ and $(x'_1, \ldots, x'_j)$. It is computed as follows:

$$d_{i,j} = |x_i - x'_j| + \min \{ d_{i,j-1}, d_{i-1,j}, d_{i-1,j-1} \} \tag{2}$$

where the terms of the minimum correspond to the cases of elongation in x, elongation in x' or matching the next elements in both time series.[7]

The matrix entries $d_{i,j}$ are computed in a column-wise manner, following the sequence: $d_{1,1}$, $d_{2,1}$, …, $d_{l,1}$, $d_{1,2}$, $d_{2,2}$, …, $d_{l,2}$, and so forth until $d_{l,l'}$. The initialization begins with the first entry, defined as $d_{1,1} = |x_1 - x'_1|$. When certain terms, such as $d_{i,j-1}$, $d_{i-1,j}$, or $d_{i-1,j-1}$, are undefined (i.e., if $i - 1 = 0$ or $j - 1 = 0$), they are ignored. In such cases, the minimum in Eq. (2) is computed only over the defined terms. Finally, the DTW distance between the two time series is given by $d_{l,l'}$.

[7] Instead of $|x_i - x'_j|$, one can calculate $(x_i - x'_j)^2$ in Eq. (2). In our study, we used the variant with $|x_i - x'_j|$.

3.3. Dynamic convolution

In time series classifiers like ROCKET, convolution serves as a local pattern detector. As discussed in [10], the combination of convolution and max pooling provides limited flexibility in pattern matching by reducing sensitivity to the exact position of a pattern within the time series. Specifically, max pooling can only maintain consistent outputs when a pattern shifts within its pooling window. If a pattern shifts beyond this range, the detection may be affected. Additionally, conventional convolution is unable to handle more complex temporal distortions, such as elongations of local patterns. These limitations motivated the development of dynamic convolution, which enhances pattern matching by incorporating time-warping mechanisms.

The core concept of dynamic convolution is to replace the standard dot product (or inner product) used in conventional convolution with the computation of Dynamic Time Warping (DTW) distances between the convolutional kernel and segments of the time series. This approach allows for more flexible pattern matching by accounting for temporal distortions, such as shifts and elongations within local patterns.

3.4. ROCKET

The "ROCKET" time series classifier begins by generating a predefined number (F) of random convolutional kernels (filters). By default, ROCKET utilizes $F = 10,000$ convolutional kernels. The parameters of each kernel, including its length, weights, bias, and dilation, are randomly sampled from appropriate distributions. Specifically:

- *The window size* (w) of the convolutional kernel is chosen uniformly from $\{7, 9, 11\}$.
- *Weights* are drawn from a standard normal distribution (mean = 0, standard deviation = 1).
- *Bias* is sampled uniformly from the interval $(-1, 1)$.
- *Dilation* is set to $\lfloor 2^d \rfloor$, where d is selected uniformly from the range $(0, \lfloor \log_2((l - 1)/(w - 1)) - 1 \rfloor)$, with l representing the length of the input time series.
- *Zero padding* is applied with a probability of 0.5 for each convolution.

This randomized kernel generation process allows ROCKET to efficiently capture diverse temporal features while maintaining computational efficiency.

Given a *single* input time series (whether from the training or test set), ROCKET applies convolution with all F randomly generated kernels, producing F convoluted time series. For each of these F convoluted time series, ROCKET performs two global pooling operations:

- *Global Max Pooling:* Extracts the maximum value from the convoluted time series.
- *Global PPV Pooling:* Computes the *Proportion of Positive Values* (PPV), which represents the fraction of values in the convoluted time series that are positive.

As a result, each input time series is transformed into a feature vector containing $2F$ real-valued features (by default, 20,000 features when $F = 10,000$).

For classification, Dempster et al. [15] apply *ridge regression* when the dataset contains fewer than $2F$ time series or *logistic regression* when the dataset is larger.

Despite its simplicity, ROCKET has been shown to outperform various deep learning-based methods, including fully convolutional networks (FCNs). This strong performance

may be attributed to its ability to efficiently capture diverse temporal patterns through a large number of random convolutional kernels.

In their follow-up work, Dempster et al. [16] found that the features resulting from "PPV pooling" are more important than the ones resulting from max pooling. In this study, we use ROCKET-PPV and ROCKET-MAX to denote those versions of ROCKET that use only the features resulting from "PPV pooling" and max pooling, respectively.

Ruiz et al. [40] noted that the multivariate extension of ROCKET, at least the one included in the sktime software library, assigns kernels to channels randomly. Although this may be useful in certain cases, this approach inherently assumes that channels are independent. Therefore, we consider an alternative version of ROCKET for multivariate time series. In particular, we generate multivariate kernels. Whenever the opposite is not explicitly stated, we use ROCKET with multivariate kernels. In the experimental evaluation, we use "ROCKET-UNI" to denote the variant that generates univariate kernels and assigns those univariate kernels to channels randomly.

4. Our Approach

In this section, we describe in detail how the proposed approach enhances ROCKET.

Given ROCKET's F convolutional kernels, we propose to calculate dynamic convolution instead of conventional convolution and apply global max pooling to the convoluted time series.

Originally, dynamic convolution was defined for univariate time series. We handle multivariate time series as follows: we generate multivariate convolutional kernels and use them when calculating convolutions. In particular, for each segment of the time series, we calculate the DTW distances for each channel separately and sum the DTW distances of the different channels. This is illustrated in Fig. 1.

As mentioned previously, in case of the original ROCKET approach, global max pooling and PPV ("portion of positive values") pooling are applied to the convoluted time series. As DTW distances are nonnegative, we omit PPV pooling. In case of conventional convolution, the role of max pooling is to select the similarity (particularly, the value of the dot product) of the time series segment that is most similar to the pattern associated with the convolutional kernel. In contrast to dot product, in case of DTW, low values mean high similarity, and high values correspond to substantial difference between the segment and the pattern associated with the kernel. Therefore, we use *min pooling* instead of max pooling.

Calculating multivariate dynamic convolution with F kernels and applying global max pooling results in F features.

Using this representation of time series, according to which each time series is represented as an F-dimensional feature vector, similarly to ROCKET, we train ridge regression or logistic regression to classify time series. Ridge regression is used in case if the number of time series in the dataset is less than the number of features, while logistic regression is used otherwise.

The computational complexity of the calculation of dynamic convolution with window size w for a time series of constant length is $\mathcal{O}(w^2)$, while the complexity of "conventional" convolution is $\mathcal{O}(w)$. This is due to the fact that DTW calculations are quadratic in the length of the input time series whereas the scalar product can be calculated in linear

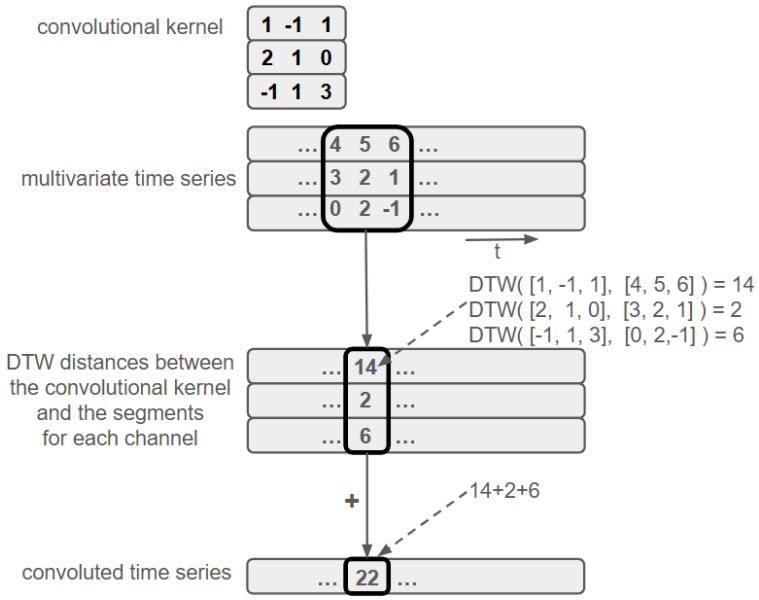


Fig. 1. Illustration of multivariate dynamic convolution

time. However, as detailed in the previous section, the window size w of the convolutional kernel is usually set to a small value, such as $w = 7, 9$ or 11, thus the actual runtime of the approach only increases moderately. On the other hand, for the case of large convolutional windows, we point out that there are various approaches to speed up DTW calculations, such as limiting the size of the warping window. (Please note that the warping window of DTW is different from the aforementioned convolutional window.) If setting the size of the warping window to a constant value, the theoretical complexity of DTW calculations becomes $\mathcal{O}(w)$ which is the same as the theoretical complexity of scalar product calculations in the "conventional" convolution.

5. Experimental Evaluation

The goal of our experiments is to assess the accuracy of the proposed approach for the recognition of epileptic tonic-clonic seizures and to compare it to relevant variants of ROCKET as well as to other prominent approaches from the literature.

5.1. Data

We performed experiments on two datasets[8], in particular: (i) the "Epilepsy" dataset [43] and (ii) the "OpenSeizure" dataset [28].

[8] See also https://timeseriesclassification.com/description.php?Dataset=Epilepsy and https://ieee-dataport.org/documents/open-seizure-database-v100 for more information about the datasets.

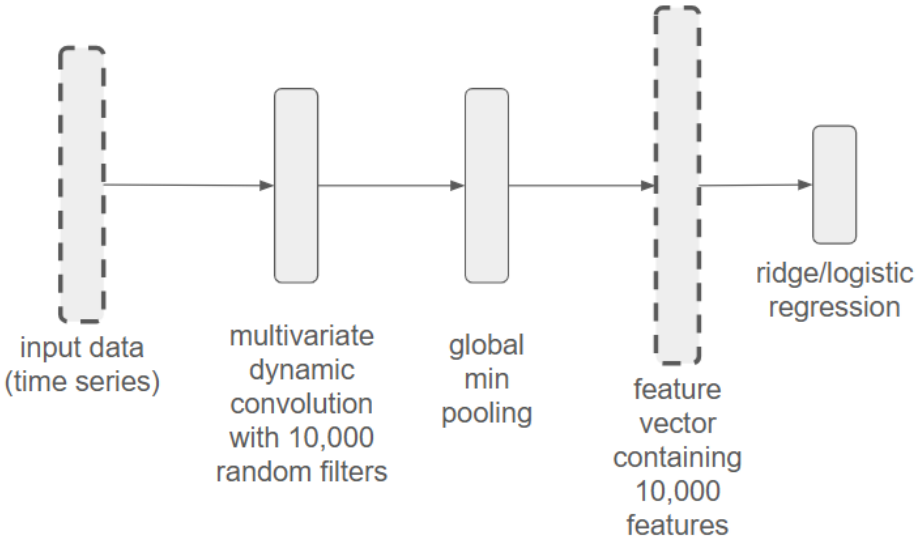


Fig. 2. Our approach, EROCKET, for the classification of multivariate time series. It takes a time series as input and performs multivariate dynamic convolution using F random kernels and global min pooling. This results in a vector of F features in total. This representation is used in ridge regression or logistic regression.

Epilepsy According to the description at timeseriesclassification.com, "the data contains multivariate time series measured by an accelerometer." It was collected from six healthy participants using a tri-axial accelerometer on the dominant wrist while they were conducting four different activities.

These activities are walking, running, sawing, and seizure mimicking. *Walking* includes different paces and gestures: walking slowly while gesturing, walking slowly, walking normally, and walking fast, each of which lasts for 30 seconds. *Running* refers to running a 40-meter-long corridor. *Sawing* was performed with a saw for 30 seconds. In the case of *seizure mimicking*, the participants were seated, and the seizure itself lasted for 30 seconds, but an additional 5-6 seconds before and 30 seconds after the mimicked seizure were recorded. Each participant performed each activity 10 times at least. The mimicked seizures were trained and controlled, following the protocol defined by a medical expert. All the activities were carried out indoors, either inside an office or in the corridor around it. The sampling frequency was 16 Hz. Some activities lasted about 30 seconds, others up to 2 minutes.

The data was truncated to the length of the shortest time series. Prefixes and suffixes of flat series were truncated to the shortest series ($\approx$13 seconds), taking a random interval of activity for series longer than the minimum. A single case from the original data (ID002 Running 16) was removed because the data was not collected correctly. After preprocessing, the data contains a total of 275 accelerometer time series, each of which has a length of $l = 206$. Out of the 275 time series, 68 belong to the class of seizure mimicking

(or epilepsy), while 74, 73, and 60 belong to the classes of walking, running, and sawing, respectively. One of the accelerometer time series from each of the classes is shown in Fig. 3.

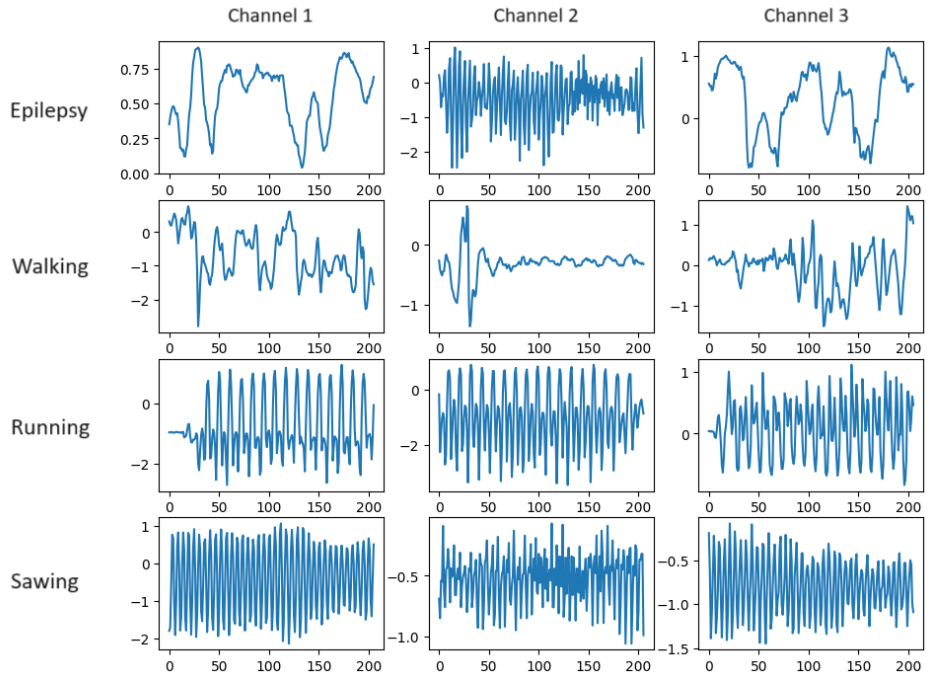


Fig. 3. An excerpt of the considered dataset: one of the accelerometer time series from each of the classes. Please note that we consider multivariate time series with three channels, each of which is shown separately.

OpenSeizure The OpenSeizure dataset is a "publicly accessible resource designed to advance non-electroencephalogram seizure detection research".[9] It contains multimodal sensor data observed in real-world, in-home environments. From the dataset, we considered tonic-clonic seizures that had associated 3-dimensional time series data. As the dataset contains labeled false alarms as well, we selected the same amount of false alarms with 3-dimensional time series. For each tonic-clonic seizure and for each of the selected false alarms, we selected 500 time points from the middle of the 3-dimensional time series. Each time series was normalized to have a mean of zero and a standard deviation of one. The script we used to preprocess the data is available in the same repository where we published the implementation of our approach, see Section 5.5.

[9] https://ieee-dataport.org/documents/open-seizure-database-v100

As we consider the classification task of distinguishing tonic-clonic seizures from false alarms, the OpenSeizure dataset is inherently more challenging compared to the Epilepsy dataset.

5.2. Baselines

Our approach is based on ROCKET, therefore, we compare its accuracy to ROCKET and its variants ROCKET-PPV, ROCKET-MAX and ROCKET-UNI (see Section 3.4). To obtain a more complete evaluation, we additionally examine other standard baseline time-series classifiers: (i) multilayer perceptron (MLP), (ii) a simple convolutional neural network (CNN), (iii) fully convolutional network (FCN), (iv) a residual network (ResNet), and (v) a classifier based on the transformer architecture (Transformer).

Although recent time series classifiers outperform multilayer perceptrons, we decided to include MLP in the experimental evaluation because feed-forward neural networks with at least one hidden layer and non-linear activation are known to be universal function approximators [5]. The MLP considered in our experiments contains a single hidden layer with 100 units and ReLU activation and an output layer with softmax activation.

The simple convolutional neural network (CNN) has a single convolutional layer, followed by a fully connected layer with 32 units and the output layer in which each unit corresponds to one of the classes. The convolutional layer contains 16 kernels with size of 32. We used softmax activation in the output layer and ReLU activation in the convolutional and fully connected layers.

The architectures of FCN and ResNet models are based on the study of Wang et al. [45], nevertheless, we omitted batch normalization, because we observed substantially higher accuracy without batch normalization both in case of FCN and ResNet.

Inspired by the success of transformer-based architectures for time series classification, see e.g. ShapeFormer [31], we implemented a computationally efficient transformer for time series classification. For simplicity, we refer to this model as Transformer. Our implementation of Transformer, as well as the implementations of other baselines are included in our code base, which we have made publicly available, see Section 5.5.

In case of the Epilepsy dataset, we trained the aforementioned neural networks (MLP, CNN, FCN, ResNet and Transformer) for 1000 epochs with a learning rate of 10^{-5} and used the Adam optimizer [29] with cross-entropy loss and set the batch size to 16. In case of the OpenSeizure dataset, we trained the neural networks for 100 epochs with a learning rate of 10^{-4} and set the batch size to 32. According to our observations, these settings allowed the neural networks to converge in all the examined cases.

5.3. Experimental Protocol

We performed experiments according to the 10×10-fold cross-validation protocol. With 10-fold cross-validation, we mean that we partition the data into 10 disjoint splits, one of these splits is used as test data, while the other 9 splits are used as training data. In the case of 10-fold cross-validation, the experiment (i.e., training and evaluation of the classifier) is repeated 10 times with a different split being used as test data each time. In case of 10×10-fold cross-validation, the original random splitting of the data is performed 10 times, therefore 10 instances of 10-fold cross-validation are performed in total.

5.4. Evaluation metrics

We used average classification error, i.e., the ratio of incorrectly classified time series, to assess the performance of our approach and the baselines. Lower values indicate better performance.

We report classification error averaged over the $10 \times 10 = 100$ folds of the cross-validation together with its standard deviation both for our approach and the baselines.

Additionally, we used paired t-test to assess whether the differences between our approach and the baselines – in terms of classification error – are statistically significant or not. We used the Nadeau and Bengio correction[10] to account for the overlapping among the repeated folds.

As low false alarm rates are of paramount importance, see e.g. [39], we consider the epilepsy class (seizure mimicking) as the positive class and report False Positive Rate (FPR), i.e., the fraction of false positives among all negative instances (time series belonging to other classes).

5.5. Implementation

We implemented our approach and the baselines in Python using the numpy and pytorch software libraries. As the dataset contains less than 10,000 time series, we used ridge regression – particularly the "RidgeClassifier" from the scikit-learn machine learning library – throughout the experiments, both in the case of our approach and other variants of ROCKET.

To calculate DTW distances quickly, we implemented DTW in Cython, which combines the efficiency of C with Python's rapid prototyping [4]. We executed the experiments in Google Colab.[11] To assist reproduction of our work, independent validation of the results and to facilitate follow-up works, we published the implementation of our approach (source codes) in our GitHub repository

https://github.com/kr7/seizure

in form of an IPython notebooks.

5.6. Experimental Results

Our results on the classification error are shown in Table 1. Our approach, EROCKET, outperforms all examined (deep) neural networks and other variants of ROCKET. We also report the significance levels (p-values) of the t-tests that measure whether the differences (in terms of classification error) between our approach (EROCKET) and its competitors are significant or not.

Furthermore, we consider false alarms, which can critically burden the acceptance of self-detection by a wider set of users, as mentioned in Section 1. Figure 4 shows the false positive rate (FPR) as a percentage in case of the Epilepsy dataset. As one can see, the proposed approach attains the smallest amount of false alarms, which makes it suitable for the examined application context.

[10] https://gist.github.com/jensdebruijn/13e8eeda85eb8644ac2a4ac4c3b8e732

[11] https://colab.research.google.com

Table 1. Average classification error, its standard deviation (after the $\pm$ sign, calculated over the 10×10 folds) and the significance level (p-value) of the t-test that measures whether the difference (in terms of classification error) between our approach (EROCKET) and its competitor is significant or not.

Approach	**Epilepsy dataset**		**OpenSeizure dataset**	
	classification error	p-value	classification error	p-value
MLP	24.11% $\pm$ 13.36	≈ 0	43.61% $\pm$ 12.64	0.0019
CNN	10.52% $\pm$ 5.07	4.30×10^{-6}	41.30% $\pm$ 13.99	0.0050
FCN	2.16% $\pm$ 2.59	0.0820	26.02% $\pm$ 12.53	0.5069
ResNet	5.60% $\pm$ 4.32	0.0012	25.05% $\pm$ 12.36	0.5362
Transformer	3.82 % $\pm$ 3.36	0.0073	36.04 % $\pm$ 14.93	0.0232
ROCKET	1.74% $\pm$ 2.20	0.1326	30.66% $\pm$ 12.51	0.1271
ROCKET-UNI	2.77% $\pm$ 2.92	0.0324	27.67% $\pm$ 12.81	0.4053
ROCKET-PPV	1.54% $\pm$ 2.54	0.2484	31.22% $\pm$ 13.42	0.1102
ROCKET-MAX	2.61% $\pm$ 3.18	0.0596	30.92% $\pm$ 12.65	0.1218
EROCKET (our approach)	**0.40% $\pm$ 1.14**		**22.84% $\pm$ 12.16**	

We also examined the performance of our approach, EROCKET, and two of the strongest baselines, ROCKET and FCN in the presence of noise in the data. In particular, we added zero-mean Gaussian noise with various standard deviations $\sigma \in \{0, 0.1, 0.25\}$ to the time series of the Epilepsy dataset and repeated the experiments with noisy data. Noise was added to both training and test data, so that our experiments simulate the real-world scenario when the sensor provides noisy data. The results are shown in Fig. 5. As one can see, with increasing noise level, the performance of all the approaches decrease, nevertheless, our approach, EROCKET, consistently outperforms its competitors in case of noisy data as well.

5.7. Discussion

The primary goal of our study was to develop a classification algorithm that outperforms state-of-the-art time series classifiers for the examined research problem. Our results are encouraging and show that the proposed classifier can be effective in the context of a self-detection mechanism of epileptic tonic-clonic seizures. Nevertheless, it cannot be claimed that similar accuracy could be observed in a clinical setting: the Epilepsy dataset contains only four types of activities and it might be more challenging to distinguish tonic-clonic seizure from *any* other activity types. On the other hand, distinguishing seizures from false alarms in case of the OpenSeizure dataset is a rather challenging task. Thus, we should interpret the results in this paper as a feasibility study regarding the use of advanced machine learning algorithms, whereas a more thorough clinical application may be subject of future works. Moreover, aspects related to privacy-preserving training of EROCKET, such as training according to the federated learning paradigm [34], comprise another possible extension.

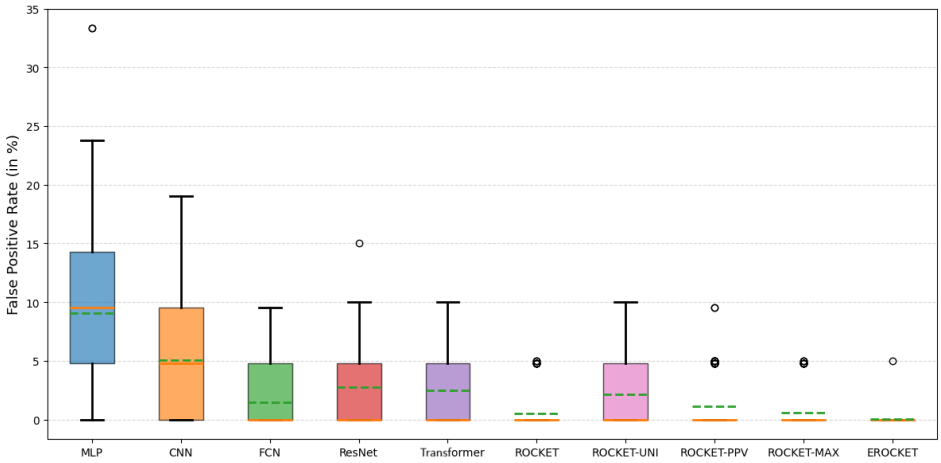


Fig. 4. Boxplots of false positive rates (FPR), measured as a percentage, for the proposed approach (EROCKET) and its competitors in case of the Epilepsy dataset. The mean FPR (over the 10×10 folds) is shown by the dashed line.

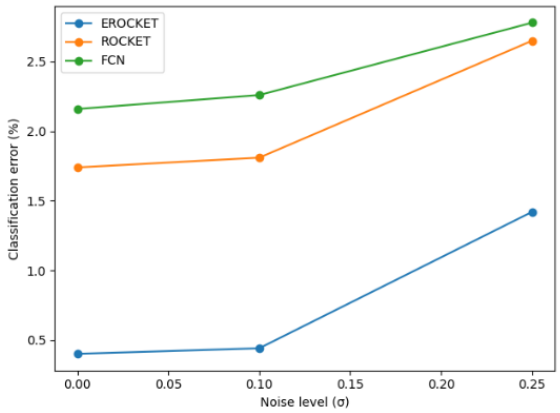


Fig. 5. Classification error of our approach (EROCKET), ROCKET and FCN in case of noisy data.

6. Conclusions

In this study, we focused on the recognition of epileptic tonic-clonic seizures based on accelerometer data. We cast this task as a time series classification problem for which recent approaches are based on "Random Convolutional Kernel Transform" (ROCKET). Considering two publicly available time series datasets containing sensor data, we observed that ROCKET achieves relatively accurate classification on its own, especially in case of

the Epilepsy dataset. Nevertheless, in medical and healthcare applications, high accuracy and low FPR are essential, therefore, we enhanced ROCKET by incorporating multivariate dynamic convolution into ROCKET. This way, our approach, EROCKET, achieved an significantly lower classification error compared to the original ROCKET algorithm. Our approach not only outperforms ROCKET, but other time series classifiers as well, both in terms of classification error and FPR. Furthermore, we point out that EROCKET may be used in other applications of multivariate time-series classification, such as the classification of electrocardiograph signals [13], user verification based on mouse dynamics [2] or person authentication based on keystroke dynamics [37].

References

1. Ahmad, I., Wang, X., Javeed, D., Kumar, P., Samuel, O.W., Chen, S.: A hybrid deep learning approach for epileptic seizure detection in EEG signals. IEEE Journal of Biomedical and Health Informatics (2023)
2. Antal, M., Denes-Fazakas, L.: User verification based on mouse dynamics: a comparison of public data sets. In: 2019 IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI). pp. 143–148. IEEE (2019)
3. Baumgartner, C., Koren, J.P.: Seizure detection using scalp-EEG. Epilepsia 59, 14–22 (2018)
4. Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D.S., Smith, K.: Cython: The best of both worlds. Computing in Science & Engineering 13(2), 31–39 (2010)
5. Bengio, Y., Goodfellow, I., Courville, A.: Deep learning, vol. 1. MIT press Cambridge, MA, USA (2017)
6. Beniczky, S., Polster, T., Kjaer, T.W., Hjalgrim, H.: Detection of generalized tonic–clonic seizures by a wireless wrist accelerometer: a prospective, multicenter study. Epilepsia 54(4), e58–e61 (2013)
7. Berg, A.T., Shinnar, S.: The risk of seizure recurrence following a first unprovoked seizure: a quantitative review. Neurology 41(7), 965–965 (1991)
8. Bruno, E., Viana, P.F., Sperling, M.R., Richardson, M.P.: Seizure detection at home: Do devices on the market match the needs of people living with epilepsy and their caregivers? Epilepsia 61, S11–S24 (2020)
9. Buza, K.: Activity recognition based on accelerometer data with enhanced rocket algorithm. In: 18th International Symposium on Applied Computational Intelligence and Informatics. pp. 321–326. IEEE (2024)
10. Buza, K., Antal, M.: Convolutional neural networks with dynamic convolution for time series classification. In: Advances in Computational Collective Intelligence: 13th International Conference, ICCCI 2021, Kallithea, Rhodes, Greece, September 29–October 1, 2021, Proceedings 13. pp. 304–312. Springer (2021)
11. Buza, K., Antal, M.: Rocket with dynamic convolution for time series classification. In: International Conference on Computational Collective Intelligence. pp. 271–282. Springer (2024)
12. Buza, K., Koller, J., Marussy, K.: PROCESS: projection-based classification of electroencephalograph signals. In: Artificial Intelligence and Soft Computing: 14th International Conference, ICAISC 2015, Zakopane, Poland, June 14-18, 2015, Proceedings, Part II. pp. 91–100. Springer (2015)
13. Buza, K., Nanopoulos, A., Schmidt-Thieme, L., Koller, J.: Fast classification of electrocardiograph signals via instance selection. In: 2011 IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology. pp. 9–16. IEEE (2011)
14. Chen, W., Shi, K.: A deep learning framework for time series classification using relative position matrix and convolutional neural network. Neurocomputing 359, 384–394 (2019)

15. Dempster, A., Petitjean, F., Webb, G.I.: Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining and Knowledge Discovery 34(5), 1454–1495 (2020)
16. Dempster, A., Schmidt, D.F., Webb, G.I.: Minirocket: A very fast (almost) deterministic transform for time series classification. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. pp. 248–257 (2021)
17. Esmael, B., Arnaout, A., Fruhwirth, R.K., Thonhauser, G.: Improving time series classification using hidden markov models. In: 2012 12th International Conference on Hybrid Intelligent Systems (HIS). pp. 502–507. IEEE (2012)
18. Faes, L., Islam, M., Bachmann, L.M., Lienhard, K.R., Schmid, M.K., Sim, D.A.: False alarms and the positive predictive value of smartphone-based hyperacuity home monitoring for the progression of macular disease: a prospective cohort study. Eye 35, 3035–3040 (2021)
19. Farooq, M.S., Zulfiqar, A., Riaz, S.: Epileptic seizure detection using machine learning: Taxonomy, opportunities, and challenges. Diagnostics 13(6), 1058 (2023)
20. Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Deep learning for time series classification: a review. Data Mining and Knowledge Discovery 33(4), 917–963 (2019)
21. Fisher, R.S., Blum, D.E., DiVentura, B., Vannest, J., Hixson, J.D., Moss, R., Herman, S.T., Fureman, B.E., French, J.A.: Seizure diaries for clinical research and practice: limitations and future prospects. Epilepsy & Behavior 24(3), 304–310 (2012)
22. Gabeff, V., Teijeiro, T., Zapater, M., Cammoun, L., Rheims, S., Ryvlin, P., Atienza, D.: Interpreting deep learning models for epileptic seizure detection on EEG signals. Artificial intelligence in medicine 117, 102084 (2021)
23. Geller, Z., Kurbalija, V., Ivanović, M.: FAP: A time series analysis and mining framework for scientific and practical applications. Computer Science and Information Systems 22(4), 1379–1403 (2025)
24. Grabocka, J., Wistuba, M., Schmidt-Thieme, L.: Fast classification of univariate and multivariate time series through shapelet discovery. Knowledge and information systems 49, 429–454 (2016)
25. Hüsken, M., Stagge, P.: Recurrent neural networks for time series classification. Neurocomputing 50, 223–235 (2003)
26. Jaiswal, A.K., Banka, H.: Epileptic seizure detection in EEG signal using machine learning techniques. Australasian physical & engineering sciences in medicine 41, 81–94 (2018)
27. Jankowski, D., Jackowski, K., Cyganek, B.: Learning decision trees from data streams with concept drift. Procedia Computer Science 80, 1682–1691 (2016)
28. Jones, G., Pordoy, J.: Open seizure database v1.0.0. IEEE Dataport (2023), https://dx.doi.org/10.21227/jftq-3e97
29. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
30. Ksiazek, K., Masarczyk, W., Glomb, P., Romaszewski, M., Buza, K., Sekula, P., Cholewa, M., Kolodziej, K., Gorczyca, P., Piegza, M.: Deep learning approach for automatic assessment of schizophrenia and bipolar disorder in patients using rr intervals. PLOS Computational Biology 21(9), e1012983 (2025)
31. Le, X.M., Luo, L., Aickelin, U., Tran, M.T.: Shapeformer: Shapelet transformer for multivariate time series classification. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1484–1494 (2024)
32. Lines, J., Davis, L.M., Hills, J., Bagnall, A.: A shapelet transform for time series classification. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 289–297 (2012)
33. Liu, T., Truong, N.D., Nikpour, A., Zhou, L., Kavehei, O.: Epileptic seizure classification with symmetric and hybrid bilinear models. IEEE Journal of Biomedical and Health Informatics 24(10), 2844–2851 (2020)

34. Lupión, M., Sanjuan, J.F., Medina-Quero, J., Ortigosa, P.M.: Epilepsy seizure detection using low-cost iot devices and a federated machine learning algorithm. In: International Symposium on Ambient Intelligence. pp. 229–238. Springer (2022)
35. Onorati, F., Regalia, G., Caborni, C., Migliorini, M., Bender, D., Poh, M.Z., Frazier, C., Kovitch Thropp, E., Mynatt, E.D., Bidwell, J., et al.: Multicenter clinical assessment of improved wearable multimodal convulsive seizure detectors. Epilepsia 58(11), 1870–1879 (2017)
36. Pavlovic, V., Frey, B.J., Huang, T.S.: Time-series classification using mixed-state dynamic bayesian networks. In: Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149). vol. 2, pp. 609–615. IEEE (1999)
37. Peška, L., Veselỳ, P., Skopal, T., Buza, K.: Person authentication using visual representations of keyboard typing dynamics. In: 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS). pp. 1–6. IEEE (2022)
38. Radovanović, M., Nanopoulos, A., Ivanović, M.: Time-series classification in many intrinsic dimensions. In: Proceedings of the 2010 SIAM International Conference on Data Mining. pp. 677–688. SIAM (2010)
39. Regalia, G., Onorati, F., Lai, M., Caborni, C., Picard, R.W.: Multimodal wrist-worn devices for seizure detection and advancing research: focus on the empatica wristbands. Epilepsy research 153, 79–82 (2019)
40. Ruiz, A.P., Flynn, M., Large, J., Middlehurst, M., Bagnall, A.: The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery 35(2), 401–449 (2021)
41. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing 26(1), 43–49 (1978)
42. Shah, S., Gutierrez, E.G., Hopp, J.L., Wheless, J., Gil-Nagel, A., Krauss, G.L., Crone, N.E.: Prospective multicenter study of continuous tonic-clonic seizure monitoring on apple watch in epilepsy monitoring units and ambulatory environments. Epilepsy & Behavior 158, 109908 (2024)
43. Villar, J.R., Vergara, P., Menéndez, M., de la Cal, E., González, V.M., Sedano, J.: Generalized models for the classification of abnormal movements in daily life and its applicability to epilepsy convulsion recognition. International journal of neural systems 26(06), 1650037 (2016)
44. Wang, X., Chuncao, L., Liu, Y., Liang, W., Kuanching, L., Poniszewska-Maranda, A.: A spatio-temporal graph neural network for EEG emotion recognition based on regional and global brain. Computer Science and Information Systems 22(3), 971–989 (2025)
45. Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: A strong baseline. In: 2017 International joint conference on neural networks (IJCNN). pp. 1578–1585. IEEE (2017)
46. Xi, X., Keogh, E., Shelton, C., Wei, L., Ratanamahatana, C.A.: Fast time series classification using numerosity reduction. In: Proceedings of the 23rd international conference on Machine learning. pp. 1033–1040 (2006)
47. Ye, L., Keogh, E.: Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. Data mining and knowledge discovery 22, 149–182 (2011)
48. Zhao, B., Lu, H., Chen, S., Liu, J., Wu, D.: Convolutional neural networks for time series classification. Journal of Systems Engineering and Electronics 28(1), 162–169 (2017)
49. Zheng, Y., Liu, Q., Chen, E., Ge, Y., Zhao, J.L.: Time series classification using multi-channels deep convolutional neural networks. In: International conference on web-age information management. pp. 298–310. Springer (2014)
50. Zhou, M., Tian, C., Cao, R., Wang, B., Niu, Y., Hu, T., Guo, H., Xiang, J.: Epileptic seizure detection based on EEG signals and cnn. Frontiers in neuroinformatics 12, 95 (2018)

Krisztian Buza is an Associate Professor at the Budapest University of Economics and Business. He received his diploma degree in computer science from the Budapest University of Technology and Economics in 2007 and earned his Ph.D. from the University of Hildesheim in 2011. In addition to his academic career, he has gained industrial experience in the automotive sector, where he worked as a Senior Deep Learning Engineer, contributing to the development and application of machine learning and deep learning solutions in industry. Dr. Buza is the co-author of more than 50 scientific publications and received the Best Paper Award at the IEEE Conference on Computational Science and Engineering in 2010. He regularly serves as a reviewer for leading journals and conferences, including Neurocomputing, Knowledge-Based Systems, Knowledge and Information Systems, the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, and the Pacific-Asia Conference on Knowledge Discovery and Data Mining. His research interests include machine learning and data mining, with a particular focus on time series classification and biomedical applications.

Alexandros Nanopoulos is a Professor of Business Informatics at Baden-Württemberg Cooperative State University, with extensive experience in data science and AI. He holds a PhD in Data Mining from Aristotle University of Thessaloniki and a Diploma in Computer Science from the same institution. With a career spanning over two decades, Dr. Nanopoulos has held senior roles at leading organizations in the industry, where he developed AI-driven solutions and data products focused on predictive modeling, price optimization, demand forecasting, and more. He has also worked as consultant on projects related to the automotive industry, and has contributed to academic research as a junior professor at the University of Hildesheim and Catholic University Eichstätt- Ingolstadt. Dr. Nanopoulos has authored more than 130 research papers and has served on the program committees of top-tier data mining conferences such as ACM KDD and ECML. His work includes the development of innovative machine learning algorithms and contributions to areas like recommendation systems, social media mining, and data mining for large-scale applications.

Noemi Agnes Varga is a clinical geneticist and a neurology fellow in the final year of training. Since completing medical school, Dr. Varga has focused on neuroscience, conducting significant research on the genetic background of autism spectrum disorders during a PhD at the Semmelweis University Szentagothai János Doctoral School. With a keen interest in understanding common neurological disorders, such as epilepsy and recognizing rare neurogenetic diseases that present with common symptoms, Dr. Varga is dedicated to advancing deeper understanding of common pathways in the neurology.

Received: September 13, 2025; Accepted: March 5, 2026.

Computer Science and Information Systems 23(2):775–800 https://doi.org/10.2298/CSIS251030023C

Research on Key Technology of Network Information Extraction Oriented to Web Topic Detection for Big Data

Mo Chen

Business College of Beijing Union University,
A3, Yanjingdongli, Chaoyang District, Beijing, 100025, P.R. China
mo.chen@buu.edu.cn

Abstract. In the context of today's big data and numerical intelligence era, this study explores an incremental network information extraction technology for Web topic detection characterized by the semi-structured or unstructured big data as important research object to promote network information detection application. This study takes Web big data as the main research object and proposes an incremental network information extraction idea for Web topic detection. In this idea, the designed algorithm of theme similarity measurement for incremental network information extraction can extract Web instances related to theme, and calculate importance of Web instances related to theme, furthermore, the designed algorithm of incremental instance extraction for Web topic detection can analyze Pattern and BasePattern according to extracted Web instance URL, and conduct segmentation for Web instance title and text content, extract keywords, which are capable of describing Web topic. Experimental results demonstrate that the framework, method, and algorithm proposed in this paper significantly outperform traditional methods in network information extraction. Particularly, the accuracy rate of extracted Web instances that are similar to the theme can reach 0.833, the F-Measure value of extracted Web instances that are similar to the theme under different threshold adjustment is close to 0.83, the accuracy rate of topic detection under the condition of determining the number of Web news instances extracted, the threshold and the parameter value is close to 0.82. The study concludes that the incremental network information extraction idea proposed in this paper is feasible, verifiable, and superior, and can play an important role in reconfiguring numerical intelligence warehouses for detecting Web topic, inferring the Web hierarchical big data propagation path.

Keywords: Incremental Network Information Extraction, Big Data, Web Topic Detection.

1. Introduction

Against the backdrop of the development of big data era, scholars are still exploring and innovating in the face of the constantly emerging heterogeneous big data and diverse demands for numerical intelligence applications [1,2,3,4]. At this stage, the network has been providing the most valuable information to various users, while the amount of network data is also growing at an astonishing rate [5,6,7]. Therefore, scholars urgently need to think about how to extract more accurate knowledge from complex big data. In this exploration process, the incremental network information extraction for Web topic detection can be an important research direction.

In the network heterogeneous big data, Internet news, as a streaming resource, has the characteristics of real-time update, wide dissemination, and high interaction from the perspective of application [8,9,10]. With the continuous occurrence of various events, the number of Internet news is showing an explosive growth trend. From a scientific perspective, it has shown the 5V characteristics of volume, variety, value, velocity and veracity for big data [11,12,13,14]. Based on the above characteristics, how to study the incremental network information extraction method for Web topic detection, it has become an urgent problem to build a numerical intelligence warehouses and provide a real-time big data source for the network information detection application.

To establish a strong foundation for the research of this paper, the next section will review existing methods related to network information extraction, and highlight their strengths and limitations. Based on the study of literature, this paper will propose an idea for in-depth exploration of the process of network information extraction, and elaborate on the main implementation methods and algorithms for Web topic detection.

This paper will make three key contributions: (1) it introduces a novel network information extraction method for Web topic detection based on big data, (2) it optimizes the process for theme similarity measurement, instance information extraction, link importance calculation, filter mode analysis and so on, (3) it provides extensive experimental validation and demonstrates the effectiveness of the proposed approach.

2. Related works

Oriented to Web application for topic detection, it needs a lot of real corpus as support. However, at present, the published Web news is showing a massive increase trend facing frequent social events [15,16,17]. So, this paper intends to use Web news source material as the research object, and discuss the effective process of extracting Web news. Some scholars have done certain research about the technology of network information extraction shown in Table 1, and these research results will be a foundation for continuing to research key technology of network information extraction for Web topic detection based on big data.

An effective paradigm is provided by the pseudo-labeling based semi-supervised learning algorithm [18], in order to alleviate the reliance on labeled data by leveraging unlabeled data. However, in the task for key information extraction, the main challenges for this algorithm are as follows, the context dependency of key information extraction results in incorrect pseudo-labels, and the intra-class variance is high, the inter-class variation is low. To this end, a similarity matrix pseudo-label bias rectification semi-supervised method is also proposed for key information extraction task, which improves the quality of pseudo-labels on key information extraction benchmarks with rare labels. More specifically, the similarity matrix bias rectification module is designed, which utilizes the contextual information of key information extraction data through the analysis of similarity between labeled and unlabeled data, in order to improve the quality of pseudo-labels. Moreover, a dual branch adaptive alignment mechanism is also designed, in order to adaptively align intra-class variance and alleviate inter-class variation on key information extraction benchmarks, which is composed of two adaptive alignment ways, one is the intra-class alignment branch, which is designed to adaptively align intra-class variance, the other one is the inter-class alignment branch, which is developed to adaptively

alleviate inter-class variance changes on the representation level. The whole research process does main contribution in the area of key information extraction, and the extensive experiment results on two benchmarks demonstrate that the pseudo-label bias rectification achieves state-of-the-art performance and its performance surpasses the previous research by 0.0211.

A solution of keyphrase extraction is proposed during learning process for personalized recommendation [19], in order to represent multi-view knowledge. In this solution, the structural features and section texts obtained from the section structure information are utilized, in order to extract keyphrase. The approach proposed consists of two main parts, one part explores the effect of structural features on keyphrase extraction models, and the other part integrates the extraction results from all section texts used as input corpora for keyphrase extraction models via a keyphrase integration algorithm to obtain the keyphrase integration result. Furthermore, the effect is also examined for the classification quality of section structure on keyphrase extraction performance. The approach proposed is different from the traditional information extraction method, the results show that incorporating structural features improves keyphrase extraction performance, though different features have varying effects on model efficacy, and the keyphrase integration approach yields the best performance, the classification quality of section structure can affect keyphrase extraction performance, then these findings indicate that using the section structure information contributes to effective keyphrase extraction. So, the whole research result does main contribution in the area of information extraction, and addresses that the length of the research objective content is limited, the noise information is introduced for the research objective content, the keyphrase extraction performance diminishes.

A method for joint entity and relation extraction is proposed [20], in this method, the tasks of entity extraction and relation classification are integrated by sharing the encoding layer. However, this method faces challenges due to incongruities in the contextual information captured by these subtasks, resulting in potential feature conflicts and adverse effects on model performance. To address this problem, a novel joint entity and relation extraction method is introduced that incorporates multi-module feature information enhancement, and a relation awareness enhancement module is employed for the entity extraction task. In this module, the model's focus is directed towards extracting entities closely related to potential relations using a potential relation extraction technology and an attention mechanism. For the relation extraction task, an entity information enhancement module is implemented that uses entity extraction results to augment the original feature information through a gating mechanism, thereby enhancing relation classification performance. The whole research process does main contribution in the method of information extraction for entity and relation, and the experiments on the multi-source datasets demonstrate that the method proposed performs well. Compared to the state-of-the-art method, the F1 score improves by 0.007.

A model is constructed for investigating how to adapt BERT to Chinese text research on reducing data volume process oriented to text classification [21], during the construction process for this model based on deep learning framework and dataset, Firstly, the Chinese corpus is compiled from various sources, the Web pages are searched through Google and Baidu search engine API belonging to the Chinese text research domain, the articles are collected through the API library provided by Wikipedia, the scientific literature written in Chinese are downloaded from CNKI, WanfangDATA, and CQVIP based

on defined query keywords, a large number of documents for archived files are stored. Next, the domain Chinese word segmentation is finished based on construction of domain pre-training corpus, a recent work informask is proposed, which optimizes the masking strategy. Next, the Pre-trained Language Model is developed, in order to further improve the performance of downstream tasks and maximize the value of a large amount of unlabeled domain data. The whole research process does main contribution in the method of text mining and information extraction for domain knowledge, and the effectiveness of the model is validated using different downstream tasks such as named entity recognition, relation extraction, and event extraction, which can perform better than general models and promote information extraction and knowledge discovery from Chinese text.

Inspired by human reasoning, a graph-based multitask information extraction framework is presented that facilitates the interaction between several information extraction tasks capable of capturing both local and global information [22]. In this framework, the graphs are constructed by selecting the most confident entity spans and coupling them with a confidence-weighted relation type and a confidence-weighted coreference. Additionally, a dynamic span graph approach is employed, where span updates are propagated across both the coreference and the relation graph, which allows useful information to be learned from a broader context by enhancing interaction across different information extraction tasks. The input data are globally shared, and the interaction between subtasks is fully exploited, in order to avoid cascading errors. The whole research process does main contribution in the method of information extraction, and the experiments demonstrate that the proposed multitask information extraction framework outperforms the state-of-the-art in multiple information extraction tasks spanning a variety of datasets. The framework proposed is different from the traditional information extraction framework, which is shown to achieve state-of-the-art results on multiple information extraction tasks across various domains and the framework's ability to enhance interaction across tasks allows it to learn valuable information from a broader context.

Based on the above research status for the information extraction area, it can be summed up that most studies have adopted following methods including the induction way based on the wrapper, the extraction way based on the Web query, the extraction way based on the Ontology, the processing way based on the natural language, and the extraction way based on the HTML structure and so on. However, the above process cannot fully consider how to design a wrapper set, in order to extract instances of different categories or theme, how to express universal and effective extraction rules, and how to iteratively learn the structure of extraction target. If above details can be studied in depth, the complexity for network information extraction process can be reduced, the extraction accuracy for network information can be improved, the research result can play an important role in describing Web topic, reconfiguring the Web topic corpus, inferring Web hierarchical big data propagation path, and providing an intelligent big data warehouse for network information detection application. To solve the research problem, this paper will propose innovative incremental element extraction method based on the theme similarity measurement, in order to describe Web topic, the next section will complete the problem definition and highlight the problem research boundaries.

Table 1. The related research status

The research method	The research limitation	The research disadvantage	The proposed methodology
The induction way based on the wrapper [18]	A wrapper can only handle instances of one category, it is necessary for a wrapper set, in order to extract instances of different categories.	The scalability is lack	The algorithm of theme similarity measurement
The extraction way based on the Web query [19]	The extraction rules need to be expressed in the XSLT way	The complexity is high for the process of the algorithm implementation, and its application is lack.	The algorithm of incremental instance extraction for Web topic detection
The extraction way based on the Ontology [20]	There are specific requirements for the structure of extraction target	The scalability is lack	The algorithm of theme similarity measurement and incremental instance extraction for Web topic detection
The processing way based on the natural language [21]	The massive instances need to be learned, in order to get effective extraction rules.	The process is difficult for automatic extraction	The algorithm of incremental instance extraction for Web topic detection
The extraction way based on the HTML structure [22]	The hyperlinks are unable to be processed, so the information is only able to be extracted with obvious range structure.	The generality is lack	The algorithm of theme similarity measurement and incremental instance extraction for Web topic detection

3. Problem definition in incremental network information extraction

From a global perspective, every Web news report can be viewed an instance node in the authoritative news network. This instance node can also link multiple related instance nodes, therefore, social events supported by a set of instance nodes can be considered as a theme. This theme belongs to column node of Web news network again, so different dimension can link theme and multiple Web news instances, and it can be regarded as a hierarchical node. However, from a local perspective, when analysing structural characteristics of a Web news instance separately, it can be found that it usually contains two parts. One part is text information related to Web news reports, and the other part is noise information which is not related to Web news reports. Therefore, if Web news instances and their relationships can be deployed in tree structure, and noise information which is not related to Web news content can be filtered in process of extracting Web news information, it will provide a hierarchical and high-quality Web news corpus for continuing to analyse Web news.

In text messages related to Web news reports, its content has presented unstructured characteristics, and it has a certain degree of difficulty for information extraction process. By analysing release template used in Web news, it can be seen that there are two parts in ¡title¿ element tag. One part is headline of Web news, another part is name of issuing organization, and two parts are separated with underline. The headings of Web news are also included in ¡h1¿ element tag, Web news release time and source are included in ¡div¿ element tag, and the text of Web news is included in ¡p¿ element tag. When analysing release template used in Web news, it can be found that the location of Web news instance data item can be determined from perspective of perceived styling features for Web news content. For example, the important content that needs to be highlighted is usually controlled by tagging elements of ¡strong¿ and ¡h1¿ in Web news. When analysing template used to publish multiple Web news on different websites, some templates can be found to have something in common. For example, the text of Web news content can be deployed in element tags of ¡p¿ and ¡div¿, and it has a certain length. Therefore, if unstructured content of Web news can be stored with semi-structured content in process of extracting Web news information, it will provide a semi-structured and high quality Web news corpus for continuing to analyse Web news.

This paper primarily defines data structures for NewsSet, HyperLinkSet, UrlSet, TopKeyWordSet, InitialUrlQueue and WaitingUrlQueue, as shown in Table 2. In addition to the defined data structures, this paper also defines Pattern and BasePattern. Pattern is a filter mode, which can filter Web news URL of sublayer. BasePattern is a base filter mode, which can filter Web news URL of brotherhood. Effective sublevel link groups can be presented with $UV = \{uv_1, uv_2, \ldots, uv_n\}$, invalid sublevel link groups can be presented with $UN = \{un_1, un_2, \ldots, un_n\}$ corresponding to toplevel link groups extracted from UrlSet set, which is $U = \{u_1, u_2, \ldots, u_n\}$. If filter mode Pattern exists, UV and UN can be detected, and UN can be filtered. If filter mode $BasePattern \in uv_i$ $(i = 1, 2, \ldots, n)$ exists, and there is no its submode that belongs to BasePattern or uv_i corresponding to valid sublayer link groups filtered from toplevel link groups, BasePattern is base filter mode of sublayer link groups.

The problem that need to be solved by the incremental network information extraction method for Web topic detection is as follows, the instances are extracted from mas-

Table 2. The data structure definition

The data structure name	The data structure description	The data structure representation
NewsSet	A URL set of Web news The seed big data source extracted from Web news information The massive and authoritative URL in Web news network	{ns1, ns2, ns3, …, nsi-1, nsi, nsi+1, …, nsn}
HyperLinkSet	The hyperlinks of massive instances for Web news contained in NewsSet	{hlsi1, hlsi2, hlsi3, …, hlsi(j-1), hlsij, hlsi(j+1), …, hlsim}
UrlSet	The big data source extracted from Web news information The massive and authoritative instances in Web news network	{us1, us2, us3, …, usi-1, usi, usi+1, …, usn}
TopKeyWordSet	A theme set of Web news The theme extracted from Web news information The keywords in social events that occur	{tkws1, tkws2, tkws3, …, tkwsi-1, tkwsi, tkwsi+1, …, tkwsn}
InitialUrlQueue	An initial queue for storing Web news URL	{iuq1, iuq2, iuq3, …, iuqi-1, iuqi, iuqi+1, …, iuqn}
WaitingUrlQueue	A pending queue for storing Web news URL	{wuq1, wuq2, wuq3, …, wuqi-1, wuqi, wuqi+1, …, wuqn}

sive Web news related theme, and the filtering mode from parent layer to sub layer and base filtering mode in brotherhood layer can be detected from these instances, in order to transfer tree structure from graph structure for interlacing complex URL network system. In this process, when the specific Web news instances belonging to certain topics are unknown, manually filtering each instance will inevitably increase computational complexity. Therefore, the problem that needs to be solved reflects uncertainty. If the set of Web news instances to be extracted is ns1, ns2, ns3, ..., nsi-1, nsi, nsi+1, ..., nsn, then each Web news instance can be considered as a node in the set. If the extraction reduction is designed in the path of extracting objects, the length of the extraction path can be shortened. In the case where each URL of a Web news instance can uniquely match the node nsi in the set of Web news instances, there is no polynomial time complexity algorithm to solve the proposed problem. The result can be described with TopWebNews=twn1, twn2, twn3, …, twni-1, twni, twni+1, …, twnk, the range for parameter i value is from one to k. The twni.url saves address for Web news, twni.title saves title for Web news, twni.pubtime saves release time for Web news, twni.pubsource saves release source for Web news, twni.content saves text for Web news, twni.dividedtitle saves segmentation result for Web news headline, twni.dividedcontent saves segmentation result for Web news text, twni.contentkeyword saves keywords for Web news content, twni.relativityvalue saves similarity between Web news content and theme, twni.parenturl saves address of parent node for Web news instances, twni.pattern saves description for Web news instance filtering mode, twni.systemtime saves system time extracted for instances.

Given the problem definition for incremental network information extraction in Section 3, the next section will introduce the framework, method, and algorithm of incremen-

tal network information extraction for Web topic detection based on big data, which aims to improve computational efficiency and accuracy.

4. Proposed Methodology: The incremental network information extraction for Web topic detection

In view of frequent events in society, the released Web news has reached at least NB level, and has shown characteristics of 5V big data [23,24,25]. Based on above problem definition, this paper proposes an incremental network information extraction method for Web topic detection, as shown in Fig. 1.

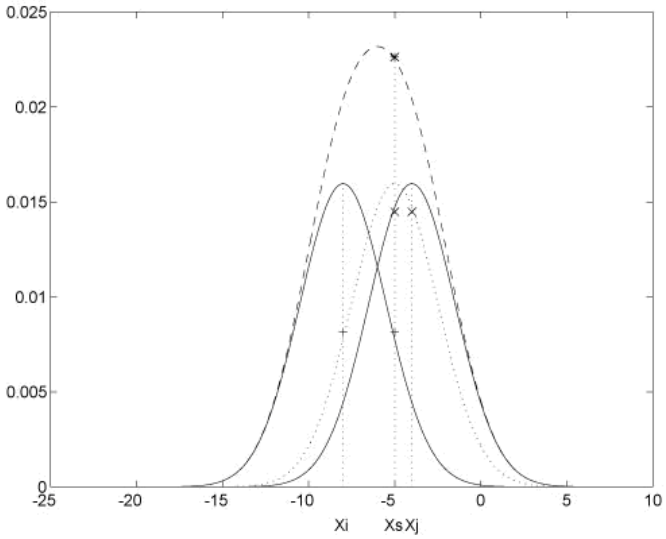


Fig. 1. The framework of incremental network information extraction

This framework completes the incremental corpus extraction for Web topic detection through using set of Web news and theme words and so on, this framework can measure Web news theme similarity. Through using selected Web news instance URL and source code related to theme, this framework can extract Web news instance information. Through using queue of InitialUrl and WaitingUrl, this framework can calculate importance of Web news links. Through using extracted Web news instance information, this framework can analyse mode of filtering and base filtering, and under background of theme, this framework can extract Web news instances incrementally, the result of keyword extracted can describe Web topic. In a word, this paper can effectively extract network information for massive Web news that report social events using this framework, and designs following algorithms in order to research a network information extraction method for Web topic detection.

4.1. The algorithm of theme similarity measurement for network information extraction

The design idea of theme similarity measurement algorithm is as follows, according to TopKeyWordSet, this algorithm can extract Web news instances related to theme from NewsSet and UrlSet, and calculate importance of Web news instances related to theme from InitialUrlQueue. The input content of this algorithm is set of Web news and theme words and so on, the output content of this algorithm is Web news instance information related to theme, the construction process is as follows.

Under background of social event occurrence, according to TopKeyWordSet, this process can form theme vector. As shown in Formula 1, si represents social event, tkwsij.weightvalue represents weight of theme word tkwsij.wordvalue and constitutes the component values of the theme vector. This vector is not constant, when massive information for Web news is extracted, according to its keyword set, this process can conduct iterative processing for theme words and weight value, and continue to learn extraction results of massive information for Web news. Among this process, the value of the number of theme words is determined in subsequent experiments.

$$\begin{gathered}\{s_i,(tkwsi_1.\text{wordvalue},tkwsi_1.\text{weightvalue}),(tkwsi_2.\text{wordvalue},\\ tkwsi_2.\text{weightvalue}),\ldots,(tkwsi_j.\text{wordvalue},tkwsi_j.\text{weightvalue}),\\ \ldots,(tkwsi_n.\text{wordvalue},tkwsi_n.\text{weightvalue})\},\\ \text{The theme vector: }[tkwsi_1.\text{weightvalue},tkwsi_2.\text{weightvalue},\ldots,\\ tkwsi_j.\text{weightvalue},\ldots,tkwsi_n.\text{weightvalue}]\end{gathered} \tag{1}$$

Through using open source library NekoHtml, this process can extract source code of Web news instance page, extract contents of ¡title¿ element label, and attribution value of content in ¡meta name=”keywords” content=”*”/¿ element tag, form element vector, and calculate similarity between Web news instance and theme as shown in Fig. 2. VectorTheme represents theme vector, VectorElement represents element vector, Relativity(D,Theme) represents the calculated result, which is compared with similar threshold value determined in subsequent experiments. If it is greater than or equal to this value, then it is similar to theme, conversely, it is not similar to theme.

As shown in Fig. 3, this process can calculate similarity between words. d1 represents the shortest distance between two words in WordNet, d2 represents depth of two words belong to same category in WordNet.

This process designs regular expressions, in order to eliminate interference of noise information for extracting Web news content. This process locates location of distribution for data items in Web news content, and extracts data items from Web news instance content related to theme. If there are URL sets pointing to next link target in extracted content, then it can be enqueued to InitialUrlQueue. This process calculates importance of similarity with theme; according to importance, instances can be ranked from high to low in InitialUrlQueue queue, and it can be enqueued to WaitingUrlQueue. When queue is not empty, this process calculates similarity between instances and theme. The importance of instance may be lower than its parent layer URL in InitialUrlQueue queue, but information contained in it may not exist in the parent layer URL instance. Therefore, this process conducts priority processing for link instances that have higher importance based on importing genetic factor, which is represented with σ, in order to ensure integrity of

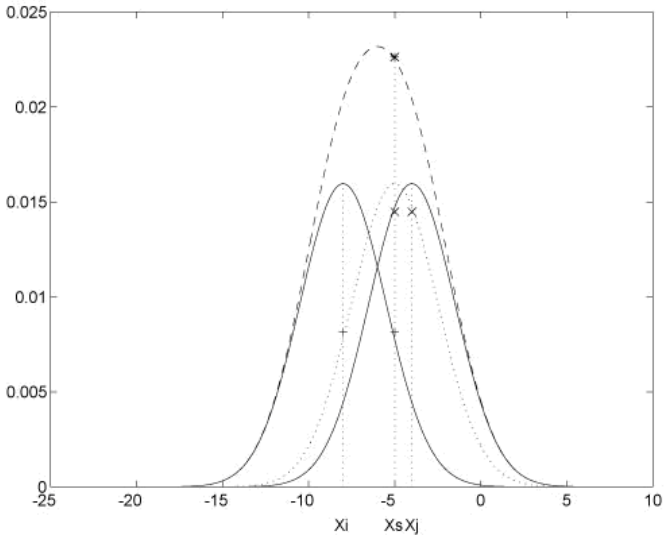


Fig. 2. The element vector and the computation process for Relativity(D,Theme)

information extraction, and consider how to do partial optimization as shown in Fig. 4. Among this process, the adjustment of factors is determined in subsequent experiments.

4.2. The algorithm of incremental instance extraction for Web topic detection

The design idea of incremental instance extraction algorithm is as follows, according to extracted Web news instance data items, this algorithm can analyze Pattern and BaseP-attern. According to TopKeyWordSet, this algorithm can conduct segmentation for Web news title and text content, and extract keywords, the result of keyword extracted can describe Web topic. Based on the imported self-adaption strategy, this algorithm can ad-just size of theme vectors and similar thresholds. The input content of this algorithm are extracted Web news instance data items and TopKeyWordSet, the output content of this algorithm are filtering mode and Web news instances content extracted under background of theme incrementally, the construction process is as follows.

This process can extract URL information from Web news instances using ”//” separa-tor as identification mark, and select its top-level nodes. Use ”/” separator as identification mark, this process can select its hierarchical nodes. The selected nodes are constructed into tree structure, and this process can analyze filtering mode Pattern among nodes for different levels. Grouping node tree structure one by one, according to filtering mode Pat-tern of previous level node for current node group, this process can analyze base filtering mode for regular hierarchical nodes as shown in Fig. 5.

This process can conduct segmentation for Web news title and text content, but in its result, there are some words that have nothing to do with topic detection. For example, the existence of words annotated as /p and /d and so on will contribute less to analysis of Web news texts, it will not only reduce efficiency and quality of Web news text analysis, and also reduce accuracy of topic detection. Based on occurrence of unexpected events, the

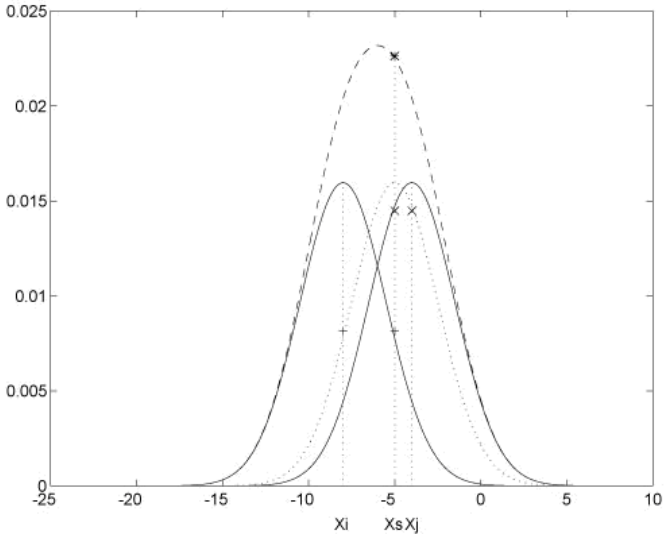


Fig. 3. The computation process for $\mathrm{Sim}(word_1, word_2)$

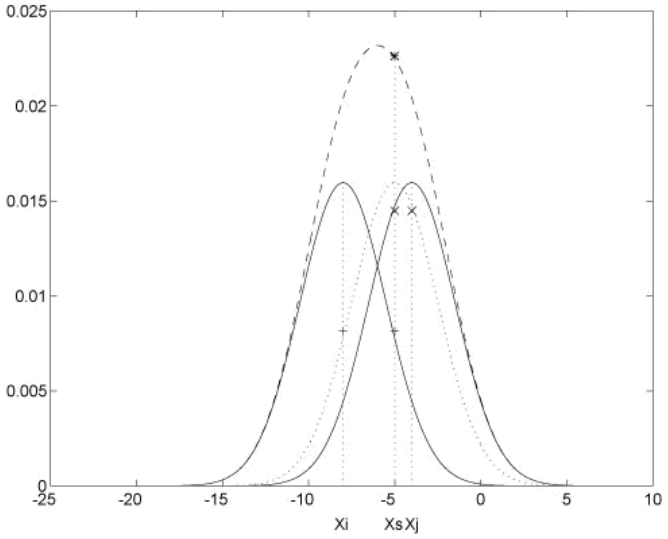


Fig. 4. The computation process for Relativity

Algorithm 1 Theme Similarity Measurement

Input: UrlSet, NewsSet, VectorTheme, Threshold, Parameters, InitialTime, T
Output: RTSet
WebNews ← (UrlSet, NewsSet)
for each $w[i]$ ($0 \leq i \leq w.\text{size}() - 1$) **do**
  sc ← Extract $w[i]$ target source code
  ec ← Extract sc element content
  VectorElement ← Generate element vector(ec, VectorTheme)
  Relativityparent ← Calculate $w[i]$ theme similarity(VectorElement, VectorTheme)
  if Relativityparent ≥ Threshold **then**
    rt ← Extract sc data items
    RTSet.addlist(rt)
    $urls$ ← Extract sc url
    EnInitialUrlQueue($urls$)
    if Time interval is T **then**
      $urlscobj$ ← Extract $urls$ target source code
      $urlecobj$ ← Extract $urlscobj$ element content
      VectorElementObj ← Generate element object vector($urlecobj$, VectorTheme)
      Relativity ← Calculate $urls$ theme similarity(VectorElementObj, VectorTheme)
      Call genetic factor strategy(Parameters)
      RankInitialUrlQueue(Relativity)
      DeInitialUrlQueue($urls$)
      EnWaitingUrlQueue($urls$)
      InitialTime ← Adjust InitialTime
    end if
  end if
end for
return RTSet

word segmentation process has some difficulties in detecting new words, so, it is impossible to accurately represent new words for emergencies in word segmentation result, these words are quite important in topic detection. The detection of these words will contribute more to Web news text analysis, it can not only improve efficiency and quality of Web news text analysis, and also improve accuracy of topic detection. Therefore, this process designs the corpus for filtering word, according to part of speech tagging and TopKeyWordSet in word segmentation result, and the corpus for filtering word, on the one hand, it can filter words that are not meaningful, on the other hand, it can detect new words with practical meaning, it will provide an effective segmentation result for Web news text analysis.

This process calculates weight of words in Web news instances, and extracts keywords as shown in Fig. 6. F(KeyWord,D) represents frequency of KeyWord occurrence in Web news instances, N represents total number of Web news instances involved in computing, n represents number of Web news instances containing KeyWord involved in computing for Web news group, Weight(KeyWord,T) represents weight in theme set for KeyWord, in order to consider importance of same word under different theme background.

On basis of traditional vector space model, this process imports self-adaption strategy, in order to consider feedback and guidance on theme in dynamically increasing Web

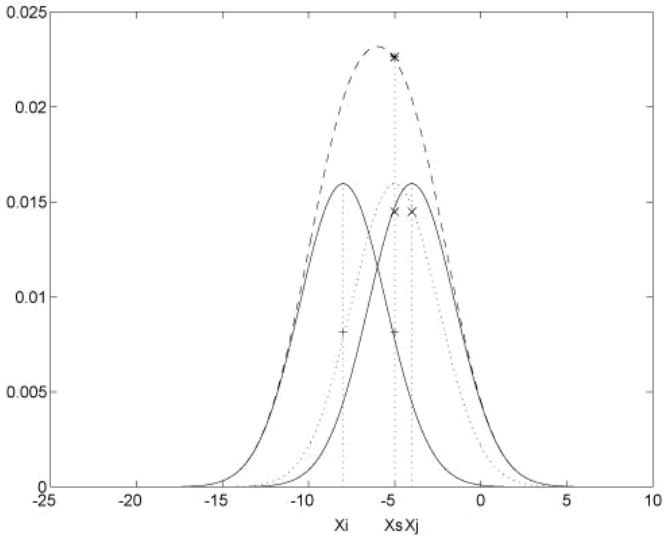


Fig. 5. URL network diagram structure is converted into tree structure

news instances. Within time interval, according to subsequent feedback information, this process automatically adjusts size of theme vectors and similar thresholds. The adjustment parameters are determined in subsequent experiments, if threshold increases, then it can improve accuracy of content extraction. However, when the calculated similarity of theme is generally low, if threshold is lowered, then scope of content extraction can be expanded. As shown in Fig. 7, Sumt1 represents the number of Web news instances extracted in t1 time, Sume represents the number of Web news instances extracted, it is expected to be within t time interval, Sumt2 represents the number of Web news instances extracted with t time interval, Sumt1/theme represents the number of Web news instances extracted that are similar to theme in t1 time.

Based on the method proposed in this paper, the effect of network information extraction for Web topic detection can be obtained, it can represent the extraction result for massive Web news in the context of social events. In the subsequent research, the analysis for semantic feature, feature evaluation and use behavior tracking on the extracted big data corpus can be conducted to further enhance the research value of incremental network information extraction for Web topic detection. To reflect the feasibility, verifiability and superiority for the method proposed in this section, the next section will complete the experimental analysis process.

5. The process of experimental analysis

This section will describe the experimental setup firstly, then analyze the evaluation metrics for the research method proposed in this paper, and then compare the performance of the experimental effect.

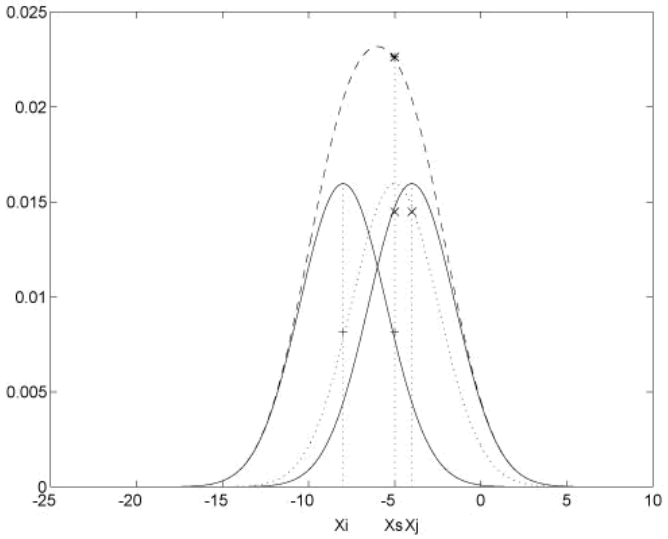


Fig. 6. The computation process for Weight(KeyWord,D)

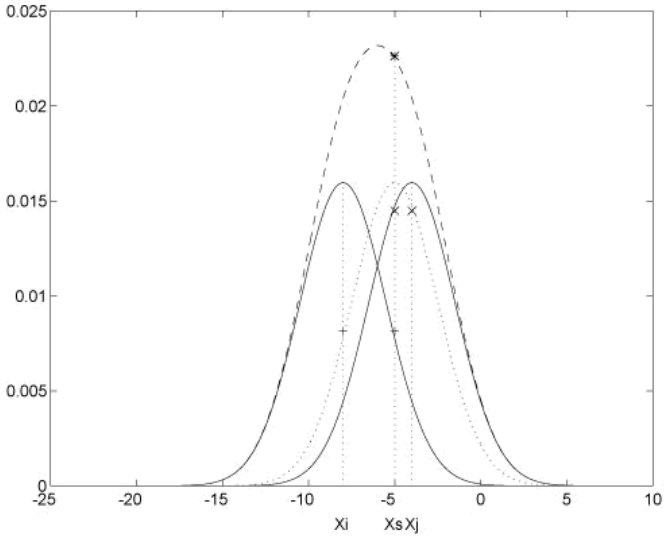


Fig. 7. The computation process for Threshold

Algorithm 2 Incremental Instance Extraction

Input: RTSet, Text corpus, Parameters, InitialTime, T
Output: Pattern, BasePattern, KeyWordSet
WebNews ← (RTSet)
for each $w[i]$ $(0 \leq i \leq w.\text{size}() - 1)$ **do**
  Pattern ← Recursive analysis for $w[i].url$ filter mode
  BasePattern ← Recursive analysis for $w[i].childurl$ filter basemode
  ws ← WordSegment($w[i].dataitems$)
  $kwobj$ ← ExtractKeyWords(ws, Text corpus)
  for each $kw[i]$ $(0 \leq i \leq kwobj.\text{size}() - 1)$ **do**
    KeyWordSet.addlist($kw[i]$)
  end for
  if Time interval is T **then**
    Call adaptive strategy(Parameters, KeyWordSet)
    InitialTime ← Adjust InitialTime
  end if
end for
Describe Web topic(KeyWordSet)

5.1. The experimental setup

Based on the design ideas and algorithms proposed in this paper, the hardware and software environments used in the experimental process are as follows. The processor is Intel 2.40GHz, the memory is 64GB, and the operating system is 64 bit Windows. The programming language is Java, mainly used for algorithm implementation. The network application research and development platform is MyEclipse, and the database management system is SQL Server, mainly used for storing and processing Web big data extracted [26, 27, 28]. The Web Project has been published on Big Data Analysis and Mining of My Teaching Classroom and My Practical Project for Tou Ge Practice Teaching Platform including program and data, due to dependency on the General Project of Science and Technology Plan of Beijing Municipal Education Commission, the Research Project on Graduate Education Science at Beijing Union University in 2025, and the Support Project of High-Level Teachers in Beijing Municipal Universities in the Period of 13th Five-Year Plan for this Web Project, therefore, it has been used in both graduate practical courses and undergraduate applied courses.

This paper uses the massive Web news generated by the German A320 aircraft crash event as the extraction target for network big data, these Web news were all published by authoritative websites. This event has gone through the process of beginning, development, and end, and the data is authentic, the experimental analysis process can verify the feasibility and effectiveness of the design ideas and algorithm design proposed in this paper.

5.2. The evaluation metrics

In the following experiments, the Precision evaluation index is used to measure not only the ratio of correctly extracted Web news instances to all extracted Web news instances that are similar to the theme, but also the ratio of correctly detected topics to all detected

topics, this ratio reflects the accuracy of extraction and detection. Based on the comprehensive consideration of Precision and Recall evaluation indicators, the F-Measure evaluation index is used to measure the overall extraction effect, this is a comprehensive performance of accuracy and comprehensiveness. Among them, the Recall evaluation index is used to measure the ratio of correctly extracted Web news instances to all Web news instances that should be extracted, this ratio reflects the comprehensiveness of extraction. In the evaluation process, the accurately extracted Web news instances can be annotated by automatically calculating the similarity between the keywords of the extracted Web news instance and the topic words.

In the following experiments, this paper also uses four other real data sets including the events of Shanghai Bund trample, Taiwan revival airliner falling river, Nepal 8.1 earthquake and Orient Star cruise overturn as the extraction target for network big data, in order to verify whether the proposed method in this paper is more universal and reliable for the evaluation index.

5.3. The performance comparison

Firstly, this section analyzes the accuracy of extracted Web news instances that are similar to the theme under different Web news information extraction methods as shown in Fig. 8 and Table 3, the accuracy represents whether the Web news instances extracted through two methods that are similar to the theme belong to the strongly or weakly associated annotation category with the theme. The red solid line represents the change in accuracy of Web news instances extracted under the theme matching method that are similar to the theme, from its trend, it can be seen that the extraction process is more open without detailed knowledge of the theme reported in Web news. With the continuous increase in the number of Web news, the accuracy has remained relatively stable, failing to break through 0.74. The blue solid line represents the change in accuracy of Web news instances extracted under the algorithm designed in this paper that are similar to the theme, from its trend, it can be seen that the information extraction filtering mode derived from continuously analyzing the extracted Web news instances has played a role in improving accuracy. Although its accuracy is similar to that of theme matching method when the number of Web news is small, as the number of Web news continues to increase, its accuracy also keeps climbing, reaching up to 0.83. So, this experiment demonstrates that the quality of Web news instances extracted by the algorithm in this paper is higher than that of theme matching method.

Next, this section analyzes the F-Measure values of extracted Web news instances that are similar to the theme under different threshold conditions to obtain the optimal threshold value as shown in Fig. 9 and Table 4, the F-Measure value represents the comprehensive quality of Web news instances extracted that are similar to the theme through the algorithm proposed in this paper, while continuously adjusting the threshold value. The red solid line represents the variation of F-Measure values under the theme background when the threshold takes different values, from its trend, it can be seen that when the threshold value is low, the comprehensive quality of Web news instances extracted that are similar to the theme is not high with the F-Measure value of approximately 0.6. As the threshold value increases, the comprehensive quality of Web news instances extracted that are similar to the theme also increases, and the F-Measure value also increases accordingly. When the threshold value is adjusted to about 0.65, the comprehensive quality of Web

news instances extracted that are similar to the theme reaches its highest level. When the threshold value is further increased, some Web news instances weakly associated with the extraction theme are not extracted, and the F-Measure value also decreases accordingly. The blue solid line represents the variation of F-Measure values under the method proposed in this paper when different threshold values are taken, from its trend, it can be seen that when the threshold value is low, the comprehensive quality of Web news instances extracted that are similar to the theme is high with the F-Measure value of approximately 0.75. As the threshold value increases, the comprehensive quality of Web news instances extracted that are similar to the theme remains stable, and the F-Measure value remains around 0.8. When the threshold value is adjusted to about 0.65, the comprehensive quality of Web news instances extracted that are similar to the theme reaches its highest level. When the threshold value is further increased, some Web news instances that are strongly or weakly associated with the extraction theme are not extracted, and the F-Measure value also decreases accordingly with a greater decrease than in the background of the theme. Before adjusting the threshold value to 0.78, the F-Measure value under the method proposed in this paper is higher than that under the theme background. In the context of the theme, when the threshold value is 0.65, the corresponding F-Measure value reaches its maximum, which is close to 0.75. Under the method proposed in this paper, when the threshold value is 0.65, the corresponding F-Measure value also reaches its maximum, which is close to 0.83. So, this experiment indicates that the optimal threshold value can be 0.65.

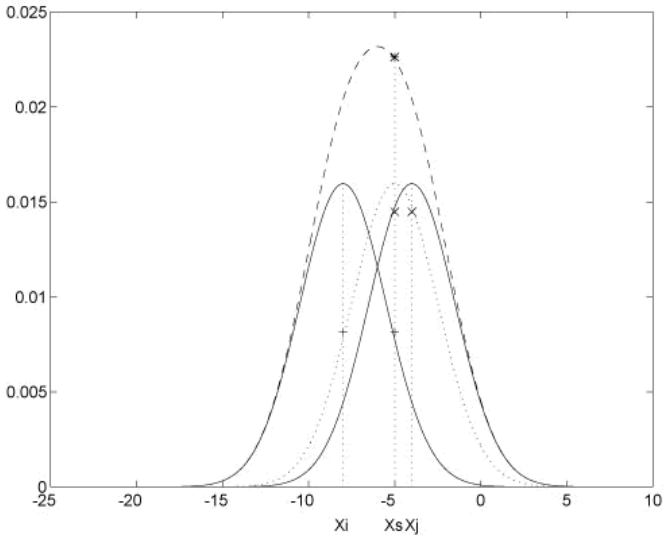


Fig. 8. The quality of Web news instances extracted that are similar to the theme under different methods

Next, this section analyzes the impact of Web news quantity extracted and threshold on the topic detection quality as shown in Fig. 10, the accuracy represents the quality of

Table 3. The comparison of accuracy rate for Web news instances extraction quality that is similar to the theme under different methods

The number of Web news (Unit: K) / The research method	20	40	60	80	100	120	140	160	180	200
Theme matching method	0.652	0.67	0.736	0.726	0.659	0.67	0.738	0.684	0.71	0.696
This paper's algorithm	0.65	0.664	0.694	0.725	0.754	0.764	0.794	0.804	0.822	0.833

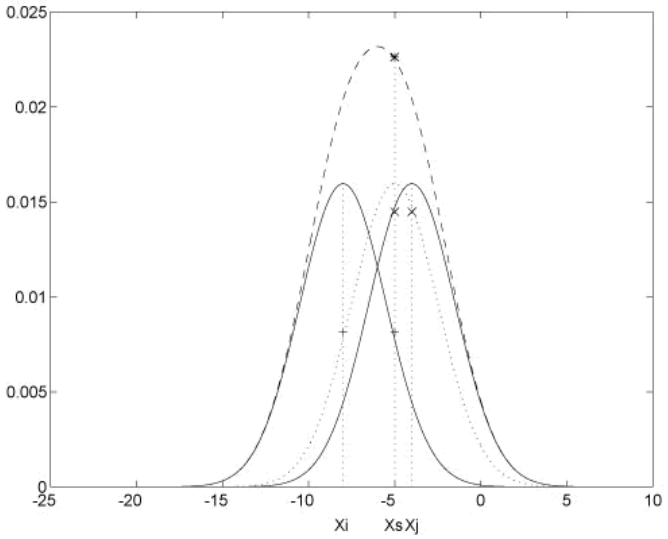


Fig. 9. The trend of F-Measure value variation with threshold value

Table 4. The comparison of F-Measure value for Web news instances extraction quality that is similar to the theme under different threshold conditions

The threshold value / The research method	0.1–0.15	0.2–0.25	0.3–0.35	0.4–0.45	0.5–0.55	0.6–0.65	0.65–0.7	0.75–0.8	0.85–0.9
The theme background	0.6–0.61 ↑	0.62–0.63 ↑	0.64–0.641 ↑	0.651–0.661 ↑	0.671–0.681 ↑	0.691–0.75 ↑	0.75–0.708 ↓	0.635–0.628 ↓	0.61–0.569 ↓
The incremental method	0.75–0.801 ↑	0.801–0.817 →↑	0.753–0.772 ↓↑	0.763–0.774 ↓↑	0.804–0.797 ↑↓	0.789–0.83 ↓↑	0.83–0.767 ↓	0.683–0.611 ↓	0.537–0.46 ↓

topic detection by adjusting the X-axis threshold and continuously increasing the number of Web news instances extracted on the Y-axis. When the threshold value is constant, it can be seen from the graph that as the number of Web news instances extracted continues to increase, the accuracy shows a trend of first increasing and then decreasing. The reason is that when the number of Web news instances extracted is small, it is not yet possible to fully understand the content reported by Web news. When there are a large number of Web news instances extracted, some Web news instances extracted that do not support topics are also analyzed, resulting in a decrease in accuracy. When the number of Web news instances extracted is constant, it can be seen from the graph that as the threshold value increases, the accuracy shows an increasing trend. The reason is that when the threshold value is low, most Web news instances extracted that do not support topics are extracted. When the threshold value is high, only a small number of Web news instances extracted that do not support topics are extracted. This experiment shows that when the number of Web news instances extracted is 100K and the threshold is 0.9, the quality of topic detection can reach the highest value, which is close to 0.82.

Next, this section analyzes the accuracy of topic detection under different numbers of theme keywords to obtain the optimal value for theme keywords quantity as shown in Fig. 11 and Table 5, the accuracy represents the quality of topic detection through the algorithm proposed in this paper with setting different numbers of theme keywords. The red solid line represents the change in accuracy of the theme background when the number of theme keywords is set differently, from its trend, it can be seen that when the number of theme keywords is set 3, the quality of topic detection tends to stabilize at approximately 0.682 to 0.685. The reason is that when the number of theme keywords is too small or too large, some Web news instances extracted that do not support topics are extracted. The blue solid line represents the change in accuracy of the incremental method when the number of theme keywords is set differently, from its trend, it can be seen that when the number of theme keywords is set 3, the quality of topic detection tends to stabilize at approximately 0.701 to 0.704 for the same reason as the red solid line trend. Overall, the accuracy under the incremental method is higher than that under the theme background, this experiment shows that the optimal value for the number of theme keywords can be set 4, so that under the algorithm proposed in this paper, the quality of topic detection is locally stable to the maximum value.

Next, this section analyzes the impact of the adjustment parameters in the algorithm on the topic detection quality to obtain the optimal adjustment range for these parameters as shown in Fig. 12, the accuracy represents the quality of topic detection when adjusting parameters for theme similarity measurement and incremental instance extraction algorithms. The red dashed line represents the variation in accuracy when the Alpha parameter takes different values for Formula 6, from its trend, it can be seen that when the Alpha value is adjusted between 1.15 and 1.3, the quality of topic detection is high and stable with an accuracy rate of approximately 0.76. The blue dashed line represents the change in accuracy when the Mu parameter takes different values for Formula 6, from its trend, it can be seen that when the Mu value is adjusted between 0.75 and 0.9, the quality of topic detection is high and stable with an accuracy rate of approximately 0.70. The green dashed line represents the change in accuracy when the Beta parameter takes different values for Formula 6, from its trend, it can be seen that when the Beta value is adjusted between 1.2 and 1.45, the quality of topic detection is high and stable with an accuracy

rate of approximately 0.77. The yellow dashed line represents the variation in accuracy when the Gamma parameter takes different values for Formula 6, from its trend, it can be seen that when the Gamma value is adjusted between 0.8 and 0.95, the quality of topic detection is high and stable with an accuracy rate of approximately 0.76. The pink dashed line represents the variation in accuracy when Sigma parameters take different values for Formula 4, from its trend, it can be seen that when the Sigma value is adjusted between 0.8 and 0.95, the quality of topic detection is high and stable with an accuracy rate of approximately 0.81. Overall, the adjustment of various parameters can locally stabilize the quality of topic detection to its maximum value, and determine the optimal adjustment range of each parameter.

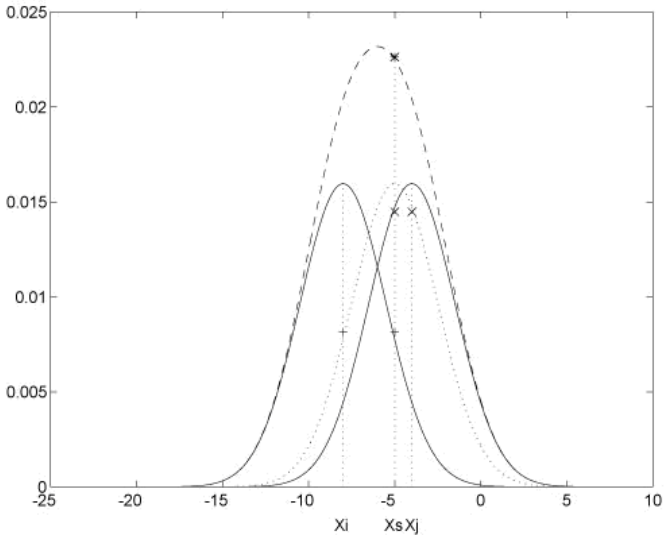


Fig. 10. The trend of accuracy changing with the number of Web news extracted and threshold value

Table 5. The comparison of accuracy rate for the topic detection quality under different numbers of theme keywords

The numbers of theme keywords / The research method	1	2	3	4	5	6	7	8
The theme background	0.625	0.684	0.682	0.685	0.682	0.685	0.662	0.662
The incremental method	0.708	0.703	0.701	0.704	0.701	0.704	0.718	0.719

Finally, this section analyzes the quality of topic detection on different data sets based on the extraction result for Web news as shown in Fig. 13 and Table 6, from its trend, it can be seen that under the method proposed in this paper, there is almost no difference in the quality of topic detection for the five events at the start, development, and end stages, this indicates that the quality of topic detection using the method proposed in this paper

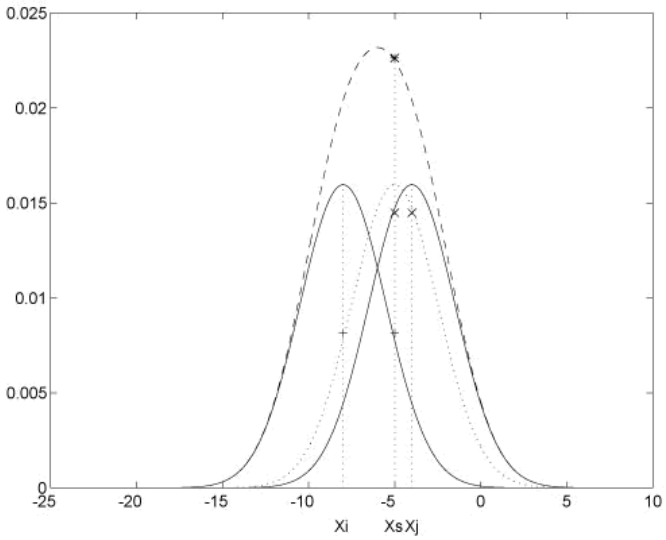


Fig. 11. The trend of accuracy changing with the number of theme keywords

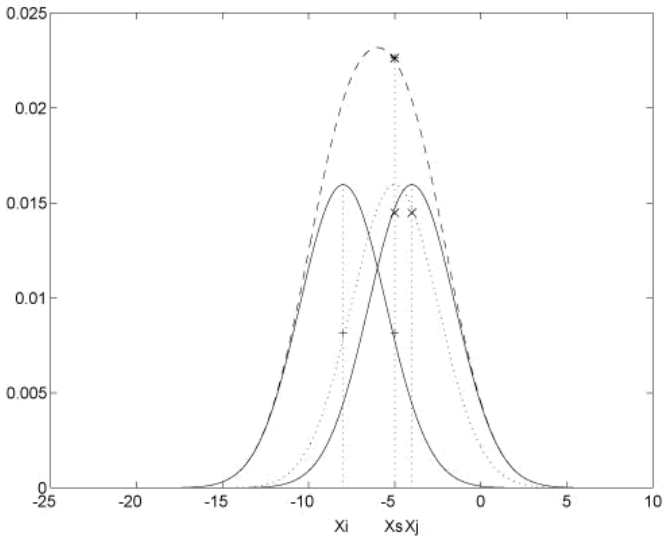


Fig. 12. The trend of accuracy changing with the adjustment for various parameters in the algorithm

is relatively stable under different events. However, as the number of Web news extracted continues to increase, it still has a certain impact on the quality of topic detection at different stages of the event. This is because, with the increase in the number of Web news extracted, the extraction result under the theme is constantly expanding, and the support for topic detection is also improving.

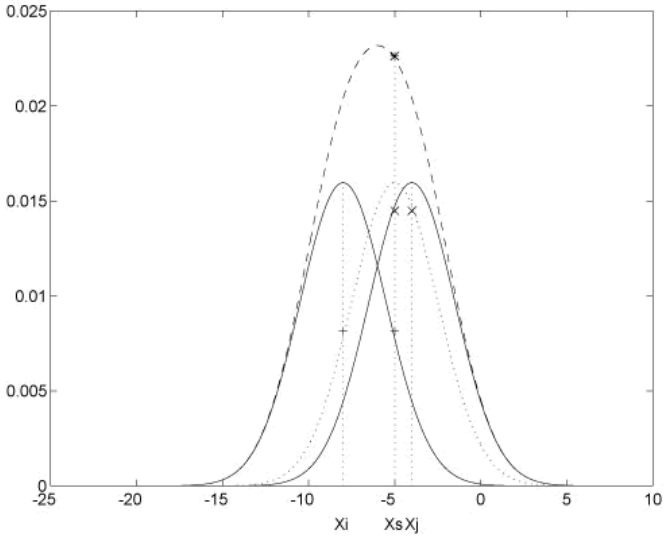


Fig. 13. The quality of topic detection on different data sets based on the extraction result

Table 6. The comparison of accuracy rate for the topic detection quality on different data sets based on the extraction result

The data sets / The evolution stage	Shanghai Bund trample	Taiwan revival airliner falling river	German A320 airliner crash	Nepal 8.1 earthquake	Orient Star cruise overturn
Start stage (About 50K)	0.701	0.698	0.706	0.692	0.704
Development stage (About 140K)	0.727	0.718	0.727	0.715	0.728
End stage (About 200K)	0.753	0.748	0.753	0.743	0.755

6. Conclusion

The research on the incremental network information extraction technology for Web topic detection, taking the network big data of Web news as the research object, has been completed, the results of this design and implementation are more valuable to scholars in related research fields. In the process of technology research, this paper proposes the theme element extraction module, the theme similarity calculation module, the instance data items extraction module, the link importance calculation module, the keywords extraction module, and the incremental instance extraction module, and proposes the algorithm

of the theme similarity measurement for incremental network information extraction and the incremental instance extraction for Web topic detection, in order to address the shortcomings in the current research status. The experimental analysis process shows that the method proposed in this paper is feasible, verifiable, and superior. It has played an important role in reconfiguring numerical intelligence warehouses for detecting Web topic, and inferring the Web hierarchical big data propagation path.

In subsequent research, the optimization of incremental network information extraction algorithm for Web topic detection can be continued, and the optimal range of parameters in the algorithm can be refined through experiments to handle fuzzy or conflicting semantic information, enhance the comprehensiveness, accuracy, and robustness of incremental network information extraction for Web topic detection. The research results can be applied to real-time Web news monitoring and multilingual Web topic detection processes, and further improve the processing efficiency for big data applications.

Acknowledgement. This paper is supported by General Project of Science and Technology Plan of Beijing Municipal Education Commission under Grant Nos. KM202011417011, Research Project on Graduate Education Science at Beijing Union University in 2025 under Grant Nos. YK202502, Support Project of High-Level Teachers in Beijing Municipal Universities in the Period of 13th Five-Year Plan under Grant Nos. CIT&TCD201704072.

References

1. Li, P., Zhang, L.: Application of big data technology in enterprise information security management. Scientific Reports 15(1), 1–5 (2025)
2. Arunkumar, M., Rajkumar, K., Jeyaseelan, W., Natraj, N.: Data mining, machine learning, and statistical modeling for predictive analytics with behavioral big data. Tehnicki Vjesnik - Technical Gazette 32(1), 72–74 (2025)
3. de Miguel, A., Sarasa-Cabezuelo, A.: A global approach to artificial intelligence. IEEE Access 13, 76946–76950 (2025)
4. Kaushik, M., Sharma, R., Koiva, P., Fister, I.J., Draheim, D.: An exhaustive multi-aspect analysis of swarm intelligence algorithms in numerical association rule mining. IEEE Access 12, 138985–138989 (2024)
5. Tang, J., Yan, Y., Bao, J., Huang, B.: Big data-driven control of nonlinear processes through dynamic latent variables using an autoencoder. IEEE Transactions on Cybernetics 55(5), 2411–2415 (2025)
6. Song, S., Pan, L., Liu, S.: A q-learning based auto-scaling approach for provisioning big data analysis services in cloud environments. Future Generation Computer Systems 154, 140–144 (2024)
7. Wang, S.: Research on the digital marketing strategies in the e-commerce logistics service mode under the influence of big data. Computer-Aided Design and Applications 21(S4), 39–43 (2024)
8. Wang, H., Zhang, S.: Research on the application of improved bert-dpcnn model in chinese news text classification. Concurrency and Computation: Practice and Experience 37(3), 1–3 (2025)
9. Tredinnick, L.: The intricate web: Network and rhizome metaphors in hypertext and the web and the epistemic challenge of fake news. Journal of Documentation 79(6), 1485–1489 (2023)
10. Li, T., Yu, J., Zhang, H.: Web of things based social media fake news classification with feature extraction using pre-trained convoluted recurrent network with deep fuzzy learning. Theoretical Computer Science 931, 65–69 (2022)

11. Li, X., Gao, N., Wang, Y.: Identifying disruptive technology using saox semantic analysis and web news data mining: A perspective of technology convergence. IEEE Transactions on Engineering Management 72, 2116–2120 (2025)
12. Mallick, P., Mishra, S., Chae, G.: Digital media news categorization using bernoulli document model for web content convergence. Personal and Ubiquitous Computing 27(3), 1087–1091 (2023)
13. Dritsas, E., Trigka, M.: Database systems in the big data era: Architectures, performance, and open challenges. IEEE Access 13, 95068–95072 (2025)
14. Angskun, T., Sritha, K., Srithong, A., Khopolklang, N., Kamollimsakul, S., Phithak, T., Angskun, J.: Using big data to assess an affective domain for distance education. Future Generation Computer Systems 160, 131–134 (2024)
15. Xi, Q., Jiang, P.: Design of news sentiment classification and recommendation system based on multi-model fusion and text similarity. International Journal of Cognitive Computing in Engineering 6, 44–47 (2025)
16. Wu, C., Wu, F., Huang, Y., Xie, X.: Personalized news recommendation: Methods and challenges. ACM Transactions on Information Systems 41(1), 1–4 (2023)
17. Xu, H., Peng, Q., Liu, H., Sun, Y., Wang, W.: Group-based personalized news recommendation with long and short term fine-grained matching. ACM Transactions on Information Systems 42(1), 1–6 (2023)
18. Guo, P., Song, Y., Wang, B., Liu, J., Zhang, Q.: Plbr: A semi-supervised document key information extraction via pseudo-labeling bias rectification. IEEE Transactions on Knowledge and Data Engineering 36(12), 9025–9036 (2024)
19. Chang, C., Tang, F., Yang, P., Zhang, J., Huang, J., Li, J., Li, Z.: Multi-view knowledge representation learning for personalized news recommendation. Scientific Reports 15(1), 1–13 (2025)
20. Li, Y., Yan, H., Zhang, Y., Wang, X.: Joint entity and relation extraction combined with multimodule feature information enhancement. Complex & Intelligent Systems 10(5), 6633–6645 (2024)
21. Serreli, L., Marche, C., Nitti, M.: Reducing data volume in news topic classification: Deep learning framework and dataset. IEEE Open Journal of the Computer Society 6, 152–163 (2025)
22. Zheng, Y., Tuan, L.: A novel, cognitively inspired, unified graph-based multi-task framework for information extraction. Cognitive Computation 15(6), 2004–2013 (2023)
23. Zhang, S., Tan, F., Peng, H.: Sample size determination for multidimensional parameters and the a-optimal subsampling in a big data linear regression model. Journal of Statistical Computation and Simulation 95(3), 628–632 (2025)
24. Matthews, S.: Review of statistical learning for big, dependent data. Journal of Official Statistics 40(4), 849–851 (2024)
25. Liu, H., Lu, F., Shi, B., Hu, Y., Li, M.: Big data and supply chain resilience: Role of decision-making technology. Management Decision 61(9), 2792–2796 (2023)
26. Ou, T., Chen, C., Tsai, W.: Establishing a dynamic recommendation system for e-commerce by integrating online reviews, product feature expansion, and deep learning. Applied Artificial Intelligence 39(1), 1–5 (2025)
27. Rui, G., Li, M.: Utilizing internet big data and machine learning for product demand forecasting and analysis of its economic benefits. Tehnicki Vjesnik - Technical Gazette 31(4), 1385–1388 (2024)
28. Zhu, Z., Sun, Y.: Personalized information push system for education management based on big data mode and collaborative filtering algorithm. Soft Computing 27(14), 10057–10060 (2023)

Mo Chen received Ph.D from School of Information, Renmin University of China in Computer Application Technology speciality, he is an associate professor of E-commerce

Department at Beijing Union University Business College. He engages in Data Structure, Database Principles and Applications, Data Acquisition and Preprocessing, Data Mining and Machine Learning, Business Big Data Analysis and Decision-Making, Big Data Technology and Application, Blockchain Application Technology and other courses teaching and researching work. His research interests are Big Data Analysis and Mining and so on, he publishes papers in the core journals, presides over research projects of the science and teaching.

Received: October 30, 2025; Accepted: March 20, 2026.

Computer Science and Information Systems 23(2):801–826 https://doi.org/10.2298/CSIS251130024L

Development and Validation of a Few-Shot Rapid Screening Model for Gastrointestinal Cancers Using AGI Large Vision Models

Lijue Liu[1], Fangjie Yin[1], Genjian Yang[2], Qi Li[3], Siya Li[4], Teng Pan[5], Ting Liu[6], Jin Tang[1,7], Ruijie Ming[8], Yu Song[9], Xue Feng[10], Dan Wang[11], Xingang Zhou[6], Wenbai Chen[2], and Jinhai Deng[11,12]

[1] School of Automation, Central South University
410083 Changsha, China
{ljliu, tjin}@csu.edu.cn, yinfangjie2023@126.com
[2] School of Automation, Beijing Information Technology Science and University
102206 Beijing, China
457706420@qq.com, chenwb@bistu.edu.cn (corresponding author)
[3] Department of Pathology, Beijing Integrated Traditional Chinese and Western Medicine Hospital
100039 Beijing, China
15201232918@163.com
[4] CAS Blue Bay Cloud Technology (Guangdong) Co., Ltd.
518001 Guangzhou, China
1004297233@qq.com
[5] Longgang District Maternity & Child Healthcare Hospital of Shenzhen City, Longgang Maternity and Child Institute of Shantou University Medical College
518172 Shenzhen, China
2570758402@qq.com
[6] Department of Pathology, Beijing Ditan Hospital, Capital Medical University
100015 Beijing, China
liuting1981_2005@126.com, zhouxg1980@126.com (corresponding author)
[7] Xiangjiang Laboratory
410205 Changsha, China
tjin@csu.edu.cn
[8] Department of Oncology, Chongqing University Three Gorges Hospital
404010 Chongqing, China
ming_ruijie@cqu.edu.cn
[9] Department of Otolaryngology, Head & Neck Surgery, Peking University First Hospital
100034 Beijing, China
syandf@163.com
[10] Department of Respiratory and Critical Care Medicine, Tianjin Chest Hospital
300222 Tianjin, China
fengxuenku@163.com
[11] Richard Dimbleby Laboratory of Cancer Research, Randall Division and Division of Cancer and Pharmaceutical Sciences, King's College London
SE1 1UL London, UK
dan.7.wang@kcl.ac.uk, jinhaideng_kcl@163.com
[12] Guangzhou Baiyunshan Pharmaceutical Holding Co., Ltd. Baiyunshan Pharmaceutical General Factory/Guangdong Province Key Laboratory for Core Technology of Chemical Raw Materials and Pharmaceutical Formulations
510515 Guangzhou, China
jinhaideng_kcl@163.com (corresponding author)

Abstract. Existing deep learning models in digital pathology typically require extensive labeled data and show limited generalization across organs. In contrast, large vision models exhibit effective feature extraction capabilities, enabling pathological image analysis for gastrointestinal cancer with relatively small sample sizes. In this study, we developed a screening framework leveraging a large vision model for coarse-grained classification of gastric and colorectal tissues. The model was evaluated on multicenter cohorts and under limited-data conditions. Using labeled tiles from only 76 whole-slide images, the model achieved class-averaged sensitivity and precision of 0.9816 and 0.9808 on the internal test set, and 0.9161 and 0.9179 on the external test set. When trained with only 200 tiles per class from 20 whole-slide images, the model maintained comparable performance, achieving sensitivity and precision of 0.9548 and 0.9518. These findings suggest that the model has reliable performance across multicenter cohorts and potential applicability in clinical pathology workflows.

Keywords: Deep Learning, Gastrointestinal Cancers, Histopathology, Unified Screening.

1. Introduction

Gastrointestinal cancers, primarily including esophageal, gastric, and colorectal malignancies, are among the most prevalent human cancers [53]. Currently, histopathological analysis remains the gold standard for diagnosis [30], providing reliable tumor typing and informing treatment decisions through the examination of tissue or cell morphology. However, traditional pathological diagnosis requires pathologists to carefully examine entire slides by pathologists, which is labor-intensive, time-consuming, and susceptible to inter-observer variability [55, 9]. These challenges, combined with the global shortage of pathology specialists [47], highlight the growing need for developing automated computational approaches to assist pathologists in histopathological diagnosis.

Convolutional neural networks (CNNs) have fueled the explosive interest in applying deep learning to histopathology, owing to their ability to learn features directly from raw data [40]. For instance, Wang et al. [48] developed an Inception-V3-based method, achieving an area under the receiver operating characteristic curve (AUC) of 0.9880 in distinguishing colorectal cancer from normal tissue. Similarly, Song et al. [39] proposed a clinical pathology diagnostic system that reached nearly 100% sensitivity and an average specificity of 80.6% in a real-world cohort. In addition, Fu et al. [14] introduced the StoHisNet, a multiscale model that attained over 94% accuracy in classifying gastric pathological images, including normal tissue and adenocarcinoma subtypes.

Nevertheless, these studies largely focused on single-organ analysis. In practical clinical settings, however, pathologists frequently encounter diagnostic tasks involving multiple organs. Images from various organs show variation in staining, tissue architecture, resolution, and imaging modality[45]. These differences create a domain shift that can cause a classifier trained on one organ to perform poorly on another despite the same underlying malignant morphology [7].

Despite this, several studies have explored multi-organ diagnostic frameworks for gastrointestinal pathology. Iizuka et al. [20] utilized the same network architecture to classify adenocarcinoma, adenoma, and non-tumor tissues from gastric and intestinal pathological images, demonstrating the potential of CNNs for unified, coarse-grained classification

across the gastrointestinal tract. Likewise, Oh et al. [32] developed a two-stage gastric cancer model that generalized well to intestinal datasets, highlighting the feasibility of joint classification. Although these approaches demonstrate the possibility of cross-domain generalization, they, along with other CNN methods, require substantial annotated data for training, ranging from 85 to 7164 whole slide images (WSIs) [48, 39, 14, 20, 32, 52, 44, 1, 51, 18, 42, 19, 43, 23, 22, 15, 2]. In practice, obtaining sufficiently annotated pathological data remains costly and challenging, making model training under low-resource conditions an important problem.

Concurrently, the rise of large vision models (also known as visual foundation models) trained with self-supervision on large-scale datasets has emerged as a promising trend in pathological image analysis, enabling broad adaptability to diverse downstream tasks [11]. Previous studies have used datasets of 32,000 to 3,100,000 WSIs to successfully train pathology large vision models with parameter scales ranging from 28 million to 11 billion [54, 6, 46, 37, 50, 56]. Among them, Chen et al. [6] developed the UNI model, which surpassed state-of-the-art (SOTA) methods in multiple single-organ and tissue subtype classifications and achieved an AUC of 0.9750 across 32 cancer types, while Wang et al. [50]'s CHIEF model demonstrated strong performance in identifying 11 cancer types, with AUCs ranging from 0.9098 to 0.9943.

Most large vision models such as UNI[6], Virchow2[56], H-optimus-0[37] employ a self-supervised learning approach called DINOv2[33] in the pretraining stage, which has been shown to yield strong, off-the-shelf representations for downstream tasks without the need for further fine-tuning with labeled data. When downstream models build upon these large vision models, substantially less data and computational resources are required[31], thereby reducing sample complexity. This characteristic closely relates to the paradigm of few-shot or low-resource learning. Few-shot learning is a specialized branch of deep learning algorithms that addresses this challenge by enabling models to learn new concepts from a few labeled examples, mimicking the human ability to generalize from limited experience[13]. Several recent large vision models [12, 6, 25] have also evaluated their performance under few-shot learning settings. In clinical practice, many histopathological tasks suffer from extremely limited annotated samples. By leveraging pre-trained models' feature representation capacity, few-shot learning can largely reduce intra-class variation, enabling models to focus on more discriminative morphological patterns[38].

1.1. Main Algorithmic Contributions

Motivated by these observations, this study leverages prior knowledge from a pathology large vision model to develop a high-performance unified screening system for both gastric and colorectal cancers, using only 76 training WSIs. It aims to reduce the clinical diagnostic workload while improving screening efficiency. The main contributions of this study are as follows:

1) Unlike existing works that focus on single-organ classification, this framework can achieve reliable performance in unified screening for both gastric and colorectal cancers even under low-resource settings.

2) This framework incorporates a gated recurrent unit (GRU) module to refine the tile-level representations extracted from pretrained pathology foundation models for enhancing the discriminative ability.

3) To address the challenge of difficult classification, this framework applies an adaptive weighted loss based on recall and the average loss over historical epochs.

2. Methods

2.1. Data Collection

In this study, a total of 149 hematoxylin and eosin (H&E)-stained slides were collected from two hospitals: slides collected from Beijing Ditan Hospital Capital Medical University (BDHCMU) were used for model development, and an independent cohort from Beijing Integrated Traditional Chinese and Western Medicine Hospital (BITCWMH) was used as an external test set to evaluate generalization across institutions. To ensure data quality, five slides with blurred regions or marker interference were excluded. The final internal cohort comprised 96 slides covering various pathological stages, including 42 gastric and 54 colorectal samples. It was randomly divided into internal training and internal testing cohorts in an approximate ratio of 8:2 prior to tile extraction. Consequently, all corresponding tiles derived from a given WSI were assigned exclusively to either the training or testing cohort. The external testing cohort consisted of 48 slides, including 15 gastric and 33 colorectal slides, also representing a range of pathological conditions. The demographic characteristics of the patients in these two cohorts are shown in Table 1. Notably, only 76 slides were used for model training in this study, which is considerably fewer than those reported in most previous studies. In the relevant literature, we investigated [48, 39, 14, 20, 32, 52, 44, 1, 51, 18, 42, 19, 43, 23, 22, 15, 2], the number of WSIs used for training in previous studies ranged from 85 to 7164. To further evaluate the generalizability of our model, an independent validation cohort comprising 100 H&E-stained slides of gastric and colorectal tissues was obtained from The Cancer Genome Atlas (TCGA) public dataset. This cohort included 50 malignant and 50 non-malignant slides. The overall cohort construction process is illustrated in Fig. 1.

Table 1. The Demographic Characteristics of the Patients

Variables		Internal Cohort	External Cohort
Sex	Female	27.40%	52.50%
	Male	72.60%	47.50%
Age	Median	56	60
	Range	30-83	31-88
	Mean $\pm$ SD	53.40 $\pm$ 11.58	60.53 $\pm$ 13.86

2.2. Data Annotation and Preprocessing

Pathologists used ASAP1.9 to annotate 144 slides at the region of interest (ROI) level and categorized them into three subclasses. Two pathologists independently labeled and diagnosed the ROIs, and a third experienced pathologist confirmed the final annotations to ensure label quality and consistency. Some examples of pathologists' labeling are shown

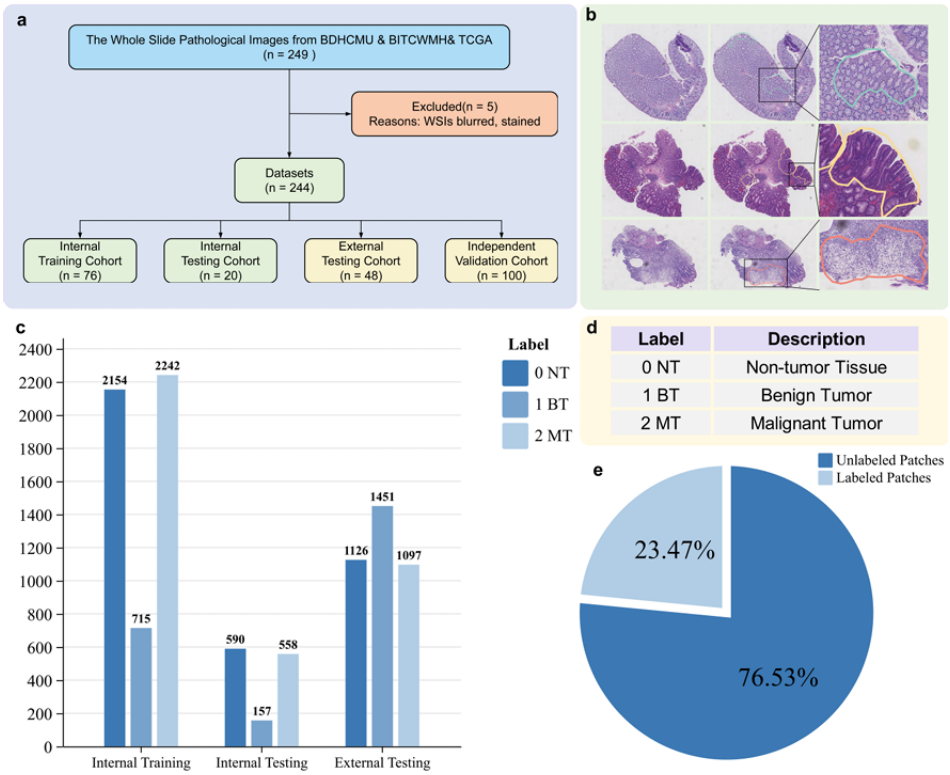


Fig. 1. Data cohort construction and its description. (a) The process of the cohort construction. (b) Some examples of pathologists' labeling. (c) The number of tiles in the internal and external cohorts. (d) Different levels of labeling and their descriptions. (e) The percentage distribution of annotated tiles.

in Fig. 1b. For the internal dataset, pathologists did not label all areas of the slide but used a selective labeling strategy. The 144 slides from BDHCMU & BITCWMH can be divided into 42,996 tiles, and the number of tiles divided after annotation is 10,090, accounting for approximately 23.47%. The proportion of annotated tiles is shown in Fig. 1e, while the total number of tiles in both the internal and external cohorts is shown in Fig. 1c. This strategy effectively controls the amount of data at the tile level while ensuring that the core information of the organization is retained, and the amount of data at the tile level is also as small as possible, thus fitting the needs of this study for limited data scenarios and providing accurate and efficient training samples for exploring the performance of the model under limited data conditions.

The different label levels and their corresponding descriptions are presented in Fig. 1d. Non-tumor tissues (Level 0, NT) include normal gastric mucosa, normal intestinal mucosa, entericized gastric mucosa, hyperplastic polyps, inflammation, fundic gland polyps, and mild atrophic glands. Benign tumors (Level 1, BT) include low-grade villous adenomas, low-grade tubular adenomas, and low-grade serrated adenomas. Malignant tumors (Level 2, MT) include high-grade tubular adenomas, signet-ring cell carcinomas, poorly differentiated adenocarcinomas, moderately differentiated adenocarcinomas, and highly differentiated adenocarcinomas. For the external dataset, pathologists assigned diagnostic labels at the WSI level to support the evaluation of the model's diagnostic performance across entire slides.

In this study, all slides were processed at a magnification of $10\times$ with a resolution of 0.8299 μm / pixel. The gastrointestinal dataset consists of ROIs annotated by pathologists. For each WSI, the annotated ROIs were independently tiled, and all resulting tiles from that WSI were used for training according to their class labels. Tiles were then fed into the model in batches with random sampling. Since the size of each WSI and the number and type of ROIs they contains vary, the number of tiles sampled from each WSI also varies, with a mean of 82 tiles per WSI (range: 5 – 1731), as shown in Table 2.

To extract the foreground from each slide and eliminate large internal cavities within the ROI, a segmentation threshold of 200 was empirically determined based on the actual characteristics of the scanned images. Subsequently, the extracted ROI is segmented into tiles of 224 $\times$ 224 pixels and saved for easy retrieval at a later stage based on the coordinate approach.

In order to achieve a relatively balanced number of tiles extracted for each type of label in the training dataset, data augmentation techniques including random rotation, horizontal flipping, and vertical flipping were applied.

2.3. Unified Screening System for Gastrointestinal Cancers

A unified screening system for gastrointestinal cancers called VGA (Virchow2-based GRU with Adaptive-weighted Loss) was developed to differentiate between non-tumor tissue, benign tumor tissue, and malignant tumor tissue. The overall workflow of the proposed screening system in this study is shown in Fig. 2.

First, since WSIs may contain up to tens of billions of pixels, feeding them directly into a neural network will be a huge challenge in terms of computational resources and storage overhead [21]. Resizing the entire image to a lower resolution would lead to the loss of cellular-level information, which will lead to a significant reduction in recognition accuracy. Therefore, it is common for researchers to adopt the strategy of dividing the WSI

into smaller tiles in order to adapt the processing capabilities of deep learning models [4, 27]. Drawing on this idea, this study analyzed WSIs at the tile level, with each tile labeled according to the pathological diagnosis of its ROI.

On the one hand, vision transformers (ViTs) have demonstrated SOTA performance in various computer vision tasks [16]. On the other hand, the large vision model prelearns collections of pathological images from different organs. Through the deep neural network structure, self-supervised learning, and other techniques, it can automatically extract meaningful features from images and construct a visual representation space with rich semantic information and strong generalization capabilities. Such pretraining enables the model to utilize existing prior knowledge when applied to downstream tasks, thereby reducing its reliance on the amount of data for the target task. Therefore, the Virchow2 large vision model [56] with ViT as the underlying architecture was selected as the feature extractor in this study, aiming at obtaining high-quality generalized image representations and providing a more discriminative and robust feature base for subsequent prediction and classification tasks. And its parameters were frozen during training.

Virchow2 is a ViT architecture with 4 registers, which consists of 32 visual transformer blocks equipped with 16 heads at the attention layer with an embedding dimension of 1280. The model was trained on a large-scale medical image dataset containing 3.1 million whole-slide histology images stained with H&E and immunohistochemistry (IHC) staining and covering tissue samples from multiple parts of the human body, including the stomach and colorectum.

After a tile-level feature is extracted by the pretrained Virchow2 encoder, a feature sequence consisting of the CLS token and the mean patch token is constructed to represent the tile. This feature sequence is then fed into an RNN for feature refinement. Specifically, the RNN model used in this study consists of two GRU [8] layers, with a hidden size of 256. The relevant hyperparameters are listed in Table 2. The GRU's update and reset gates enable adaptive interaction between global information (CLS token) and aggregated local patch tokens, producing the refined tile-level feature. The final feature is passed through a linear classifier to generate tile-level prediction. Tile-level predictions are then aggregated across all tiles to obtain ROI- and WSI-level predictions.

Tiles are extracted from the ROIs corresponding to each label category. However, due to the imbalance in the number of ROIs across different label levels, the resulting number of tiles per class is also unevenly distributed. To address this imbalance, this study introduces an adaptive weighted loss function in addition to employing data augmentation. This function is able to automatically adjust the weights of different categories according to the training state of the model, thus preventing the model from favoring easily classifiable categories and ignoring other categories. Specifically, it dynamically updates class weights based on per-class recall using an exponential moving average (EMA), ensuring that underrepresented classes receive higher emphasis. Furthermore, it combines cross-entropy and focal loss in a weighted manner, with the balance between them adjusted according to the losses from recent epochs.

Formally, let y denote the true label and p the predicted probability vector for a sample. The detailed formulation of the adaptive weighted loss is presented below:

$$L_{AW}(p, y) = w_y \times [\beta L_{CE}(p, y) + (1 - \beta) L_{\text{Focal}}(p, y)] \tag{1}$$

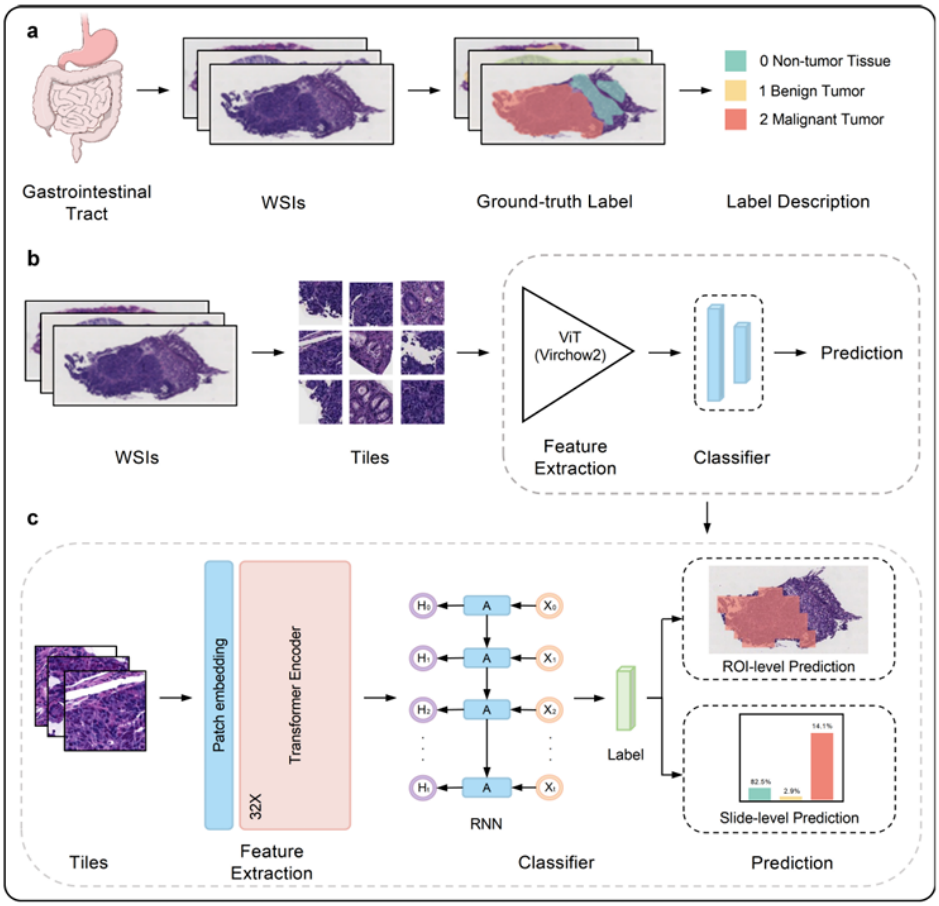


Fig. 2. The workflow of the unified gastrointestinal cancers screening system developed in this study. Subfigure (a) shows the segmentation of tissue regions on pathology slides based on expert annotations, with diagnostic labels assigned to each region, including non-tumor (blue for Level 0), benign tumor (orange for Level 1), and malignant tumor (red for Level 2). Subfigure (b) shows the extraction of tiles from the segmented tissue regions and the use of the system for training and inference. Subfigure (c) shows the training and inference process of the system. The tiles extracted from the WSI are first processed by the pre-trained Virchow2 model for feature extraction. The resulting features are then fed into a recurrent neural network (RNN)-based classifier to generate final predictions, which are subsequently compared with the pathologists' diagnoses for validation

where L_{AW} denotes the adaptive weighted loss; L_{CE} denotes the cross-entropy loss; L_{Focal} denotes the focal loss; w_y denotes the weight of the true class; β controls the relative contribution of cross-entropy loss and focal loss, dynamically adjusted based on recent epoch loss trends.

Each class c is assigned a weight w_c with the range of 1 to w_{max}, which is updated based on its recall using an exponential moving average:

$$w_c^{e_{\text{current}}} = \lambda w_c^{e_{\text{last}}} + (1-\lambda)\frac{1}{\mathrm{r_c} + \epsilon} \tag{2}$$

where $w_c^{e_{\text{current}}}$ denotes the w_c of the current epoch; $w_c^{e_{\text{last}}}$ denotes the w_c of the last epoch; λ denotes the EMA momentum; ϵ prevents division by zero.

The balance weight β is dynamically updated based on the ratio of the latest epoch loss to the historical average loss, which is computed over the past t epochs. If the latest epoch loss decreases by no more than τ compared to the historical average, the balance weight β is decreased by $\triangle\beta$ at each epoch, starting from 1 and with a lower bound of 0. The hyperparameters w_{max} = 2.0, λ = 0.9, t = 3, τ = 0.05, $\triangle\beta$ = 0.02 were determined based on preliminary experiments.

2.4. Training Strategy

In order to improve the robustness and generalization ability of the model training, this study uses a 3-fold cross-validation strategy to optimize the training dataset. Also, the Adam optimizer and cosine scheduler are used for learning rate and weight decay with an initial learning rate of 0.0001. During the training iterations, an adaptive weighted loss function is utilized to calculate the loss and update the model weights based on the minimum loss achieved at the end of each iteration. The training process is configured for a maximum of 200 iterations, with an early stopping mechanism implemented. If the validation loss is not further reduced after 10 consecutive training iterations, the training is terminated. These parameter settings are determined according to experimental performance and may significantly influence the final diagnostic outcomes. Therefore, selecting and tuning appropriate parameters during the experimental process is essential for achieving optimal results. All experiments were performed on an NVIDIA GeForce RTX 3090 using Python version 3.8.19 and Pytorch version 2.3.1. The model hyperparameters and settings are shown in Table 2.

2.5. Quantitative and Statistical Analysis

To evaluate the performance of multi-class classification at the tile level, this study uses class-by-class tile-level sensitivity, specificity, and precision, which are defined as follows, in conjunction with the requirements of medical research and commonly used deep learning evaluation metrics:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

Table 2. The Model Hyperparameters and Settings

Hyperparameter	Setting	Notes
Learning rate	1e-4	Initial learning rate for trainable layers
Optimizer	Adam	Weight decay = 1e-4
Batch size	32	
Maximum training epochs	200	
Early stopping patience	10	Monitored on validation loss
Number of sampled tiles per WSI	82	Average (range:5-1371)
Virchow2 encoder	Frozen	Pretrained encoder
GRU input dimension	1280	Matches feature embedding
GRU hidden dimension	256	
GRU layers	2	Number of stacked GRU layers
GRU dropout	0.3	Applied within GRU and before FC layer
GRU weight init	Xavier(input), Orthogonal(hidden)	Bias zeros, update gate bias=1
GRU output	Last hidden state	Passed to linear classifier

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

where TN, FP, TP, and FN represent true-negative, false-positive, true-positive, and false-negative situations for each category, respectively.

The evaluation at the ROI-level and WSI-level was completed by aggregating the tile-level prediction results within each ROI or WSI. During the training process, no information from the test ROI and WSI is used at all.

To evaluate the network performance at the region-of-interest level, this study draws on the idea of the paper [32] to utilize the tile-level accuracy of each ROI, with the metric defined as follows:

$$Accuracy_{ROI} = \frac{1}{|R_c|} \sum_{r \in R_c} \left(\frac{1}{|P_r|} \sum_{p \in P_r} M_p \cdot \mathbf{1} \left(\widehat{y_p} = y_p \right) \right) \tag{6}$$

where R_c is the set of samples for each region-of-interest level truth class; P_r is the set of tiles within the r_{th} ROI; $\widehat{y}_p$ is the predicted value of the p_{th} tile; y_p is the corresponding tile set label; M_p is the foreground mask, which is set to 1 when a tile accounts for more than 50% of the foreground pixels.

To evaluate the network performance at the WSI-level, this study follows clinical practice guidelines, a slide was classified as high-grade if at least one high-grade tile was identified, with the formula defined as follows:

$$Y_{slide} = \max_{p \in P_w} \mathbb{I}(\hat{y}_p = H) \tag{7}$$

where Y_{slide} is the label of the WSI; P_w is the set of tiles within the WSI; $\widehat{y}_p$ is the predicted label of the p_{th} tile; H denotes the high-grade label.

Considering the inconsistency in the number of samples in each category in the test dataset, this study used the weighted average method to find all the class-averaged assessment metrics, which were calculated using the following formula:

$$Avg = \frac{\sum_{i=1}^{C} n_i \times \text{metric}}{\sum_{i=1}^{C} n_i} \tag{8}$$

where metric is the corresponding metric value; C is the number of categories; n_i is the number of samples in the i_{th} category.

In addition, this study used AUC to compare the classification performance of different models. All metrics were calculated using the Scikit-learn package [35]. To ensure reliable performance assessment, 95% confidence intervals (CI) were calculated for the class-averaged metrics using the standard error method (mean $\pm 1.96 \times$ standard error). To evaluate whether the differences in model performance were statistically significant, the DeLong test was applied to compare AUC values between models on the internal test dataset. A two-sided p-value < 0.05 was considered statistically significant.

3. Results

3.1. Ablation Study

Table 3. Component configurations

Components	Linear Head	Mean Patch Token	GRU	Adaptive Weighted Loss
①	✓	-	-	-
②	✓	✓	-	-
③	✓	✓	-	✓
④	-	✓	✓	-
⑤	-	✓	✓	✓

To assess the contribution of each component in our model, we conducted an ablation study, and the relevant results are summarized in Table 3 and Table 4 [1]. The baseline (①) only uses the CLS token as input to the linear head for classification. Its performance provides a reasonable starting point for the ablation study. The combination of concatenating the CLS token and the mean patch token (②) shows improvements in all metrics, demonstrating that the mean patch token is useful for supplementing global information. Comparisons of combinations ③, ④, and ⑤ indicate that the use of GRU and adaptive weighted loss further enhances the model's discrimination ability. In summary, the ablation results confirm that each module has a positive contribution, and their integration is helpful for improving classification performance.

3.2. Tile-level Classification of Gastrointestinal Tissues Using VGA

In digital pathology image analysis, dividing WSI into tiles for analysis has become a widely accepted strategy in AI-assisted pathology diagnostic systems [49, 10, 5, 26]. This strategy for analysis not only significantly reduces computational complexity but also captures localized pathological features at a finer level. In addition, the tile level facilitates the generation of heat maps, accurate lesion localization, and informed region-level

[1] Note: Sens denotes sensitivity, Spec denotes specificity, Prec denotes precision, and Avg denotes average class indicator value.

Table 4. Ablation study results on the internal test dataset

Components	Level	Sens	Spec	Prec	AUC
①	0 NT	0.9610	0.9919	0.9789	0.9985
	1 BT	0.8769	0.9746	0.8514	0.9892
	2 MT	0.9833	0.9610	0.9729	0.9880
	Avg(95%)	0.9604±0.0104	0.9766±0.0082	0.9610±0.0103	0.9929±0.0045
②	0 NT	0.9633	0.9954	0.9884	0.9987
	1 BT	0.8938	0.9886	0.8662	0.9902
	2 MT	0.9815	0.9864	0.9642	0.9884
	Avg(95%)	0.9627±0.0102	0.9907±0.0052	0.9634±0.0100	0.9933±0.0044
③	0 NT	0.9695	0.9961	0.9902	0.9990
	1 BT	0.9236	0.9901	0.8865	0.9956
	2 MT	0.9827	0.9897	0.9728	0.9958
	Avg(95%)	0.9696±0.0093	0.9927±0.0046	0.9703±0.0090	0.9972±0.0029
④	0 NT	0.9830	0.9952	0.9882	0.9919
	1 BT	0.9002	0.9956	0.9451	0.9780
	2 MT	0.9845	0.9872	0.9678	0.9910
	Avg(95%)	0.9808±0.0086	0.9950±0.0049	0.9743±0.0086	0.9898±0.0054
⑤	0 NT	0.9898	0.9973	0.9932	0.9997
	1 BT	0.9172	0.9968	0.9600	0.9992
	2 MT	0.9910	0.9899	0.9736	0.9962
	Avg(95%)	0.9816±0.0072	0.9941±0.0042	0.9808±0.0074	0.9981±0.0023

decision-making. It serves as a critical intermediate step between low-level feature extraction and high-level clinical inference. Therefore, this study begins with a systematic evaluation of the model's performance at the tile level to verify that it better understands the tile information.

We compared the VGA model developed in this study with three other approaches: H-optimus-0 [37], CHIEF [50], and InceptionV3 [41]. Among them, H-optimus-0 and CHIEF are two models that perform well in existing pathological large vision models [54, 6, 46, 37, 50, 56], while InceptionV3 is a backbone or baseline network that has been widely adopted in previous pathological image analysis studies [48, 10, 29, 24, 34]. All models were trained and tested using the same dataset, following identical experimental procedures to ensure a fair comparison.

The ROCs of the four methods and their AUCs on both the internal and external test datasets are shown in Fig. 3. The tile-level performance metrics of the four methods on the internal and external test datasets are summarized in Table 5 [2] and Table 6 respectively. The VGA method developed in this study achieved an average sensitivity of 0.9816 (95% [CI]: 0.9744, 0.9888) on the internal test dataset, an average specificity of 0.9941 (95% [CI]: 0.9899, 0.9983), an average precision of 0.9808 (95% [CI]: 0.9734, 0.9882) and an average AUC of 0.9981 (95% [CI]: 0.9958, 1.0000). In order to focus on the multi-category classification ability of AI systems in the clinic, we used category-averaged sensitivity as the primary evaluation metric. It can be seen that the VGA method developed in this study achieves higher performance compared to similar methods. Meanwhile, the VGA model developed in this study showed robust generalization performance during external validation, with a class-averaged sensitivity of 0.9161 (95% [CI] 0.9071,0.9251)

[2] Note: p-values are calculated using the DeLong test with VGA as the reference model.

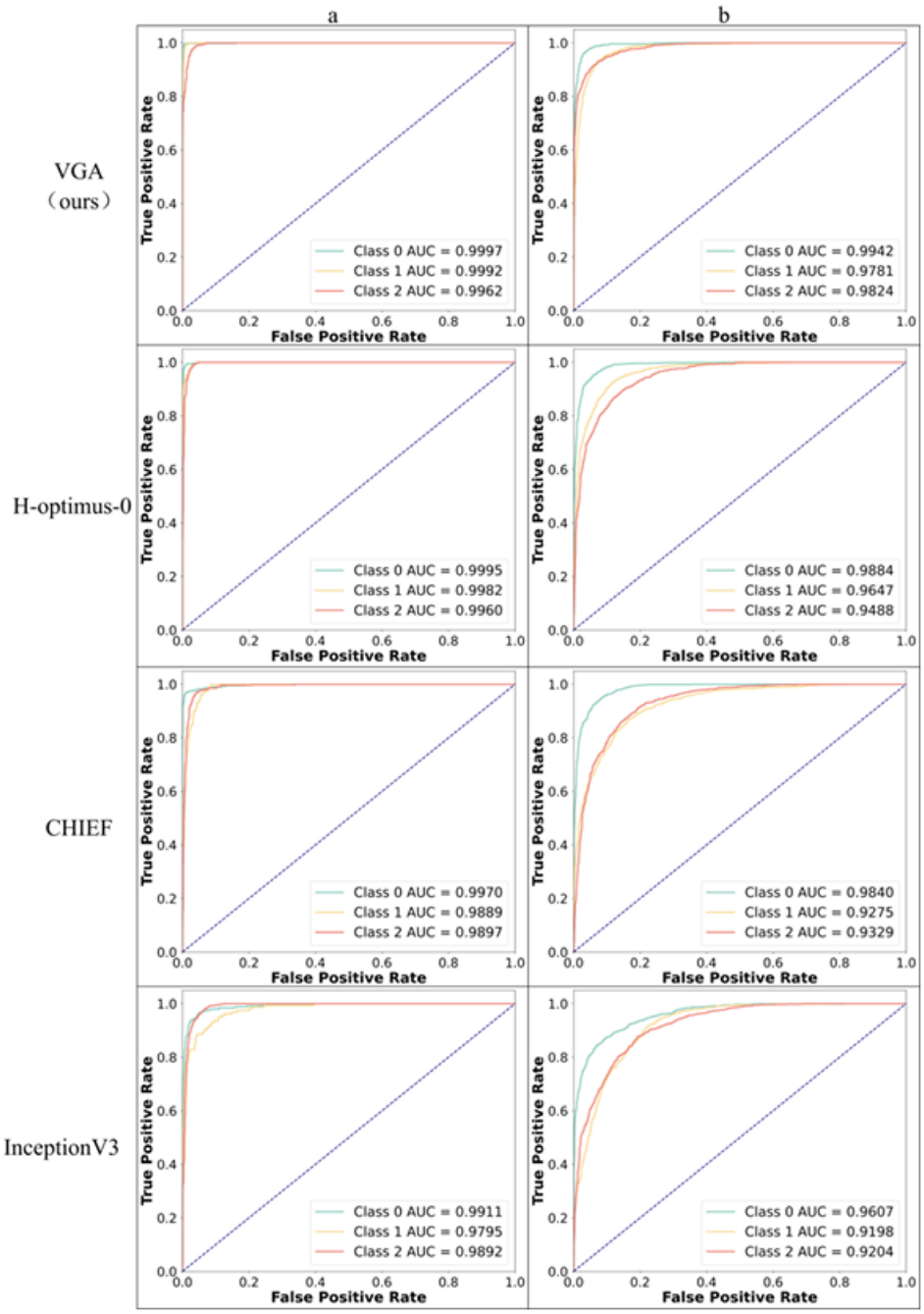


Fig. 3. Receiver operating characteristic curves (ROCs) and their AUCs for the four methods. (a) ROC and AUC of four methods on the internal test dataset. (b) ROC and AUC of four methods on the external test dataset

Table 5. Tile-level performance metrics of different methods on the internal test dataset

Models	Level	Internal Test Dataset				
		Sens	Spec	Prec	AUC	p-value
VGA (Ours)	0 NT	0.9898	0.9973	0.9932	0.9997	-
	1 BT	0.9172	0.9968	0.9600	0.9992	-
	2 MT	0.9910	0.9899	0.9736	0.9962	-
	Avg(95%)	0.9816±0.0072	0.9941±0.0042	0.9808±0.0074	0.9981±0.0023	-
H-optimus-0	0 NT	0.9847	0.9911	0.9781	0.9995	0.0073
	1 BT	0.8981	0.9958	0.9463	0.9982	0.0097
	2 MT	0.9803	0.9893	0.9716	0.9960	0.0270
	Avg(95%)	0.9724±0.0088	0.9909±0.0052	0.9715±0.0090	0.9978±0.0025	-
CHIEF	0 NT	0.9475	0.9883	0.9705	0.9970	0.0005
	1 BT	0.8662	0.9831	0.8095	0.9889	< 0.0001
	2 MT	0.9642	0.9852	0.9607	0.9897	0.0014
	Avg(95%)	0.9449±0.0123	0.9863±0.0063	0.9469±0.0118	0.9929±0.0045	-
Inception V3	0 NT	0.8695	0.9671	0.9144	0.9911	< 0.0001
	1 BT	0.7325	0.9847	0.7986	0.9795	0.0001
	2 MT	0.9409	0.9469	0.8692	0.9892	< 0.0001
	Avg(95%)	0.8835±0.0170	0.9606±0.0105	0.8811±0.0174	0.9889±0.0057	-

Table 6. Tile-level performance metrics of different methods on the external test dataset

Models	Level	External Test Dataset			
		Sens	Spec	Prec	AUC
VGA (Ours)	0 NT	0.9218	0.9765	0.9454	0.9942
	1 BT	0.9021	0.9357	0.9015	0.9781
	2 MT	0.9289	0.9616	0.9114	0.9824
	Avg(95%)	0.9161±0.0090	0.9559±0.0066	0.9179±0.0089	0.9843±0.0040
H-optimus-0	0 NT	0.8934	0.9776	0.9464	0.9884
	1 BT	0.8980	0.9172	0.8763	0.9647
	2 MT	0.8879	0.9530	0.8895	0.9488
	Avg(95%)	0.8936±0.0100	0.9464±0.0072	0.9017±0.0096	0.9672±0.0057
CHIEF	0 NT	0.9405	0.9195	0.8403	0.9840
	1 BT	0.7733	0.9568	0.9212	0.9275
	2 MT	0.8915	0.9305	0.8453	0.9329
	Avg(95%)	0.8598±0.0110	0.9375±0.0078	0.8737±0.0107	0.9464±0.0072
Inception V3	0 NT	0.7469	0.8905	0.7509	0.9607
	1 BT	0.5782	0.9208	0.8266	0.9198
	2 MT	0.8778	0.9872	0.6454	0.9204
	Avg(95%)	0.7194±0.0140	0.9313±0.0081	0.7493±0.0138	0.9325±0.0081

and a class-averaged precision of 0.9179 (95% [CI] 0.9090, 0.9268) on the external test dataset. In contrast, the other methods exhibited a substantial decline in performance, primarily due to overfitting on the internal training dataset. The better tile-level prediction performance of the VGA model can be largely attributed to the use of the Virchow2 large vision model as a feature extractor, which was pre-trained on 3.1 million WSIs, significantly surpassing the pre-training dataset sizes of H-optimus-0 and CHIEF used in the comparison models. It is worth emphasizing that only 5,111 labeled tiles from 76 WSIs (out of a total number of 29,457 tiles) were used for training in this study. Consequently, InceptionV3 exhibited suboptimal performance in this task, which further underscores the advantages of the large vision model as a feature extractor in downstream tasks and provides strong support for the development of subsequent AI-assisted systems.

3.3. ROI-level Classification of Gastrointestinal Tissues Using VGA

Accurate classification at the ROI level is essential for identifying tumor regions and guiding precise pathological diagnoses. It facilitates localized analysis within WSIs, enabling finer-grained decision-making and improving the accuracy of clinical workflows. Table 7[3] shows the average classification accuracy of the ROI-level of the different models on the internal and external test datasets. While the H-optimus-0 model achieved the highest ROI-level class average accuracy on the internal test dataset, its performance on the external test dataset was lower than the VGA model developed in this study. Overall, the VGA model developed in this study demonstrated the stronger generalization performance among the four models.

Table 7. ROI-level performance metrics for different models on the test dataset

Model	$\text{Accuracy}_{\text{ROI}}$	
	Internal Test Set	External Test Set
VGA(Ours)	0.9613	0.9175
H-optimus-0	0.9659	0.9029
CHIEF	0.9301	0.8901
Inception V3	0.8834	0.7121

Some of the ROI-level prediction results of the four models on the test dataset are given in Fig. 4. As illustrated, the VGA model produces predictions that most closely align with the ground truth annotations, demonstrating competitive accuracy compared to the other three models. In contrast, the remaining three models show varying degrees of deviation from the true labels, with InceptionV3 exhibiting the greatest discrepancy. From the clinical point of view, the VGA model developed in this study appears more suitable for application in gastrointestinal cancers screening, where it can assist pathologists in improving diagnostic efficiency and accuracy.

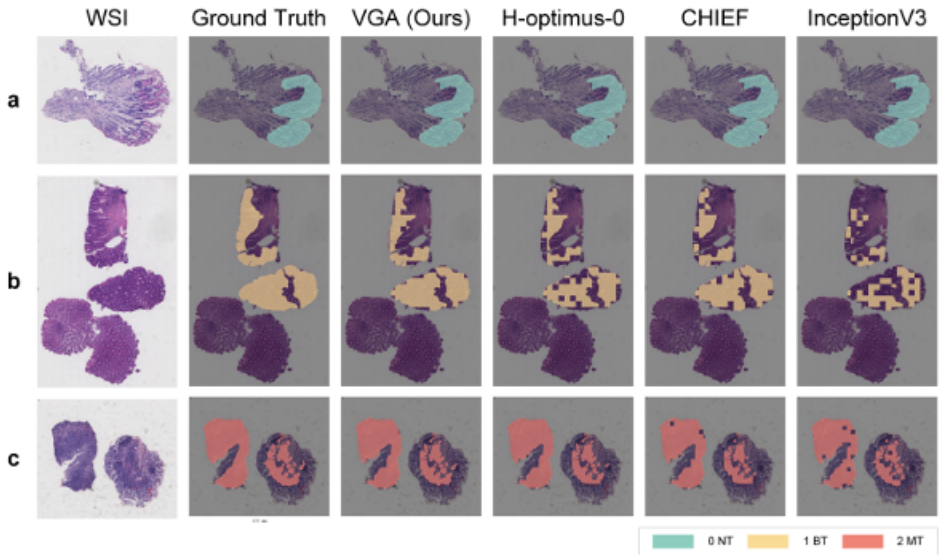


Fig. 4. ROI-level prediction results for different models on the test dataset

3.4. VGA's WSI-level Cancer Screening Capabilities

To evaluate the generalizability of the proposed algorithm across multi-center datasets, this study performed WSI-level testing on an external test dataset and an independent validation dataset selected from TCGA. Following clinical practice guidelines, a slide was classified as high-grade if at least one high-grade tile was identified. Among the 48 WSIs in the external test dataset, 28 were classified as non-malignant and 20 as malignant. At the same time, we invited a pathologist to review these 48 WSIs, confirming that among the malignant predictions, 5 were non-malignant, while the remaining were correctly classified. In the WSI-level test, the sensitivity of the VGA model for malignant tumor slide detection was 100%, and the specificity was 84.85%, indicating its ability to accurately exclude non-malignant tumor slides. In the independent validation dataset (100 slides), the VGA model achieved 98.00% sensitivity and 82.00% specificity for malignant tumor slide detection, further demonstrating its robust generalization performance. Fig. 5a and Fig. 5b illustrate the confusion matrix plots of the VGA model on the external test dataset and the independent validation dataset. Meanwhile, the average time for the VGA model to analyze a WSI is 25 seconds. This means that in combination with the initial screening results from the VGA model, pathologists may focus on high-risk areas in WSIs in a short period of time, thus improving the efficiency and accuracy of gastrointestinal cancers screening.

Fig. 6 shows a heatmap generated from a subset of slides during WSI-level testing of the VGA model on the external dataset. More heatmaps are presented in Fig.1 and Fig.2 of the Supplementary Materials. The diagnostic predictions of the VGA model are based on the classification probability outputs of all tiles within each slide, which can be used

[3] Note: $\text{Accuracy}_{\text{ROI}}$ indicates the average accuracy of the ROI class.

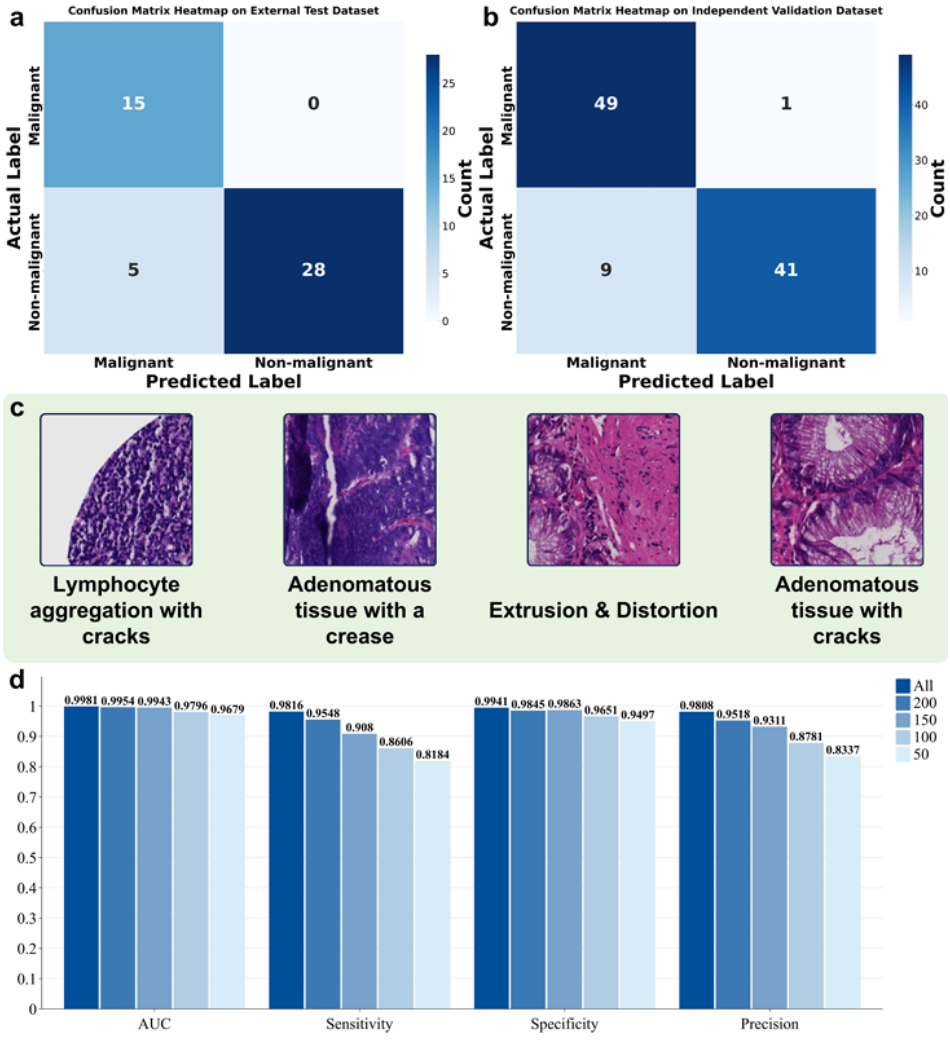


Fig. 5. Results visualization. (a) Confusion matrix for the external test dataset. (b) Confusion matrix for the independent validation dataset. (c) Representative error tiles for the VGA model when tested at the WSI level. (d) Performance of the VGA model on the internal test dataset under different sample-scarcity conditions (All indicates that the VGA model is trained using the full training dataset, 200 indicates that 200 samples of each class in the training dataset are taken in equal quantities to form a new training dataset to train the VGA model, and so on; all the metrics in the graphs are class averages)

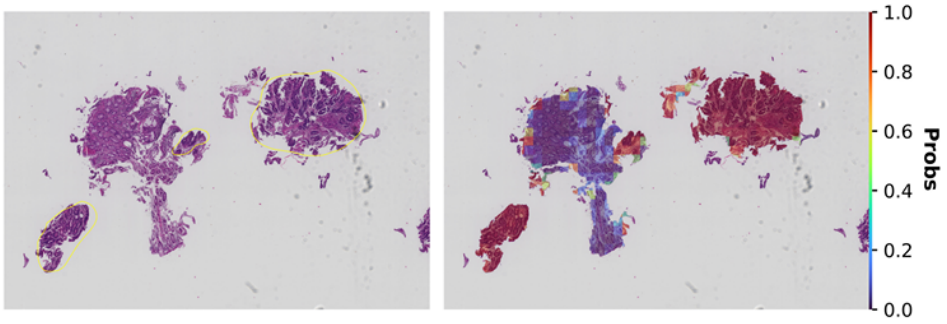


Fig. 6. The heatmap of a slide during WSI-level testing of the VGA model.

to visualize the localization of highly suspicious foci on malignant tumor slides. In the heatmap, warmer colors indicate areas where themodel assigns a higher probability of malignant tumor presence.

In this study, we recorded the error cases during WSI-level testing and further analyzed the frequently occurring misclassifications. In the five slides that were misdiagnosed as malignant, extrusion, distortion, creasing, or cracking of the tissue had a significant impact on the diagnostic results. As shown in Fig. 5c, an area of lymphocyte aggregation accompanied by cracks was misdiagnosed as malignant. Similarly, the adenomatous tissue was also misdiagnosed as malignant due to the presence of an obvious crease. In another case, the morphological deformation of the fibrous mesenchyme after extrusion led to the misdiagnosis. Additionally, the adenomatous tissue was again misdiagnosed as malignant due to the accompanying cracks. The results indicate that the VGA model developed in this study is susceptible to errors when dealing with poor-quality slides or cases involving partially benign tumors and requires adjunctive examination by a pathologist for accurate determination.

3.5. Sample Less Scenario Testing

To further validate the applicability and robustness of the proposed model under low-resource conditions, we conducted sample adaptation evaluation experiments. Specifically, we randomly selected 50, 100, 150, and 200 tiles per category from the internal training dataset to simulate clinical scenarios with varying levels of data availability. The number of WSIs used has also gradually increased from 9 to 20 accordingly. In each setting, we evaluated the classification performance and confusion matrix performance of the model in a three-level organizational activity classification task to analyze the trend of the impact of changes in data volume on model performance. Fig. 5d shows the performance metrics of the VGA algorithm for the internal test dataset for the three organizational activity levels under different sampling conditions. As the number of samples per class increased from 50 to 200, the classification performance of the model also improved accordingly. When using 200 samples per class, the model achieved an average AUC of 0.9954 (95% [CI] 0.9917,0.9991), an average sensitivity of 0.9548 (95% [CI] 0.9436,0.9660), an average specificity of 0.9845 (95% [CI] 0.9778,0.9912), and a mean precision of 0.9518 (95% [CI] 0.9402,0.9634). Combined with Table 5, it can be seen

that the performance of the model on the internal test dataset at this point is close to the results at full training, with small differences observed in the evaluation metrics. Increasing the number of tiles further yields marginal perfomance gains. Therefore, we limit the few-shot evaluation to up to 200 tiles per class.

Beyond the empirical results, it is also important to understand why the model remains effective when the number of training tiles is limited. This can be explained in part by the strong representation reuse enabled by large-scale pretraining. Through the pretrained feature extractor, the encoder retains general visual representations learned from extensive pretraining, which can be effectively reused for downstream classification tasks[36]. Furthermore, freezing the encoder while training only a lightweight classification head reduces the number of trainable parameters, leading to faster training and lower computational cost compared to fine-tuning the entire model. The reduced number of trainable parameters also helps mitigate the risk of overfitting on smaller target datasets[17]. Consequently, when the model is trained using only a small number of image tiles for each category, it can maintain stable performance rather than experiencing a sharp decline.

4. Discussion

Timely diagnosis and screening of gastrointestinal cancers are important to improve patient outcomes and survival. Currently, histopathologic analysis remains the gold standard for the diagnosis of gastrointestinal cancers. However, on the one hand, there is a shortage of pathologists and a long training period for pathologists all over the world, including the United States and low- and middle-income countries [3]. On the other hand, insufficient diagnostic experience can lead to missed diagnoses and misdiagnoses, significantly impacting subsequent treatment [55]. Therefore, these challenges highlight the urgent need for reliable tools to assist in pathological image analysis and gastrointestinal cancers screening, with the goal of enhancing diagnostic efficiency. At the same time, advances in digital pathology have created a practical foundation for the integration and deployment of AI models in clinical diagnostic workflows [28].

WSIs from the stomach and colorectum were collected from different institutions for training, testing, and external validation of a universal screening model for gastrointestinal cancers, referred to as VGA, developed in this study. Unlike previous studies that either trained separate models for stomach and colorectum data using the same network architecture or used a model trained on one organ to validate its generalization on the other, this study simultaneously incorporated data from both organs for model training and performed classification to achieve unified screening.

The model was trained using a simplified three-class system (NT, BT, MT) to facilitate rapid screening. However, this simplification overlooks important clinical distinctions. For example, the NT class includes both normal mucosa and inflammatory conditions, which have different clinical implications, while the MT class groups high-grade adenomas with adenocarcinomas, which would normally require more granular subtyping. This simplified classification is suitable for initial screening, but future work will focus on extending the model or integrating it into workflows that require finer diagnostic resolution to better reflect clinical complexity.

The results showed that the AI screening model developed in this study achieved a class-averaged sensitivity of over 0.9161 and a class-averaged precision of over 0.9179

on both the internal and external test datasets, demonstrating improvements over the other three baseline networks. Meanwhile, in the WSI-level test on the external dataset, the sensitivity of malignant tumor slides reached 100%, which demonstrates the model's potential to assist pathologists in confirming diagnoses or prompting further investigation when discrepancies arise in the initial assessment. The AI screening model developed in this study could potentially be integrated into the digital pathology workflow by automatically predicting the diagnosis and highlighting the suspected malignant areas after slide scanning. This enables pathologists to prioritize suspicious cases, thereby speeding up the diagnosis. Additionally, we observed that occasional misdiagnoses tended to occur when it processed slides of poor quality, such as those with tissue extrusion, distortion, creases, or cracks. Future work could consider marking such slides to prompt pathologists to re-evaluation, rescanning or recutting when necessary. However, it cannot be denied that the performance of the current model relies on high-quality input.

Despite the promising performance of the proposed AI screening model, several limitations and deployment assumptions should be acknowledged. First, the current framework operates at the tile level and is not yet capable of directly interpreting individual ROIs or entire WSIs. Future work will focus on developing effective tile-level feature aggregation methods to enable WSI-level prediction and interpretation. Second, this study was limited to gastrointestinal organs including the stomach and colorectum only, and data from more organs will be included in the future to construct a generalized screening model for gastrointestinal cancers. Third, the dataset used was derived from only three medical centers, and further validation on more diverse datasets is required to assess robustness and reduce potential dataset bias.

From a deployment perspective, several practical considerations also exist. The model assumes access to high-quality WSIs; variations in staining protocols, scanning resolution, or slide preparation may affect prediction accuracy. Furthermore, real-time integration into digital pathology workflows may face computational constraints, such as inference speed and hardware requirements, which could affect scalability in high-throughput settings. Finally, the current framework relies on frozen encoders with pretrained representations, which may limit adaptability to rare or out-of-distribution cases. Addressing these limitations through broader multi-center datasets, improved WSI-level modeling, prospective clinical validation and optimization for real-time clinical deployment will be important directions for future work.

5. Conclusion

In summary, previous AI-assisted systems have tended to be organ-specific tasks or for pan-cancer detection. We have developed a unified screening model for gastrointestinal cancers named VGA, which is capable of reliably classifying WSIs from the stomach and colorectum in multiple categories. With the help of the developed AI system, pathologists can perform more rapid screening of gastrointestinal cancers and lay the foundation for subsequent treatment.

Code Availability. The code are available from the corresponding author upon reasonable request.

References

1. Barmpoutis, P., Waddingham, W., Yuan, J., Ross, C., Kayhanian, H., Stathaki, T., Alexander, D.C., Jansen, M.: A digital pathology workflow for the segmentation and classification of gastric glands: Study of gastric atrophy and intestinal metaplasia cases. Plos one 17(12), e0275232 (2022)
2. Bilal, M., Tsang, Y.W., Ali, M., Graham, S., Hero, E., Wahab, N., Dodd, K., Sahota, H., Wu, S., Lu, W., et al.: Development and validation of artificial intelligence-based prescreening of large-bowel biopsies taken in the uk and portugal: a retrospective cohort study. The Lancet Digital Health 5(11), e786–e797 (2023)
3. Black-Schaffer, W.S., Morrow, J.S., Prystowsky, M.B., Steinberg, J.J.: Training pathology residents to practice 21st century medicine: a proposal. Academic pathology 3, 2374289516665393 (2016)
4. Cai, C., Shi, Q., Li, J., Jiao, Y., Xu, A., Zhou, Y., Wang, X., Peng, C., Zhang, X., Cui, X., et al.: Pathologist-level diagnosis of ulcerative colitis inflammatory activity level using an automated histological grading method. International Journal of Medical Informatics 192, 105648 (2024)
5. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine 25(8), 1301–1309 (2019)
6. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine 30(3), 850–862 (2024)
7. Cheung, J., Savine, S., Nguyen, C., Lu, L., Yasin, A.S.: Transfer learning from one cancer to another via deep learning domain adaptation (2026)
8. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734. Association for Computational Linguistics, Doha, Qatar (Oct 2014), https://aclanthology.org/D14-1179/
9. Choi, S., Kim, S.: Artificial intelligence in the pathology of gastric cancer. Journal of Gastric Cancer 23(3), 410 (2023)
10. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A.: Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine 24(10), 1559–1567 (2018)
11. Da, Q., Wang, S., Wang, W., Yang, C., Wang, B., Ruan, M., Fu, Z., Xu, Y., Zhou, Y., Wang, C., et al.: Progress and challenges of pathological artificial intelligence in the era of large models. Zhonghua bing li xue za zhi= Chinese journal of pathology 54(3), 305–309 (2025)
12. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., Williamson, D.F.K., Robertson, H., Chen, B., Almagro-Perez, C., Doucet, P., Sahai, S., Chen, C., Chen, C.S., Komura, D., Kawabe, A., Ochi, M., Sato, S., Yokose, T., Miyagi, Y., Ishikawa, S., Gerber, G., Peng, T., Le, L.P., Mahmood, F.: A multimodal whole-slide foundation model for pathology. Nature Medicine 31(11), 3749–3761 (Nov 2025)
13. Du, Y., Liu, X., Yue, L., Feng, L., Tao, P., Jing, Q.: Minidigpath: A new standard for pathology images few-shot learning classification. In: 2023 2nd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE). pp. 136–140 (2023)
14. Fu, B., Zhang, M., He, J., Cao, Y., Guo, Y., Wang, R.: Stohisnet: A hybrid multi-classification model with cnn and transformer for gastric pathology images. Computer Methods and Programs in Biomedicine 221, 106924 (2022)

15. Griem, J., Eich, M.L., Schallenberg, S., Pryalukhin, A., Bychkov, A., Fukuoka, J., Zayats, V., Hulla, W., Munkhdelger, J., Seper, A., et al.: Artificial intelligence–based tool for tumor detection and quantitative tissue analysis in colorectal specimens. Modern Pathology 36(12), 100327 (2023)
16. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence 45(1), 87–110 (2022)
17. Hasan, K.R., Kim, S., Cho, J., Han, H.S.: Prototypical few-shot learning for histopathology classification: Leveraging foundation models with adapter architectures. IEEE ACCESS 13, 86356–86379 (2025)
18. Hinata, M., Ushiku, T.: Detecting immunotherapy-sensitive subtype in gastric cancer using histologic image-based deep learning. Scientific reports 11(1), 22636 (2021)
19. Huang, B., Tian, S., Zhan, N., Ma, J., Huang, Z., Zhang, C., Zhang, H., Ming, F., Liao, F., Ji, M., et al.: Accurate diagnosis and prognosis prediction of gastric cancer using deep learning on digital pathological images: A retrospective multicentre study. EBioMedicine 73 (2021)
20. Iizuka, O., Kanavati, F., Kato, K., Rambeau, M., Arihiro, K., Tsuneki, M.: Deep learning models for histopathological classification of gastric and colonic epithelial tumours. Scientific reports 10(1), 1504 (2020)
21. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. Computational and structural biotechnology journal 16, 34–42 (2018)
22. Korbar, B., Olofson, A.M., Miraflor, A.P., Nicka, C.M., Suriawinata, M.A., Torresani, L., Suriawinata, A.A., Hassanpour, S.: Deep learning for classification of colorectal polyps on whole-slide images. Journal of pathology informatics 8, 30 (2017)
23. Lan, J., Chen, M., Wang, J., Du, M., Wu, Z., Zhang, H., Xue, Y., Wang, T., Chen, L., Xu, C., et al.: Using less annotation workload to establish a pathological auxiliary diagnosis system for gastric cancer. Cell Reports Medicine 4(4) (2023)
24. Le Page, A.L., Ballot, E., Truntzer, C., Derangère, V., Ilie, A., Rageot, D., Bibeau, F., Ghiringhelli, F.: Using a convolutional neural network for classification of squamous and non-squamous non-small cell lung cancer based on diagnostic histopathology hes images. Scientific Reports 11(1), 23912 (2021)
25. Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., Parwani, A.V., Zhang, A., Mahmood, F.: A visual-language foundation model for computational pathology. Nature Medicine 30(3), 863–874 (Mar 2024)
26. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering 5(6), 555–570 (2021)
27. Ma, M., Zeng, X., Qu, L., Sheng, X., Ren, H., Chen, W., Li, B., You, Q., Xiao, L., Wang, Y., et al.: Advancing automatic gastritis diagnosis: An interpretable multilabel deep learning framework for the simultaneous assessment of multiple indicators. The American Journal of Pathology 194(8), 1538–1549 (2024)
28. Moxley-Wyles, B., Colling, R.: Artificial intelligence and digital pathology: where are we now and what are the implementation barriers? Diagnostic Histopathology (2024)
29. Mudeng, V., Farid, M.N., Ayana, G., Choe, S.w.: Domain and histopathology adaptations–based classification for malignancy grading system. The American Journal of Pathology 193(12), 2080–2098 (2023)
30. Nagtegaal, I.D., Odze, R.D., Klimstra, D., Paradis, V., Rugge, M., Schirmacher, P., Washington, K.M., Carneiro, F., Cree, I.A., et al.: The 2019 who classification of tumours of the digestive system. Histopathology 76(2), 182 (2019)
31. Neidlinger, P., El Nahhas, O.S.M., Muti, H.S., Lenz, T., Hoffmeister, M., Brenner, H., van Treeck, M., Langer, R., Dislich, B., Behrens, H.M., Rocken, C., Foersch, S., Truhn, D., Marra, A., Saldanha, O.L., Kather, J.N.: Benchmarking foundation models as feature extractors for weakly supervised computational pathology. Nature Biomedical Engineering (Oct 2025)

32. Oh, Y., Bae, G.E., Kim, K.H., Yeo, M.K., Ye, J.C.: Multi-scale hybrid vision transformer for learning gastric histology: Ai-based decision support system for gastric cancer treatment. IEEE journal of biomedical and health informatics 27(8), 4143–4153 (2023)
33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024)
34. Park, S.y., Ayana, G., Wako, B.D., Jeong, K.C., Yoon, S.D., Choe, S.w.: Vision transformers for low-quality histopathological images: A case study on squamous cell carcinoma margin classification. Diagnostics 15(3), 260 (2025)
35. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. the Journal of machine Learning research 12, 2825–2830 (2011)
36. Rahman, T., Baras, A.S., Chellappa, R.: Evaluation of a task-specific self-supervised learning framework in digital pathology relative to transfer learning approaches and existing foundation models. Modern Pathology 38(1), 100636 (2025)
37. Saillard, C., Jenatton, R., Llinares-López, F., Mariet, Z., Cahané, D., Durand, E., Vert, J.P.: H-optimus-0 (2024), https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0
38. Song, Y., Wang, T., Cai, P., Mondal, S.K., Sahoo, J.P.: A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. ACM Comput. Surv. 55(13s) (Jul 2023)
39. Song, Z., Zou, S., Zhou, W., Huang, Y., Shao, L., Yuan, J., Gou, X., Jin, W., Wang, Z., Chen, X., et al.: Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nature communications 11(1), 4294 (2020)
40. Srinidhi, C.L., Ciga, O., Martel, A.L.: Deep neural network models for computational histopathology: A survey. Medical Image Analysis 67, 101813 (2021)
41. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016)
42. Tsuneki, M., Kanavati, F.: Weakly supervised learning for poorly differentiated adenocarcinoma classification in gastricendoscopic submucosal dissection whole slide images. Technology in Cancer Research & Treatment 21, 15330338221142674 (2022)
43. Tung, C.L., Chang, H.C., Yang, B.Z., Hou, K.J., Tsai, H.H., Tsai, C.Y., Yu, P.T.: Identifying pathological slices of gastric cancer via deep learning. Journal of the Formosan Medical Association 121(12), 2457–2464 (2022)
44. Veldhuizen, G.P., Röcken, C., Behrens, H.M., Cifci, D., Muti, H.S., Yoshikawa, T., Arai, T., Oshima, T., Tan, P., Ebert, M.P., et al.: Deep learning-based subtyping of gastric cancer histology predicts clinical outcome: a multi-institutional retrospective study. Gastric Cancer 26(5), 708–720 (2023)
45. Vinay Kumar, Abul K˙ Abbas, D.J., Aster, J.C.: Robbins & Cotran Pathologic Basis of Disease. Elsevier, Illinois, USA (2020)
46. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature medicine 30(10), 2924–2935 (2024)
47. Wang, J., Liu, X.: Medical image recognition and segmentation of pathological slices of gastric cancer based on deeplab v3+ neural network. Computer methods and programs in biomedicine 207, 106210 (2021)
48. Wang, K.S., Yu, G., Xu, C., Meng, X.H., Zhou, J., Zheng, C., Deng, Z., Shang, L., Liu, R., Su, S., et al.: Accurate diagnosis of colorectal cancer based on histopathology images using artificial intelligence. BMC medicine 19, 1–12 (2021)

49. Wang, S., Zhu, Y., Yu, L., Chen, H., Lin, H., Wan, X., Fan, X., Heng, P.A.: Rmdl: Recalibrated multi-instance deep learning for whole slide gastric image classification. Medical image analysis 58, 101549 (2019)
50. Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. Nature 634(8035), 970–978 (2024)
51. Wang, Z., Peng, H., Wan, J., Song, A.: Identification of histopathological classification and establishment of prognostic indicators of gastric adenocarcinoma based on deep learning algorithm. Medical molecular morphology pp. 1–13 (2024)
52. Wei, J.W., Suriawinata, A.A., Vaickus, L.J., Ren, B., Liu, X., Lisovsky, M., Tomita, N., Abdollahi, B., Kim, A.S., Snover, D.C., et al.: Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. JAMA network open 3(4), e203398–e203398 (2020)
53. Xie, Y., Shi, L., He, X., Luo, Y.: Gastrointestinal cancers in china, the usa, and europe. Gastroenterology report 9(2), 91–104 (2021)
54. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature 630(8015), 181–188 (2024)
55. Yang, Z., Wei, T., Liang, Y., Yuan, X., Gao, R., Xia, Y., Zhou, J., Zhang, Y., Yu, Z.: A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. Nature Communications 16(1), 2366 (2025)
56. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)

Lijue Liu is a associate professor at the School of Automation, Central South University, with research interests in image processing, data mining, and intelligent information processing.

Fangjie Yin is a master's candidate at the School of Automation, Central South University, with research interests in image processing.

Genjian Yang is an Engineer at China Electronic Product Reliability and Environmental Testing Research Institute. His research interests include multimodal algorithms and game learning.

Qi Li is an attending physician in the Department of Pathology, Beijing Integrated Traditional Chinese and Western Medicine Hospital. Her main research focuses on the pathological diagnosis of digestive tract, thyroid, gynecological and other diseases.

Siya Li is a researcher at CAS Blue Bay Cloud Technology (Guangdong) Co., Ltd focusing on artificial intelligence and medical research and development, with interests in AI-driven medical image analysis and AI applications in precision medicine.

Teng Pan has dedicated her master's, doctoral, and postdoctoral research to investigating the tumor microenvironment and angiogenesis in breast cancer, with extensive research experience in breast cancer biology. She obtained her PhD in Oncology from Tianjin

Medical University. Her research work has been published in several journals, including Cancer Letters, British Journal of Cancer, and Theranostics.

Ting Liu is employed at Beijing Ditan Hospital, with research interests including data mining and hepatocellular carcinoma-related studies.

Jin Tang is a professor at the School of Automation, Central South University, with research interests in computer vision, industrial intelligence, and application of large language models.

Ruijie Ming is a physician in the Department of Oncology at Chongqing University Three Gorges Hospital, with research interests in tumor therapy resistance, the tumor microenvironment, and tumor multi-omics analysis.

Yu Song is an associate chief physician and deputy director in the Department of Otolaryngology Head and Neck Surgery, Peking University First Hospital, with research interests in surgical treatment of allergic rhinitis and diagnosis and treatment of rhinobasal diseases.

Xue Feng is an attending physician in the Department of Respiratory and Critical Care Medicine, Tianjin Chest Hospital, with research interests in respiratory critical care medicine and respiratory central regulation.

Dan Wang is a postdoc at Richard Dimbleby Laboratory of Cancer Research, Randall Division and Division of Cancer and Pharmaceutical Sciences, King's College London, with interests in the application of AI in medicine, tumor microenvironment.

Xingang Zhou is a chief physician in the Department of Pathology, Beijing Ditan Hospital, Capital Medical University, specializing in digital and artificial intelligence pathology as well as liver pathology.

Wenbai Chen is a professor at the Beijing Information Technology Science and University, with research interests in include multimodal algorithms and pattern recognition.

Jinhai Deng specialized in the tumor microenvironment throughout his master's and doctoral studies, and possesses extensive research expertise in tumor biology and tumor immunology. He obtained his PhD from King's College London, UK. His research findings have been published in journals including EMBO Molecular Medicine, British Journal of Cancer, Theranostics and so on.

Received: November 30, 2025; Accepted: March 30, 2026.

Computer Science and Information Systems 23(2):827–860 https://doi.org/10.2298/CSIS251020021K

A Pilot Study of Multi-Method Evaluation of Machine Translation in Macedonian

Jana Kuzmanova[1] and Katerina Zdravkova[2]

[1] Faculty of Computer Science and Engineering
jana.kuzmanova@finki.ukim.mk
[2] Faculty of Computer Science and Engineering
katerina.zdravkova@finki.ukim.mk (corresponding author)

Abstract. This pilot study offers a linguistic evaluation of six machine translation systems: GPT-4o, GPT-5, Gemini 2.5 Flash, Google Translate, Microsoft Translator, and NLLB-600M applied to the translation of a short excerpt of Orwell's "1984" into Macedonian. The analysis consisted of three interconnected experiments: manual annotation of translation errors and comparison with human output, evaluation using eight popular MT metrics, and sentence-level similarity analysis via cosine similarity, Jaccard similarity, and Levenshtein distance. Manual annotation revealed that stylistic errors (48.47%) and linguistic errors (34.54%) were the most common. The LLMs outperformed other systems, particularly GPT-5, while NLLB-600M performed poorly, often introducing incomprehensible sentences or non-existent words. Metrics-based evaluation showed that lexical metrics sometimes penalized fluent and accurate translations that deviated from the reference. Sentence similarity analysis confirmed that accurate translations were more consistent, while wrong–wrong sentence pairs were more divergent, especially in Levenshtein scores. The findings underscore the importance of combining manual and metric-based evaluation to fully understand MT quality, particularly in low-resource settings.

Keywords: Machine Translation, Manual Evaluation and Annotation, Linguistic Similarity, Low Resource Language

1. Introduction

Bibliophiles enjoy reading novels in the original language when they are fluent in it, but they also appreciate high-quality translations, particularly when they capture the spirit of the original. In some cases, translation can even surpass the source text, especially when it brings creativity and cultural insight to the work [1]. An excellent translation is both an expertise and an impressive reproductive art [2]. It not only conveys the meaning and nuances of the original text, but also flows naturally, reflecting the translator's deep understanding of the culture, history, and linguistic subtleties of the source [3,4].

Unfortunately, achieving high-quality translation is not always an easy activity [5]. Many readers have, at some point, given up reading an otherwise excellent literary work due to poor translations [6]. A major turning point in improving translation quality has been the development of machine translation (MT) tools. These tools can significantly facilitate the process of converting a text from one language to another, while striving to preserve the meaning of the original.

Recently, MT systems have improved rapidly, particularly due to their shift to neural machine translation (NMT). NMT is mainly based on deep learning (DL) and large training datasets [7]. DL involves training artificial neural networks with many layers, enabling the models to automatically learn hierarchical features and complex representations from vast amounts of data.

New MT systems have become an essential support for readers who attempt to understand texts written in languages entirely unfamiliar to them. Yet, a critical question remains: can MT reach the quality of a skilled human translator? This study seeks to explore this question by comparing a human translation of Orwell's "1984" into Macedonian with translations produced by six leading MT systems.

Building on this inquiry, our pilot research investigates how the following MT systems: ChatGPT (GPT-4o and GPT-5), Gemini (2.5 Flash), Google Translate, Microsoft Translator and NLLB-600M handle literary language, by focusing on a small excerpt from a well-known and widely translated novel "1984" by George Orwell. This novel was selected not only for its literary and political significance, but also because it forms the basis of the MULTEXT project [8]. Moreover, as part of the MULTEXT-East project [9], a professionally produced human translation of "1984" has been part of speech (POS) [10] and morpho-syntactically annotated and mutually aligned multiple languages, including Macedonian [11,12]. These resources allowed us to conduct a detailed manual evaluation of machine-translated outputs against a trusted human reference. The Macedonian reference translation was used with written permission from the publisher for academic research purposes within the MULTEXT-East framework. In accordance with these conditions, the study does not redistribute the full text and includes only short illustrative examples for analysis.

For each machine-generated sentence, we assessed whether the translation was accurate or contained errors. By an accurate translation, we refer to one that faithfully preserves the meaning, intent, and relevant nuances of the source text in the target language, without distortion, omission, or unwarranted addition [13,14].

The identified translation errors were grouped into four clusters, reflecting both the linguistic levels and key dimensions of translation quality:

1. Linguistic accuracy, referring to grammatical accuracy regardless of the meaning;
2. Stylistic errors, addressing issues such as inappropriate word choice, tone, or awkward phrasing;
3. Lexical and semantic errors, evaluating how well the intended meaning of the source text is preserved;
4. Fluency and naturalness, evaluating the readability and native-like quality of the translation.

Inspired by Tolstoy's famous opening line from Anna Karenina "All happy families are alike; each unhappy family is unhappy in its own way" [15], we adapted the sentiment to reflect our findings: "Every accurate translation is alike; every wrong translation is inaccurate in its own way." This phrase emerged from our manual evaluation of machine translation results, where we noticed that accurate translations tended to converge in form and quality, while inaccurate ones differed significantly in their errors.

This motivated us to explore sentence-level similarity analysis using three different methods:

1. Lexical/surface similarity (BLEU, TER, chrF)
2. Embedding-based similarity (BERTScore, Sentence-BERT, cosine similarity)
3. Edit-based distance (Levenshtein)

We further examined the variance in similarity scores, hypothesizing that accurate translations would be more consistent and clustered closely, while erroneous outputs would show greater divergence. In addition, we investigated how specific error types relate to automated metric scores and assessed the extent to which metric-based rankings align with human judgments. We also evaluated the literalness of machine translation, comparing it to both automatic quality estimates and human evaluations.

Finally, we visualized these results to reveal the underlying patterns and clarify the relationships between translation quality, error characteristics, and translation similarity.

2. Review of Recent Studies that Evaluate MT Systems

Several studies have evaluated and compared commercial and publicly available MT systems. As an example,[16] performed an automatic evaluation of the systems submitted to the WMT22 General Machine Translation task, employing chrF (Character n-gram F-score) [17], BLEU (Bilingual Evaluation Understudy) [18], and COMET (Crosslingual Optimized Metric for Evaluation of Translation) [19] metrics. A total of 185 systems were evaluated across 11 language pairs and 21 translation directions. These pairs included Czech – English, Czech – Ukrainian, French – German, German – English, English – Chinese, English – Croatian, English – Japanese, English – Russian, English – Ukrainian, as well as more distant and low-resource combinations such as Russian – Yakut and English – Livonian. The systems were ranked using the same three metrics after which a statistical significance testing was performed. Additional statistics were computed for each system, including: the number of sentences identical to the reference; the number of sentences that differed between the MT output and the reference, despite being the same in the source; and sentence normalization to assess how punctuation changes affect BLEU and COMET scores. The results revealed significant discrepancies between BLEU and COMET scores for some systems. It is worth mentioning that COMET did not agree with the top-ranked BLEU and chrF systems in 11 out of 21 translation directions. BLEU and chrF gave identical rankings for only 5 language pairs, although their rankings were generally similar. Statistical significance testing indicated that a difference of 0.9 BLEU points was considered meaningful.

Similarly,[20] found that a difference of 2 to 3 BLEU points is statistically significant. However,[21] demonstrated that achieving a pairwise accuracy of 85%, defined as the proportion of system pairs for which the automatic metric ranks systems in the same order as human evaluators, requires a minimum difference of 3.35 BLEU points between system scores. It is important to note that BLEU never reaches 90% pairwise accuracy, even when the score differences are larger. Because different metrics operate on different scales, the score improvement required to reach a given pairwise accuracy threshold can vary significantly. For example, chrF can achieve 90% pairwise accuracy with an improvement of 3.05 points, while COMET can reach 95% accuracy with an increase of only 1.18 points for some models. In contrast, for lexical metrics like BLEU and chrF, the required score difference is typically larger, and pairwise accuracy decreases when comparing unrelated systems. Additionally, Kocmi et al. (2021), who first proposed the pairwise accuracy

metric, noticed that in cases where human evaluations disagree with BLEU rankings, the median BLEU score difference between systems is 1.3 points.

[22] compared two versions of ChatGPT (3.5 and 4), Google Translate, and DeepL using BLEU, chrF, and TER (Translation Edit Rate) [23] for all translation combinations between English, German, Chinese, Japanese, and Romanian. They also computed word prediction accuracy based on frequency, grouped scores by sentence length, and conducted a human evaluation of the number of errors each system produced. Human evaluation revealed that GPT-4 was the best-performing model, despite having lower BLEU scores than Google Translate.

[24] conducted a study that analyzed the quality of literary translations from English into Dutch. A 500-word short story was translated using three open-access NMT engines: DeepL, Systran, and Google NMT. The target translations were evaluated for accuracy, fluency, and style. The evaluation was performed by human reviewers and through automated metrics, including BLEU scores, supplemented by a Dutch literariness algorithm. The study also highlighted specific words and phrases in the source text that demanded careful handling across the assessed categories. Consistent with the findings of Jiao et al., a significant discrepancy was observed between human judgments and BLEU scores.

[25] also evaluated generative pre-trained transformer (GPT) models using a combination of automatic metrics and human judgment. The evaluation covers translations between English and a range of other languages, including French, German, Czech, Icelandic, Chinese, Japanese, Russian, Ukrainian, and Hausa. The evaluation metrics included two versions of COMET: COMET-22 and the reference-free COMETkiwi, as well as BLEU and chrF scores. Translations were assessed at both the sentence and document level. The results were grouped according to prompt design and selection, language resource level (high vs. low), and domain. Consistent with previous studies, human evaluation indicated that traditional lexical metrics do not fully capture improvements in translation quality, whereas COMET exhibited a stronger correlation. The study further analyzed distinctive characteristics of GPT-generated translations compared to those produced by NMT systems, focusing on dimensions such as non-monotonicity, fluency, punctuation insertion, and dropped or inserted content.

Other studies also examine the distinguishing characteristics of MT systems. Focusing on German-to-English translation,[26] investigate translation artifacts that differentiate human, NMT, and LLM-generated outputs. They classify sentences as either original or translated and employ explainability methods to identify the features contributing to these classifications. Although BLEU and COMET scores are reported, no correlation is found between classification accuracy and translation quality. The authors apply leave-one-out and integrated gradient techniques to analyze both feature overlap and feature frequency. Feature overlap is assessed by identifying the most influential features and calculating Jaccard similarity scores to compare them across systems. Feature frequency is analyzed through POS distributions grouped by sentence length. The findings indicate that LLM translations exhibit artifacts more similar to those observed in human translations than in NMT outputs, although notable differences remain.

A similar comparison between machine and human translations is conducted by[27]. They perform word-based and arc-based analyzes of English–Chinese, English–French, and English–German translations. They identify POS patterns in the source text and examine the corresponding target-side patterns, focusing on distribution, conditional entropy,

and convergence. These metrics are aggregated according to frequency. Their findings suggest that translation quality and structural divergence are not directly related; however, human translations consistently exhibit greater variability and divergence than machine-generated ones. [28], meanwhile, investigate the literalness of translations produced by GPT models in comparison to NMT systems. Literalness is measured by the number of unaligned source words and by non-monotonicity, defined as deviations in word order. Their analysis spans English – German and English – Russian translations in both directions and is supported by human evaluation. The results show that GPT translations contain more unaligned source words overall, while higher non-monotonicity is observed only when translating from English into other languages.

Given the limitations of automated metrics noted above, MT evaluation is typically carried out by human annotators. However, the WMT24 Translation Task employed a preliminary system ranking based on automated metrics, due to the high volume of submissions [29]. Two top-performing metrics from the WMT23 Metrics Task were used: MetricX-23-XL and COMETkiwi-DA-XL. The latter, a reference-free metric, was included specifically to mitigate reference bias. The evaluation covered a wide range of language pairs, including translations from English into Czech, German, Spanish, Hindi, Icelandic, Chinese, Japanese, Russian, and Ukrainian, as well as translations between Czech and Ukrainian and between Japanese and Chinese.

However, human evaluation is also far from straightforward. Scoring practices and evaluator expertise can significantly influence results. Direct Assessment (DA), i.e., the practice of assigning scores on a 0–100 scale, can be unreliable; however, this limitation can be partially mitigated by supplementing the scale with descriptive labels. A more comprehensive evaluation framework is the Multidimensional Quality Metrics (MQM) system [30], which focuses on error annotation. Freitag et al. (2021) [31] proposed an MQM-based scheme that incorporates categories for accuracy, fluency, style, and locale, along with error severities and a method for computing aggregate scores.

[32] adapted MQM for Slavic languages by incorporating agreement features such as person, number, gender, and case into the core tagset. Both studies found MQM to yield higher inter-annotator agreement than DA or Scalar Quality Metrics (SQM). However, MQM annotation is time-consuming and costly. An alternative approach, error-span annotation [33], focuses only on marking problematic segments and their severity. This method achieves strong alignment with MQM rankings, demonstrates higher inter-annotator agreement, and is more efficient to conduct.

The human evaluation of literary translation is also explored by [34], with a focus on the performance of large language models (LLMs). The study evaluates translations involving English, Polish, Russian, Czech, French, Japanese, and Chinese as source languages, with English, Japanese, and Polish as targets. Evaluations are conducted at both the sentence and paragraph levels. Annotators are presented with two MT outputs (one generated at the sentence level, the other at the paragraph level) and are asked to mark error spans from categories including mistranslation, untranslated segments, grammar, inconsistency, register, and formatting. They are also required to select their preferred version and provide freeform justifications. To investigate potential data memorization, the authors assessed whether masked named entities can be predicted by the model, which was generally not the case. While paragraph-level translation yields improvements over sentence-level output, critical errors may still persist.

An alternative method for MT evaluation involves the use of test suites, as demonstrated by[35]. Their approach employs a curated set of linguistic phenomena alongside regular expression rules to automatically detect correct and incorrect translations, supplemented by manual evaluation where necessary. The study focuses on translations from English into German and Russian. Accuracy is calculated as the proportion of correct translations. Statistical significance is also assessed. The results are aggregated by linguistic category and phenomenon. For both languages, case agreement, prepositional multi-word expressions, and date formatting are among the most accurately translated categories. Conversely, idioms and semantic role assignments remain challenging. Additionally, German translations struggle with rare verb tenses, while Russian systems face difficulties with compounds and verbal multi-word expressions.

The above-mentioned papers, particularly recent WMT shared-task evaluations [36], illustrate substantial variation in MT performance across language pairs, depending on resource availability. While some pairs involving traditionally lower-resource languages (e.g., Czech–Ukrainian) achieve relatively strong results with human evaluation, likely aided by linguistic relatedness, other pairs such as English–Icelandic exhibit markedly lower performance compared to high-resource benchmarks. These results highlight that translation quality in lower-resource settings is influenced not only by data size but also by factors such as language similarity and available training corpora.

Recent research reviewed in this section highlights both the potential and the limitations of current MT systems, particularly in literary and multilingual contexts. While LLMs and NMT systems show notable progress, traditional lexical metrics often fail to capture nuanced improvements in quality. Human evaluation remains essential, although methods vary in reliability and cost. Emerging alternatives such as error-span annotation and test suites offer promising, more scalable solutions. Overall, fine-grained, linguistically informed evaluation remains crucial for assessing translation quality across diverse languages and text types.

Building on the insights and challenges identified in the literature, the following section presents our methodology and methods, which draw upon and extend the approaches reviewed to provide a comprehensive evaluation framework for MT quality.

3. Methodologies and Methods

Our study was initially designed to evaluate the quality of five MT systems, selected for their widespread use and demonstrated proficiency in the Macedonian language. Based on an extensive review process, these five systems were confirmed as the focus of our initial evaluation: GPT-4o [3], Gemini 2.5 Flash [4], Google Translate [5], Microsoft Translator [6], and NLLB-600M [37]. Unfortunately, one of the most popular NMT services in Europe, DeepL [7], still does not support the Macedonian language. However, in early August 2025, the highly anticipated release of GPT-5 [8] was announced, generating significant interest

[3] https://chatgpt.com/
[4] https://gemini.google.com/
[5] https://translate.google.com/
[6] https://translator.microsoft.com/
[7] https://www.deepl.com/en/translator
[8] https://chatgpt.com/

in both academic and technological communities. Recognizing its potential impact, we promptly expanded the scope of our research to include GPT-5, ensuring that our study reflects the most up-to-date trends in the field.

All MT systems were used through their respective web interfaces as available in August 2025, with the exception of NLLB, which was used through HuggingFace. The systems were used with their default parameters. The prompt used for GPT-4o, GPT-5, and Gemini was a simple zero-shot prompt as follows: "Translate the following sentences from English into Macedonian. Return one sentence per line:", followed by the sentences, each one in a new line.

To avoid exceeding the free translation limits of these MT systems, we focused on the first 100 sentences of Orwell's novel. An additional reason for selecting these specific sentences is that, without exception, they are aligned 1:1 in the human translation from English to Macedonian. For the LLM prompts, the sentences were chunked in two batches of 44, and a remaining batch of 22 sentences.

The linguistic evaluation consisted of two phases: manual annotation of machine translations and sentence-level similarity assessment. Each phase is explained in more detail in the following subsections.

3.1. Manual Annotation of Translation Errors

Manual error annotation and classification were performed by a single expert annotator, who was in continuous consultation with two senior linguists from the Institute of Macedonian Language, who acted as domain experts and external reviewers. The evaluation was further supported by systematic cross-linguistic comparison with existing human translations in Serbian, Croatian, and Bulgarian available within the MULTEXT-East framework [9].

All translations were compiled in a Google Docs spreadsheet consisting of six worksheets, one dedicated to each MT system. Each worksheet included five columns: the source English sentence, the corresponding MT output, the official human translation into Macedonian, a column for mnemonic labels identifying observed errors, and a section for detailed comments explaining those errors.

The corpus of 100 sentences was examined in detail, and all observed errors were initially recorded and labelled in the comments. These errors were then reviewed, filtered, and consolidated through an iterative process, resulting in a final set of 39 distinct error types that were characteristic of all evaluated translations, many of which were language-specific (LS).

To facilitate evaluation, the identified errors were categorized into four main clusters, as outlined in the introduction: Linguistic inaccuracy, Stylistic errors, Lexical and semantic errors, and Fluency and naturalness. Each cluster was further divided into three subgroups, based on shared characteristics and recurring patterns observed across the corpus.

The resulting taxonomy was established through a systematic comparison of existing error-classification frameworks, with particular attention to their cognitive and systemic characteristics. Although this comparative analysis informed the structure of the taxonomy, it was developed independently and was not directly validated against the MQM framework.

According to Polio (1997) [38], linguistic accuracy refers to the degree to which the rules of a language are correctly applied, which in the case of MT includes grammar, vo-

cabulary, and syntax. We determined that the corresponding cluster of errors encompasses the following broad categories: morphological disagreement, verb and tense accuracy, and syntactic structure.

Morphological disagreement refers to mismatches in grammatical categories between words that are expected to agree [39]. Within this category, we identified four recurring error types: gender disagreement, incorrect plural formation, and number disagreement.

Verb and tense accuracy relate to the correct use of verb forms to ensure clarity and temporal consistency in translation [40]. Irregular verb form refers to the incorrect use of the active voice when the passive voice would be more appropriate. Past tense selection in Macedonian language depends on whether the speaker directly witnessed or participated in the event or is simply reporting it indirectly [41]. This subcategory includes not only tense and verb form errors, but also person disagreement, which refers to a mismatch between the subject and the verb's inflectional marking.

Syntactic structure is a broad term encompassing grammatical rules, word order, and the construction of well-formed sentences [42]. We limited this group to four specific error types: missing accusative clitic, word order errors, wrong definiteness, and wrong preposition. The key reason why missing accusative clitics form part of this classification is that they are usually associated with the incorrect use of prepositions, whose function in sentence structures is crucial for its linguistic accuracy.

Stylistic errors are writing choices that impair clarity or appropriateness, even if they are not strictly ungrammatical [43]. These errors make the text harder to read or less natural, although the intended meaning usually remains understandable.

The use of calques, literal translations, foreign language insertions, and pleonasms falls under the subgroup of literalism. These errors involve transferring structures or redundant expressions from the source language, resulting in unnatural or inaccurate target-language output [44].

Punctuation and formatting errors include missing and wrong punctuation, unnecessary use of quotation marks, and use of lowercase letters in formal expressions. These errors occur when the translator does not follow the target language's conventions or fails to appropriately adapt the source text's style [45]. They can significantly impair readability, alter meaning, and undermine the professionalism of the translated text.

Finally, we include spelling errors, misused synonyms, and incorrect phrasing under the subgroup of vocabulary inaccuracies. Vocabulary inaccuracy refers to the incorrect use of words in terms of spelling, form, or contextual meaning. Such errors can lead to miscommunication, obscure ideas, and indicate an imprecise understanding of the target vocabulary [46].

Translation errors sometimes stem from a lack of knowledge, carelessness, or insufficient competence. MT systems are prone to such errors, as they are often embedded in the training data they rely on [47]. These types of errors are grouped in the third cluster named Lexical and semantic errors. It comprises three broad subcategories: terminological errors, untranslatability, and contextual misinterpretation.

Terminological errors involve the incorrect use of terms specific to specialized domains [48]. When such terms are mistranslated, the result is a word or phrase with an inaccurate or inappropriate meaning in the target language. In our experiment, these errors appeared in the form of incorrect adverbs, medical term, misinterpretation of Orwell's Newspeak, and improper word usage.

Untranslatability refers to the difficulty or even impossibility of conveying the exact meaning of a source text in the target language [49]. In our case, this subcategory includes inconsistent translations of the same word or phrase, untranslated terms that alter the original meaning, and instances where transliteration was used in place of accurate translation.

Contextual misinterpretation occurs when the translator fails to accurately interpret the cultural, situational, or linguistic context of the source text [3]. This can result in translations that are inaccurate, misleading, or even inappropriate. In the MT systems we analyzed, such errors typically involved issues with sentence structure, such as incorrect POS or mistranslations caused by a lack of information from preceding context.

The last group deals with errors related to fluency and naturalness in translation. These criteria assess how well the translated text can be understood in the target language. Fluency refers to grammatical correctness and the smooth flow of a sentence [50], while naturalness encompasses broader cultural and idiomatic appropriateness, ensuring that the translation reads as if it were originally written in the target language, rather than as a literal translation from another language [50].

Grammatical correctness is often compromised by omissions. In the MT systems we examined, the omissions included missing conjunctions, prepositions, verbs, and entire phrases. Notably, many observed errors also involved the invention of non-existent words.

We grouped such errors alongside untranslated phrases and words into a subcategory labeled non-existent or untranslated items. The final subcategory, unnatural expressions, included meaningless translations, overly descriptive phrasing, and misuse of conjunctions, all of which disrupt the natural flow and tone of the target text.

After manually evaluating and annotating all the six MT systems, we compared machine and human translations by estimating sentence similarity. In parallel, the five best-performing MT were also compared with one another using sentence similarity techniques. These techniques are introduced in the following subsection.

3.2. Assessing Translation Quality with Automated Metrics

To assess the quality of MT outputs produced by various systems, we employed a comprehensive set of both lexical and learned evaluation metrics. Lexical metrics operate primarily at the surface level, focusing on the overlap between the generated translation and a human reference. Among these, we included BLEU [18], a precision-oriented metric that measures the n-gram overlap between system output and reference translations. BLEU computes a geometric mean of modified n-gram precisions (typically up to 4-grams) and applies a brevity penalty to penalize overly short outputs.

In addition to BLEU, we used chrF [17], which calculates F-scores based on character-level n-gram matches. chrF has been shown to correlate better with human judgments than BLEU for morphologically rich languages and translations involving high lexical variation, as it captures finer-grained patterns of correspondence that word-level metrics may miss.

Another lexical metric we used was TER [23], which quantifies the number of edits: insertions, deletions, substitutions, and shifts needed to convert a system translation to reference. For these three metrics, their sacrebleu implementation [51] was used.

We also considered METEOR (Metric for Evaluation of Translation with Explicit Ordering [52]), which aligns words using not only exact matches, but also stems, synonyms, and paraphrases, and incorporates a fragmentation penalty to account for word order.

Beyond surface-level lexical evaluation, we also integrated a set of embedding-based and learned metrics that leverage deep neural models to estimate translation quality with a stronger focus on meaning and contextual understanding. Among these, we used COMET [19], a regression-based metric trained on human quality judgments. We also employed XCOMET-XL [53], a recent extension of COMET that leverages larger pretrained language models, showing improved correlation with human evaluation, especially for high-resource language pairs and longer documents.

To further complement our evaluation suite, we included COMETkiwi [54], a reference-free variant of COMET. This metric estimates the quality of a translation using only the source and the system output, without requiring a human reference.

Finally, we used BERTscore [55], which computes similarity between token embeddings derived from BERT models, aligning words in the hypothesis and reference based on cosine similarity. BERTscore captures semantic similarity more effectively than traditional lexical metrics and has been shown to correlate well with human judgments at both the sentence and the system levels.

3.3. Sentence Similarities Between Different Machine Translation Systems

In addition to evaluating translation quality through standard metrics, we also investigated the pairwise sentence-level similarities between the outputs of different MT systems. The goal was to check if systems fail in the same or different ways when they make errors, as well as to profile the overall diversity between the machine translations of different systems.

To quantify these similarities, we employed three complementary metrics: cosine similarity based on multilingual MPNet sentence embeddings (Masked and Permuted Pre-training for Language Understanding) [56], Jaccard similarity, and Levenshtein distance.

The cosine similarity metric was computed using sentence embeddings generated by a multilingual variant of MPNet, a transformer-based model known for its effectiveness in capturing contextual semantic information. MPNet combines masked language modelling and permuted language modelling for improved sentence representation. A higher cosine similarity indicates that two translations share similar meanings, even if they use different words or structures, making this metric especially useful for identifying cases where different systems arrive at semantically equivalent translations via different surface forms.

In contrast, Jaccard similarity operates on a set-based lexical level, measuring the overlap between the unique tokens (e.g., words or character n-grams) in two translation hypotheses. It is calculated as the size of the intersection divided by the size of the union of the token sets. This metric reflects vocabulary-level similarity and is sensitive to synonymy and word order changes, often underestimating similarity in semantically equivalent but lexically diverse translations. Nevertheless, it provides a simple and interpretable measure of how much two systems reuse similar words, which can be helpful in studying system redundancy or diversity.

The third metric, Levenshtein distance (also known as edit distance), quantifies the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one translation into another. We normalized this value to allow comparability across sentences of varying lengths. While Levenshtein distance is primarily a surface-level similarity metric, it offers an intuitive view of how closely aligned the outputs are in terms of literal sequence. In cases where translations are nearly identical except for minor

variations or errors, Levenshtein distance provides a direct estimate of edit effort, which is especially relevant in post-editing scenarios or for assessing system robustness.

To summarize the relationships between system outputs, we computed the mean similarity scores across all sentence pairs for each metric, effectively creating a system-level similarity matrix. Furthermore, to explore how output similarity correlates with translation quality, we analyzed the distribution of similarity scores within sentence pairs grouped according to manual evaluation labels (e.g., accurate vs. inaccurate translations).

4. Results and Discussions

The linguistic evaluation of machine translations from English to Macedonian, which is the subject of this paper, consists of three interconnected experiments. The first experiment involved defining the key error categories and the clusters they naturally belong to. Following this, all translations were manually annotated and compared with human translations. In the second experiment, the quality of each system was assessed using the eight metrics, which were introduced in Section 3.2. The third experiment focused on evaluating sentence similarity through complementary linguistic metrics: cosine similarity, Jaccard similarity and Levenshtein distance.

The results of all three experiments will be presented and discussed in a unified section, with the findings of the first experiment illustrated in detail using examples from both machine and human translations.

4.1. Analysis of Manually Annotated Translation Errors

This section is organized into four subsections, each presenting the results and corresponding discussion for one of the three experiments.

Linguistic Inaccuracy The number of manually annotated errors across all ten types of linguistic inaccuracies is presented in Table 1. In the remainder of this subsection, each type is explained in detail and, where necessary, illustrated with representative examples, demonstrating that these errors not only compromise grammatical correctness but also hinder comprehension of the translated text.

Linguistic inaccuracy accounts for 34.64% of all observed errors, with the majority of 72.99% originating from Microsoft Translator and NLLB-600M. Most of these errors stem from incorrect tense usage, which is not entirely unexpected. Even native Macedonian speakers occasionally confuse the so-called L-form with the traditional past tense, such as the aorist. Although both refer to past actions, the aorist is used when the speaker directly witnessed or participated in the event, whereas the L-form implies the speaker learned of the action indirectly [41]. In the context of Orwell's 1984, Winston recounts events he personally experienced, making the use of the L-form inappropriate. Incorrect tense selection also appeared in GPT-4o during the final ten sentences, suggesting a LLM decline in consistency toward the end of the output. Interestingly, GPT-5 corrected nearly all tense selection errors, with only one exception, similarly to Gemini and Google Translate. These systems appear to favor the L-form, which is stylistically appropriate for the narrative prose of the novel and therefore does not result in frequent verb selection errors.

Table 1. Distribution of linguistic inaccuracy

Linguistic inaccuracy		GPT-4o	GPT-5	Gemini 2.5-Flash	Google Translate	Microsoft Translator	NLLB-600M
Morphological disagreement	Gender disagreement	1	2	1	0	7	20
	Incorrect plural formation	0	0	0	1	0	0
	Number disagreement	0	0	0	0	1	2
	Person disagreement	0	0	0	1	2	1
Verb and tense accuracy	Incorrect verb voice	0	2	0	1	1	0
	Wrong tense selection	8	1	1	2	73	46
	Missing accusative clitic	1	1	2	2	1	2
Syntactic structure	Word order	3	3	0	2	2	4
	Wrong definiteness	5	5	7	12	2	8
	Wrong preposition	1	0	0	2	1	8
Total errors		19	14	11	23	90	91

In contrast, plural formation was generally accurate, except for the noun *виор (transliterated as vior / English: eddy)*, where Google Translate "invented" the plural form *виорови (viorovi)* instead of the correct *виори (viori / eddies)*. Namely, the plural suffix -ови is rarely used in Macedonian and typically appears with monosyllabic masculine nouns, such as *бит – битови (bit – bitovi / bit – bits)*, *рог – рогови (rog – rogovi / horn – horns)*, and *слон – слонови (slon – slonovi / elephant – elephants)*. The noun *виор* is disyllabic *ви-ор* [57].

Number and person agreement were handled successfully by all three LLMs: both GPT versions and Gemini. The three MT systems also performed confidently. However, gender disagreement between personal pronouns, adjectives, and their corresponding nouns was widespread. This issue was especially prevalent in Microsoft Translator and NLLB-600M, with the latter displaying nearly three times as many occurrences. Microsoft's system misclassified typically feminine nouns ending in a, such as *книга (kniga / book)* and *хартија (hartija / paper)*, as masculine. Meanwhile, NLLB-600M tended to treat most nouns and adjectives as neuter, regardless of their actual grammatical gender.

The wrong definiteness of Macedonian nouns and adjectives was the second most frequent linguistic inaccuracy MT issue with 5.45% presence among all errors. However, it is a frequent problem even to native speakers [58]. They often overlook or misuse definiteness, largely due to the influence of English, which handles definiteness differently and more simply. This tendency is reinforced by frequent exposure to English through media and education, code-switching, and informal digital communication, leading to structural simplification and reduced grammatical awareness in native usage [59].

Incorrect voice selection (active vs. passive) did not pose significant problem to most MT systems, except in the case of GPT-5, which mistakenly used the active voice instead of the passive on two occasions. We note that a change in verb voice does not necessarily result in an incorrect translation, as active–passive alternations may preserve propositional meaning. In our annotation scheme, however, incorrect voice selection was marked only in cases where the choice of voice affected discourse structure, information focus, or stylistic appropriateness in the target language. In Macedonian, passive constructions, such as the reflexive passive and, less frequently, the periphrastic passive is commonly used in colloquial and narrative contexts to background the agent or maintain textual cohesion. In the annotated instances, the source sentence employed a passive construction with no explicit

agent, while the MT output rendered it as an active clause with an overt or implied agent, thereby altering the information structure and narrative perspective. Such cases were therefore classified as incorrect voice selection, despite the semantic content remaining largely intact.

Word order errors are diverse. For example, both GPT versions translated the sentence There were no windows in it at all. as *Во него немаше прозорци воопшто (Vo nego nemashe prozorci voopshto)* instead of the more natural *Во него воопшто немаше прозорци. (Vo nego voopshto nemashe prozorci)*. This reflects a literal translation of the phrase at all, which in English is typically placed at the end of a clause to add emphasis in negative statements, conditional clauses, and questions [9]. In Macedonian, however, adverbs such as *воопшто* should be placed next to the verb they modify [60]. Since *воопшто* complements *нема - немаше (nema – nemashe / there is not – there was not)*, it should immediately precede or follow the verb, not the object. Another error, found in both GPT versions and NLLB-600M, is the use of *било кој* instead of the correct *кој било (koj bilo / whichever)*. This reversed order of the pronoun *кој* and the particle *било* is typical of Serbian and frequently appears in colloquial Macedonian, but it is not standard.

A closer examination of the observed linguistic inaccuracies reveals that weaknesses in gender recognition and tense selection manifest differently across MT systems. Gender-related errors are primarily linked to morphological agreement and long-distance dependency resolution, reflecting persistent difficulties in correctly propagating gender features in morphologically rich languages such as Macedonian. In contrast, tense-related inaccuracies, often involving the choice between aorist, imperfect, and L-form constructions appear to stem more from stylistic normalization and system-level preferences rather than from a failure to identify temporal reference per se. While both types of errors affect grammatical correctness, gender mismatches tend to be more disruptive to local syntactic well-formedness, whereas tense-related errors more subtly undermine narrative coherence and stylistic authenticity.

Stylistic Errors While linguistic inaccuracies affect grammatical structure and clarity, stylistic errors undermine the tone, register, and expressive intent of the original text. In this subsection, such issues are analyzed in terms of literalism, punctuation and formatting, and vocabulary inaccuracies (Table 2).

Stylistic errors represent the core weakness of the MT systems examined, accounting for 48.32% of all annotated errors. More than half of them involve the use of distant synonyms, primarily for nouns, where the intended meaning becomes difficult to infer without knowing the original novel. For example, the noun landing from the sentence “On each landing, opposite the lift-shaft, the poster with the enormous face gazed from the wall.” is translated as *меѓукатие (megjukatie / mezzanine)* by both GPT translators, as *под (pod / floor)* and *слетување (sletuvanje / flight landing)* by all other MT systems. The translation of lift-shaft is even more confusing: *лифт-шахта (lift-shahta / lift manhole)*, *шахта на лифтот (shahta na liftot / manhole for the lift)*, *вратило за лифт (vratilo za lift / lift axle)* all appear. Two translators simply skipped to translate it, while the correct translation *отвор за лифт (otvor za lift / elevator shaft)* was produced only by Microsoft Translator. Even the verb gazed was mistranslated using the close synonym *гледа (gleda / watch, see)*,

[9] www.collinsdictionary.com/dictionary/english/at-all

Table 2. Distribution of stylistic errors

Stylistic errors		GPT-4o	GPT-5	Gemini 2.5-Flash	Google Translate	Microsoft Translator	NLLB-600M
Literalism	Calque	0	0	1	2	2	3
	Foreign language	2	0	0	0	4	33
	Pleonasm	0	0	1	1	2	0
	Too literal translation	1	0	2	2	1	0
Punctuation and formatting	Missing punctuation	2	2	0	1	4	0
	Quotation marks	0	0	2	3	1	0
	Lowercase letters in formal expressions	0	1	0	0	1	0
	Wrong punctuation	4	2	0	0	3	0
Vocabulary inaccuracies	Spelling error	0	1	1	4	1	24
	Misused synonyms	23	21	33	45	23	40
	Incorrect phrasing	7	7	8	7	2	16
Total errors		39	34	48	65	44	116

although the verb *зјана (zjapa)* would more accurately reflect Orwell's intention to convey the act of staring.

Nearly 70% of Google Translate's stylistic errors were due to misused synonyms (45 out of 65 errors in total), indicating that this type of mistake is the primary contributor to its stylistic inaccuracies. Even the most accurate systems among the six studied GPT-4o and GPT-5 averaged more than one stylistic error per five sentences. Integrating a high-quality interpretive bilingual resource, such as the Digital Dictionary of the Macedonian Language [10] could significantly reduce these errors. Unfortunately, due to copyright restrictions, it is currently unavailable for implementation in MT systems.

Incorrect phrasings were also frequently observed, particularly in NLLB-600M. For example, in half of the MT systems, the phrase black-mustachio'd face was mistranslated as a face (made) of a black moustache, rather than a face with a black moustache. Macedonian offers several dedicated adjectives for this expression, including *црномустаќест (crnomustaќjest)*, *црномустак (crnomustak)*, and *црномустаклест (crnomustaklest)*. Notably, both GPT versions produced *црномустак (crnomustak)*, demonstrating that MT systems can generate this less frequent but legitimate Macedonian word correctly, even if it is not the most common form.

Foreign language errors were significantly recorded only with NLLB-600M. This system is directly integrated with Meta's social network, which, according to the company, supports 25 billion translations per day across more than 200 languages This extensive multilingual reach was clearly reflected in our experiment. Specifically, 39 foreign words were detected in the output: 28 in Bulgarian, 9 in Serbian, 1 in Russian, and 1 in Old Church Slavonic. Many of these intrusions were immediately spotted due to the use of Cyrillic characters, which do not exist in the Macedonian alphabet. They include, for example, Bulgarian *одвън (odvŭn / outside)*, *гладък (gladŭk / smooth)*, or *жълта (zhŭlta / yellow)*. Most other Slavic words, for example the Bulgarian *бутилка (butilka / bottle)*, *етажи (etazhi / flights)*, *парцални (parcalni / rag)*, the Russian *слева (sleva / to the left)* and Old Slavonic *ввирено (vvireno / corrugated)* do not exist even in the dialects, thus the Macedonian reader cannot recognize them. The Serbian words: *бели (beli / white)*

[10] http://drmj.eu/

in *бели бетон (beli beton / white concrete)* or *тупи (tupi / blunt)* are acceptable, mainly because the Macedonian adverbs are almost identical: *бел (bel)* and *тап / tap)*. The word *њушкајќи (njushkajkji)*, which, judging by the suffix *-јќи* could be interpreted as a verbal adverb formed from the Serbian verb *њушкати (njushkati, "to snoop")* combined with the Macedonian suffix *-јќи*, is nevertheless unclear. Specifically, the word *њушка (njushka)* exists in several dialects only as a noun, meaning snout or trunk, and is predominantly used in reference to animals.

Calques appeared only rarely. However, without exception, all MT systems that are not LLMs mistranslated the phrase indoor display, each in a different way. The accurate translation: *затворен простор (zatvoren prostor)* is a well-established expression in Macedonian, yet it was not used by any of the systems.

Punctuation errors, particularly missing or misused commas, were also recorded, with 9 instances of each type. This issue is not unique to MT systems; it is also prevalent across many Macedonian news portals, especially those that do not employ official proofreaders. Once legislation enforcing mandatory editing of published texts is fully implemented, the frequency of such errors is expected to decline.

Incorrect phrasing accounts for 13.50% of all stylistic errors. Except for Microsoft Translator, whose performance was almost perfect, such errors occurred nearly ten times in each of the other systems, highlighting a serious and recurring issue. A potential solution would be the integration of a bilingual English – Macedonian phraseological dictionary [11], although it is currently available only in a searchable format and remains under copyright protection. Interestingly, pleonasms, overly literal translations, wrongly used quotation marks and lowercase letters in formal expressions were rare and therefore fall outside the scope of our proposed strategies for improving MT accuracy.

Stylistic errors, ranging from imprecise synonym choices to incorrect phrasing, represent a major obstacle for producing natural, fluent translations. Addressing these challenges will require both better linguistic modelling and access to high-quality bilingual lexical resources.

Lexical and Semantic Errors Lexical and semantic errors arise when the system fails to convey the basic meaning of the source text, in our study due to terminological errors, untranslatability, and context misinterpretation (Table 3.). These errors go beyond grammar or word choice. Namely, they reflect a fundamental misunderstanding of the original sentence and sometimes significantly distort the intended message.

These errors, accounting for just 4.46% of the total, are significantly fewer than those in other categories, particularly since a quarter of them stem not from mistranslation, but from the transliteration of Newspeak terms. During manual annotation, we encountered five pairs of inconsistently translated words and phrases using the NLLB-600M, an inconsistency we did not expect to be on such a large scale.

GPT-4o proved to be a highly competent translator, avoiding the typical errors associated with this category. Its upgraded version, GPT-5, had less consistent handling of these structural elements. For example, the named entity (NE) Victory Mansions at the beginning of the novel was translated as *Победничките Палати (Pobednichkite Palati / Victory Palace)*, while later it appeared as *Победничките Згради (Pobednichkite Zgradi / Victory Buildings)*. According to standard English – Macedonian bilingual dictionaries,

[11] https://zoze.mk/en-mk/

Table 3. Distribution of lexical and semantic errors

Lexical and semantic errors		GPT-4o	GPT-5	Gemini 2.5-Flash	Google Translate	Microsoft Translator	NLLB-600M
Terminological errors	Wrong adverb	0	0	0	0	1	1
	Wrong medical term	0	0	0	1	0	1
	Wrong Newspeak translation	0	0	1	1	0	0
Untranslatability	Inconsistent translation of same word or phrase	0	1	1	0	0	5
	Not translated word which affects the meaning	0	0	1	0	0	0
	Transliteration instead of translation	0	0	0	4	1	3
Context misinterpretation	Wrong POS	0	0	1	2	0	1
	Wrong translation due to lack of informationin the previous context	0	2	1	1	1	1
Total errors		0	3	5	9	3	12

the only translation of mansion is *палата (palata / palace)*. The conjunction *дека (deka / that)* is a better choice than *да (da / that, to)* in the translation of the sentence "It was even conceivable that they watched everybody all the time." This mistake is relatively subtle.

On the other hand, the omission of the preposition *со (so / with)* in the phrase *небото беше со јасно сина боја (neboto beshe so jasno sina boja / the sky a harsh blue)* is a serious issue, as it significantly affects the meaning and clarity of the sentence.

Gemini's inconsistency involves the abbreviation of the named entity *Министерство за изобилство (Ministerstvo za izobilstvo / Ministry of Plenty)*, which should be rendered as *Минизоб (Minizob)* or *Миниизоб (Miniizob)*, rather than *Миниобил (Miniobil)*. Still, the abbreviation used is arguably acceptable, as the adjectives *изобилен (izobilen / abundant)* and *обилен (obilen / abundant)* are often used interchangeably. In its first occurrence, NLLB correctly translated the word caption as *натпис (natpis)*, but in the second instance, it used the Bulgarian form *надпис (nadpis)*. The term Thought Police appears both as *Мислова Полиција (Mislova Policija / Thinking Police)* and *Полиција на мисла (Policija na misla / The Ministry of Thought)*. The term telescreen is translated as both *телевизор (televizor / TV)* and *телескрин (teleskrin / Macedonian pronunciation of telescreen)*. The word alcove was either translated into the non-existent term *алкова (alkova)* or left untranslated, indicating a lack of familiarity with its meaning.

A common issue across most observed systems, except for both versions of GPT, was the misinterpretation of the part of speech (POS) in the source language at the beginning of the fourth sentence: It depicted simply an enormous face. Specifically, the adverb simply was confused with the adjective simple. In Macedonian, both the adverb simply and the neuter singular form of the adjective simple are rendered as *едноставно (ednostavno)*. Incorrect translations placed *едноставно* in front of the noun *лице (lice / face)*, resulting in the incorrect phrase *прикажуваше едноставно огромно лице (prikazhuvashe ednostavno ogromno lice / depicted a simple enormous face)*, instead of the correct translation *едноставно прикажуваше огромно лице (ednostavno prikazhuvashe ogromno lice / depicted simply an enormous face)*.

The case of mistranslation due to a lack of contextual understanding involves the named entity Airstrip One, which was predominantly rendered as *Писта Еден (Pista Eden)* or *Авионска писта Еден (Avionska pista Eden)*, both of which mean Runway One. Even a Macedonian human translator interpreted it as *Воздушниот коридор Еден (Vozdushniot*

koridor Eden / Air Corridor One). In reality, according to Orwell, Airstrip One refers to one of the provinces of Oceania in Orwell's 1984.

Overall, structural and consistency errors were relatively infrequent, with NLLB-600M showing the lowest level of internal coherence among the evaluated models. While some inconsistencies could be explained by lexical ambiguity or lack of context, others highlight the need for improved handling of named entities and domain-specific terminology in MT systems.

Fluency and Naturalness Fluency and naturalness errors occur when the translated text, although grammatically correct, sounds awkward, stilted, or unidiomatic in the target language. These issues often result from meaningless translations, omitted words, non-existent or untranslated units, and unnatural expressions, ultimately reducing the readability and stylistic quality of the output (Table 4).

Table 4. Distribution of fluency and naturalness errors

Fluency and naturalness		GPT-4o	GPT-5	Gemini 2.5-Flash	Google Translate	Microsoft Translator	NLLB-600M
Omissions	Missing conjunction	0	0	1	1	0	1
	Missing phrase	0	0	0	0	0	1
	Missing preposition	1	1	0	2	1	2
	Missing verb	0	0	0	0	1	1
Non-existent or untranslated units	Non-existent word	0	1	2	2	0	21
	Not translated phrase	0	1	0	0	0	2
	Not translated words	0	0	2	2	9	4
	Meaningless translation	1	0	2	2	6	10
Unnatural expressions	Too descriptive	0	0	2	2	2	1
	Misuse of conjunctions	1	1	0	0	0	0
Total errors		3	4	9	11	19	43

GPT-4o produced only one example of a meaningless translation. The phrase: *Лицето му го обликуваше во израз на тивок оптимизам (Liceto ti go oblikuvashe vo izraz na tivok optimizam)* can be back translated as His face was shaped into an expression of quiet optimism, which is lexically different from the original phrase: He had set his features into the expression of quiet optimism. Their deeper analysis shows that they describe the same expression, but they do not express the same meaning.

It seems that GPT-5 on two occasions responded with overconfidence, either by inventing the non-existent and nonsensical word *бљаболеше (bljaboleshe)*, or by omitting words from larger expressions, for example, translating *во овој час (vo ovoj chas / at this time)* instead of the full phrase *во овој час од денот (vo ovoj chas od denot / at this time of day)*.

The quarto-sized term, which refers to a book format created by folding a single sheet of paper into four leaves, is entirely obscure. GPT-4o translated it as *големина кварто (golemina kvarto / size quarto)*; GPT-5 rendered it as *квартов формат (kvartov format / quarto format)*; Gemini used *со големина на четвртина (so golemina na chetvrtina / with a size of a quarter)*; and Google Translate offered *големина на четврт парче (golemina na chetvrt parche / size of a quarter piece)*. Each translation is partially true in its own way, and the last two can be easily visualized by the reader.

Microsoft Translator and NLLB-600M chose not to translate the term at all. This approach is arguably better than inventing incorrect or non-existent words. In relation to this phenomenon, it is worth noting that both Gemini and Google Translate generated two non-existent words each; the most amusing among them is *холиоза (holioza)*, used as a translation of the noun hallway. In contrast, four systems translated hallway correctly as *ходник (hodnik)*, and one as *коридор (koridor)*.

Microsoft Translator decided not to translate some of the Newspeak-named entities, for example Ministry of Truth and Victory Gin and all the abbreviations: Minitrue, Mini-pax, Miniluv, and Miniplenty. This system generated six meaningless translations, the most interesting are the translations of the phrases picked out on its white face, which is *избрани на неговото бело лице (izbrani na negovoto belo lice / selected on his white face)* and He had set his features into the expression, which was transformed to *Тој ги постави своите карактеристики во израз (Toj gi postavi svoite karakteristiki vo izraz / He set his characteristics in the expressions)*. While pick up on is a commonly used phrasal verb meaning to notice or respond to something subtle, selected on is only correct when followed by a specific criterion, such as selected on the basis of merit. In the second example, although back translation of the phrase is almost identical to the original, the Macedonian phrase has no logical interpretation.

Overly descriptive translations are exemplified by the adjective black-mustachio'd, which has already been discussed as an example of incorrect phrasing. Another example is the phrase day in April, which can be translated literally as *ден во април (den vo april)*, a solution used by Gemini, Microsoft Translator, and NLLB-600M, although the more natural expression is *априлски ден (aprilski den)*, as rendered by both GPT models and Google Translate.

NLLB-600M has proven to be a brilliant inventor of non-existent words, most of which are completely incomprehensible, such as *долборот (dolborot)*, *патрула (patrula)*; *шпикувајќи (shpikuvajkji)* or *контејната (kontejnata)*. They contributed to the creation of meaningless translations.

In addition to non-existent words, this MT also invented completely incomprehensible translations. For example, the sentence Winston made for the stairs. was translated as *Уинстон го направи тоа за скалите. (Uinston go napravi toa za skalite. / Winston did it for the stairs.)*. However, the most notable example is the phrase *тревата од вила се држеше над купки од рушеви (trevata od vila se drzeshe and kupki od rushevi)*. Although willow-herb is correctly translated as *врбовка (vrbovka)*, the use of *тревата (trevata / the grass)*, even *тревата од вила (trevata od vila / the fairy grass)*, at least retains some sense. The aorist form of the reflexive verb *се држи (se drzhi)* is *се држеше (se drzheshe)* and in the context of herbs, the English verbs hold, support, keep, stick, and maintain can be used, none of which corresponds closely to the source verb straggle. The phrase *купки од рушеви* contains the non-existent word *рушеви (rushevi)*, which should probably be *рушевини (rushevini / rubble)*, a perfect match. However, the plural noun *купки (kupki / bath)* has nothing in common with the original word heaps. If you cannot guess the source English phrase, here it is: the willow-herb straggled over the heaps of rubble.

Fluency and naturalness remain challenging aspects for machine translation systems, with notable variability in performance among different models. While some LLMs, such as GPT-4o, demonstrate strong capabilities with only minor issues, others like NLLB-

600M frequently produce nonsensical or invented words that undermine overall translation quality.

The nuanced handling of idiomatic expressions, specialized terms, and stylistic subtleties often reveals the limits of current technology. Nevertheless, the relatively low overall error rates highlight encouraging progress toward producing translations that are both accurate and readable.

Comparative Analysis of Manual Error Annotations To assess similarities in error behavior across systems, this section conducts a correlation-based comparison of manual error annotations. The analysis aims to reveal whether different MT systems exhibit shared or distinct error profiles when translating from English to Macedonian.

Table 5 presents the distribution of translation errors across four error categories for three LLM-based systems (GPT-4o, GPT-5, and Gemini 2.5 Flash) and three conventional NMT systems (Google Translate, Microsoft Translator, and NLLB-600M). In total, 526 errors were identified across all systems. The share of errors of LMS- based systems is only 26.74%. Their results are incomparably better than the NMT system.

Table 5. Distribution of translation errors

Linguistic evaluation of MT systems	GPT-4o	GPT-5	Gemini 2.5-Flash	All LLMs	Google Translate	Microsoft Translator	NLLB-600M	All NMT systems
Linguistic inaccuracy	19	14	11	44	23	90	91	204
Stylistic errors	40	35	48	123	65	44	116	225
Lexical and semantic errors	0	3	6	9	9	3	12	24
Fluency and naturalness	3	4	9	16	11	19	43	73
Total inaccuracies	62	56	74	192	108	156	262	526

Across all systems, stylistic errors constitute the most frequent error type (348 instances overall), affecting both LLMs and NMT systems. However, stylistic issues are particularly prominent in NMT output, with 225 errors compared to 123 errors for LLMs. A similar pattern is observed for linguistic inaccuracies, where NMT systems collectively produced 204 errors, compared to only 44 errors for all LLMs combined. This suggests that LLM-based systems handle grammatical and lexical correctness more robustly in English–Macedonian translation.

In contrast, Lexical and semantic errors are relatively rare across all systems, though they appear slightly more often in NMT systems than in LLMs. Errors related to fluency and naturalness show the clearest divergence between paradigms: NMT systems exhibit 73 such errors, whereas LLMs register only 16, reinforcing the observation that LLM-generated translations tend to be more fluent and natural. Overall, the table highlights a consistent trend in which LLM-based systems outperform conventional NMT systems across all evaluated error categories, both in total error count and in qualitative dimensions related to fluency and stylistic adequacy.

The correlation matrix (Table 6.) reveals very strong internal consistency within the LLM group. GPT-4o, GPT-5, and Gemini 2.5 Flash show near-perfect correlations with one another, ranging from 0.93 between GPT-4o and Gemini 2.5Flash to 0.99 between both

Table 6. Correlation matrix

Mutual correlation between all MT systems	GPT-4o	GPT-5	Gemini 2.5-Flash	All LLMs	Google Translate	Microsoft Translator	NLLB-600M	All NMT systems
GPT-4o	1.00	0.99	0.93	0.98	0.97	0.53	0.95	0.91
GPT-5	0.99	1.00	0.97	1.00	0.99	0.41	0.90	0.85
Gemini 2.5Flash	0.93	0.97	1.00	0.98	0.99	0.18	0.79	0.71
All LLMs	0.98	1.00	0.98	1.00	1.00	0.37	0.89	0.83
Google Translate	0.97	0.99	0.99	1.00	1.00	0.32	0.86	0.80
Microsoft Translator	0.53	0.41	0.18	0.37	0.32	1.00	0.74	0.82
NLLB-600M	0.95	0.90	0.79	0.89	0.86	0.74	1.00	0.99
ALL NMT systems	0.91	0.85	0.71	0.83	0.80	0.82	0.99	1.00

GPT versions. The aggregated "All LLMs" score, computed by summing the identified errors across all LLM-based MT systems, is even higher, ranging from r = 0.98 to r = 1.00. This indicates that, despite architectural and training differences, LLM-based systems exhibit highly similar error distributions across the four error categories, suggesting a shared translation behavior and error profile when translating from English to Macedonian.

A similarly strong pattern is observed within the "All NMT" group, defined as the aggregate of manually identified errors across NMT systems only, particularly between NLLB-600M and the resulting aggregated score (r = 0.99). NLLB-600M also correlates strongly with Google Translate (r = 0.86) and Microsoft Translator (r = 0.74), indicating that conventional NMT systems likewise share a broadly comparable error distribution. However, the correlations among NMT systems are generally lower and more variable than those observed among LLMs, pointing to greater heterogeneity within the NMT paradigm.

Cross-paradigm correlations between LLMs and NMT systems are mixed. Google Translate shows very high correlations with LLMs (r = 0.97–0.99), suggesting that its error profile is closer to that of LLM-based systems than to other NMT systems. In contrast, Microsoft Translator exhibits consistently weak correlations with LLMs, ranging from r= 0.18 with Gemini 2.5 Flash to r=0.53 with GPT=40. This divergence reflects systematic differences in how error types are distributed rather than overall quality alone.

4.2. Automatic Evaluation of Observed Machine Translation Systems

According to the manual annotation, Microsoft Translator and NLLB-600M performed noticeably worse than the other systems. Google Translate achieved average results, whereas LLMs were more successful, with GPT-5 outperforming all others. Similarly, the automatic metrics confirmed that the two lowest-performing MT systems also produced the weakest results (see Table 7). Some versions of COMET also identified Google Translate as weaker than the LLM-based systems, a distinction not reflected in the lexical metrics. Additionally, TER rated Gemini and Google Translate as more divergent from the reference translation compared to the GPT models.

These scores are broadly consistent with the manual evaluation, which found that Microsoft Translator and NLLB produced the highest number of errors, with Google Translate trailing slightly behind the remaining systems. However, automatic metrics do not fully capture the qualitative differences within these groups. We also calculated COMETkiwi scores for the human reference translation. Its mean score was 0.7889, lower than most of

Table 7. Mean metric scores for each system.

MT systems / Evaluation metrics	BLEU	chrF	METEOR	TER	BERT score (F1)	COMET	COMET kiwi	XCOMET-XL
GPT-4o	27.28	56.30	0.4953	53.36	0.8669	0.8503	0.8168	0.8611
GPT-5	27.47	57.65	0.5006	51.65	0.8693	0.8488	0.8179	0.8670
Gemini2.5Flash	27.30	56.74	0.5060	54.04	0.869	0.8448	0.8154	0.8501
Google Translate	27.58	56.41	0.5004	54.01	0.8678	0.8322	0.8154	0.8363
Microsoft Translator	22.72	52.34	0.4429	57.64	0.8551	0.8207	0.8054	0.8379
NLLB-600M	17.83	47.54	0.3844	66.44	0.8376	0.7602	0.7722	0.7288

the machine translation outputs, highlighting that automated metrics do not always rank translations with more human-like features at the top.

The distribution of per-sentence scores is shown in Figures 1 and 2. The weaker systems exhibit particularly wide score ranges when evaluated with the COMET models, whereas their ranges remain relatively narrow when assessed using lexical metrics. In contrast, the lexical metrics tend to assign a broader range of scores to translations produced by Google Translate and the LLM systems. Notably, these metrics also assigned very low scores to some systems that were identified through human evaluation as having the fewest errors. This suggests that lexical metrics may underestimate the quality of more fluent or freer translations that, while correct, do not closely resemble the reference.

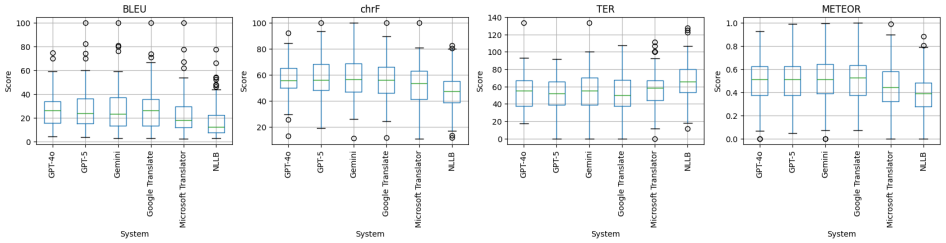


Fig. 1. Distribution of scores for lexical metrics

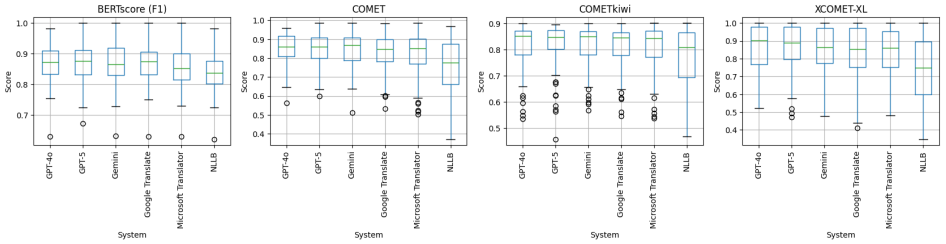


Fig. 2. Distribution of scores for embedding-based and learned metrics

To get a better picture of how each metric compares to the manual evaluation, we computed the correlation between them. We quantified the manual evaluation with a formula inspired by the MQM and ESA scoring systems. Errors were given weights depending on their severity, -1 for minor errors and -5 for major errors. Accuracy errors that change the meaning of the sentence were considered more major than grammatical, punctuation, or stylistic errors. For example, errors from the categories Literalism and Morphological disagreement were given a weight of -1, while missing words, transliteration instead of translation and including non-existent words were given a weight of -5. Correct translations were assigned a weight of 0. The score for each sentence was computed as the weighted sum of errors in that sentence.

The correlation between the manual evaluation and automated metrics is given in Table 8. The first two columns show the sentence-level Spearman correlation between the manual evaluation and each metric, and the second two columns show the system-level Spearman correlation. Note that TER is negatively correlated because a lower TER score indicates better translation quality.

Table 8. Correlation with manual evaluation and pairwise accuracy

	Sentence-level		System-level	
	Correlation with manual evaluation	Pairwise accuracy	Correlation with manual evaluation	Pairwise accuracy
Manual evaluation	1	1	1	1
COMET	0.41	0.528	0.943	0.93
chrF	0.407	0.53	1	1
BERTscore	0.4	0.515	0.943	0.93
METEOR	0.372	0.485	0.486	0.67
XCOMET-XL	0.372	0.52	0.943	0.93
BLEU	0.347	0.41	0.829	0.87
TER	-0.341	0.448	-0.943	0.93
COMETkiwi	0.233	0.45	1	1

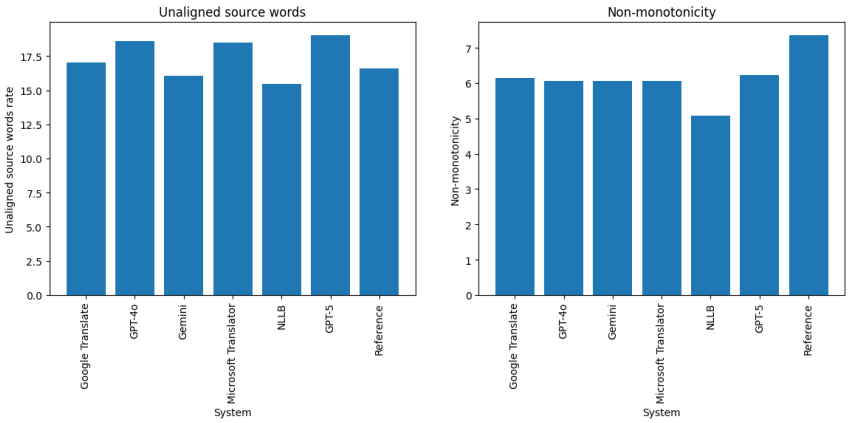
While the system-level correlations are strong, the sentence-level correlations are low to moderate. This is expected, since the human evaluation metric can be prone to fluctuations and is more sensitive to errors that might not be captured by the automated metrics. Among the metrics, COMET and chrF show the high correlations on both the sentence and system level, while BLEU and METEOR lag behind. COMETkiwi and XCOMET-XL, as well as TER (in terms of magnitude) show strong correlations on the system level but are ranked in the middle of the pack or lower on the sentence level. Overall, these results show that the included metrics, especially the highly correlated ones, can adequately rank the systems' translation quality, but highlight the fact that a gap between them and human judgment exists on the sentence level.

We also computed the pairwise accuracy using the human evaluation as the gold standard. A pair of systems is considered to be accurately ranked by a metric if that metric ranks them in the same order as the human evaluator, using the same scoring scheme as above. The pairwise accuracy results follow the trends of the Spearman correlation, with

the same set of metrics generally achieving the best results. One exception is XCOMET-XL, which is ranked low by the Spearman correlation on the sentence level, but is among the best when evaluated through pairwise accuracy.

Further evaluation including multiple human evaluators, which might stabilize the fluctuations in the human evaluation on the sentence level is left for future work.

Inspired by Raunak et al. (2023) [28], we also computed the number of unaligned source words and the degree of non-monotonicity for all MT systems and compared these to the corresponding scores for the reference translation. Additionally, we examined the number of unaligned source words that were not stopwords as defined in the nltk module. To evaluate the aligner model, we used a sample of 60 sentences, 10 from each MT system. While the model made occasional errors, these were rare and consistent across systems. Most errors involved multi-word expressions, where each word in the expression was aligned with all words in the target language. There were almost no instances where a word that should have been aligned was missed.



(a) Unaligned source word rate **(b)** Non-monotonicity score

Fig. 3. Comparison between machine and human translation performance

Figure 3a presents the unaligned source word rate for all MT systems alongside the human reference translation. In the sample analyzed, this metric appears to be unrelated to overall translation quality. Human reference contains fewer unaligned source words than most machine translations. While the relative ordering of systems remains largely the same when stopwords are excluded, the top-ranked system changes, ranking first in this case, compared to second when stopwords are included as potential unaligned words. The non-monotonicity metric, however, reveals that the reference translation has a significantly freer word order compared to the machine translations (Figure 3b). Among the MT systems, there were no substantial differences in scores, except for NLLB-600M, which had by far the lowest non-monotonicity score.

4.3. Comparison of the Similarity of Machine Translations

The last experiment examined sentence-level similarities between the different MT systems. For this analysis, we used cosine similarity based on multilingual MPNet embeddings [56], Jaccard similarity, and Levenshtein distance. We calculated the mean similarity between all pairs of systems and analyzed the distribution of each similarity metric for pairs grouped by translation correctness, as determined by the manual evaluation. Figure 4, presents the mean cosine similarity. As expected, cosine similarity is high across all systems, given that they are translations of the same source text. Nonetheless, certain systems, most notably NLLB-600M, exhibit considerably lower similarity to the translations produced by the other systems.

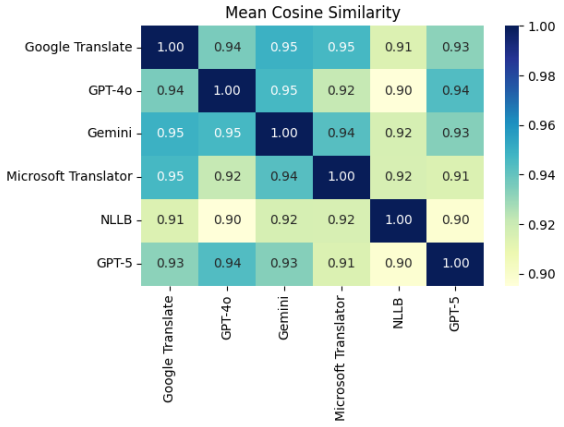
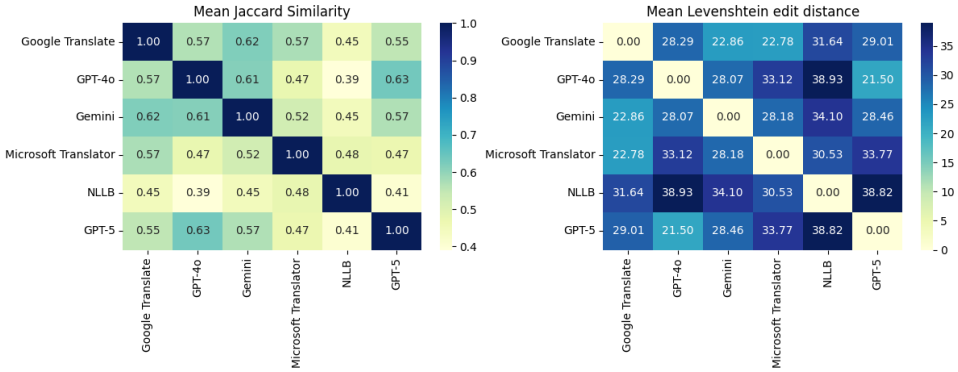


Fig. 4. Mean cosine similarity

The Jaccard similarity and Levenshtein distance, which focus more on the exact words used in each translated sentence, may provide a clearer indication of which systems tend to produce similar translations. The mean scores for these metrics are shown in Figure 5. Both metrics highlight that NLLB-600M stands out as particularly dissimilar from the other systems. At the same time, they reveal a strong similarity between the GPT systems, as expected. Interestingly, the Google-developed systems Google Translate and Gemini also produce notably similar translations, despite their differing architectures.

To determine whether correctly translated sentences tend to be more similar to each other and how different incorrect translations are, we analyzed the distribution of sentence-level similarities between all sentence pairs, regardless of the system, based on their correctness. Sentences with no identified errors in the manual evaluation were labeled as correct, while all others were marked as incorrect. The results are shown in Figure 6.



(a) Jaccard similarity **(b)** Levenshtein distance

Fig. 5. Comparison of exact wording

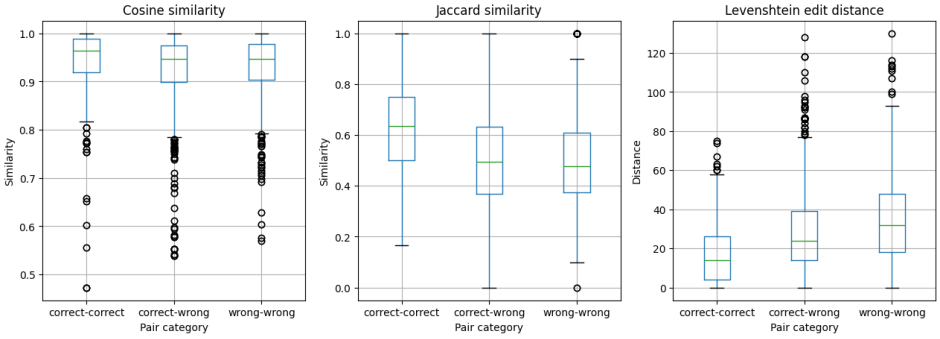


Fig. 6. Distribution of scores for embedding-based and learned metrics

As expected, pairs of correctly translated sentences exhibit higher mean similarity across all metrics, with generally smaller score ranges. For cosine and Jaccard similarity, there is no significant difference between the scores of correct–wrong and wrong–wrong sentence pairs. Both categories have similar means, though correct–wrong pairs tend to show slightly higher variance. A notable difference appears in the Levenshtein distance: wrong–wrong pairs receive higher scores, indicating they are less similar to each other than correct–wrong pairs. Additionally, the variance in Levenshtein scores is considerably higher for wrong–wrong pairs compared to correct–wrong pairs.

We preformed paired Mann-Whitney U test on these results to check their statistical significance. They shows that the difference between the correct-correct and the other groups is significant ($p \ll 0.0001$), while the differences between the correct-wrong and wrong-wrong groups sometimes aren't significant, particularly in the case of cosine similarity and Jaccard distance.

Overall, these findings highlight that sentence-level similarity metrics can reflect patterns in translation quality and system behavior, though their sensitivity varies depending

on the type of metric used. They show that wrong translations tend to diverge more than correct ones. While embedding-based and lexical similarity measures confirm expected trends, such as greater consistency among accurately translated sentences, only certain metrics, like Levenshtein distance, capture more nuanced differences between error types. These insights further underscore the importance of combining multiple evaluation approaches when analysing machine translation output.

5. Conclusions

This study provides a comprehensive linguistic evaluation of six prominent machine translation systems applied to the translation of George Orwell's 1984 into Macedonian. The three LLMs (GPT-4o, GPT-5, and Gemini 2.5 Flash) consistently outperformed traditional MT systems (Google Translate, Microsoft Translator) and especially the social media NLLB-600M, which produced the weakest output in both human and automatic evaluations.

Manual annotation revealed an average of 1.2 errors per sentence, with stylistic (48.47%) and linguistic (34.54%) errors being the most frequent. While LLMs demonstrated superior fluency and syntactic control, evaluation metrics, particularly lexical ones, sometimes penalized these systems for producing accurate but freer translations.

Echoing our initial premise inspired by Tolstoy, while accurate translations often converge, the diversity in error patterns reveals each system's unique translation behavior. Our findings show that accurate translations tend to converge across systems, exhibiting higher consistency and lower variance, especially when assessed using Levenshtein distance, while inaccurate translations display far more variation in both form and structure.

Sentence similarity metrics confirmed these patterns, reinforcing the idea that accurate outputs share common traits, whereas errors are more system-specific. Notably, this discrepancy highlights a key limitation of existing automatic metrics, which still tend to favor surface-level similarity over deeper linguistic accuracy or stylistic nuance. While metrics such as cosine similarity and Jaccard index revealed overall alignment trends, particularly among systems with similar architectures, only Levenshtein distance was sensitive enough to capture nuanced differences between error types.

What we have recognized within our research is that the performance of machine translation systems when applied to Orwell's 1984 is relatively satisfactory, especially when compared with the high-quality translations these systems produced in professional and technical domains. As teachers of several courses delivered in parallel in Macedonian and English, we frequently rely on such systems to translate slides, examination tasks and advertisements. Apart from errors arising from the absence of standardized terminological dictionaries, the translations are almost impeccable. The development of such dictionaries, once common practice but now largely abandoned, would enable experts to rely more confidently on machine translation, particularly for translations from major languages. By contrast, machine-generated translations found on Wikipedia are often of questionable quality, largely because they have not undergone editorial revision. What is particularly unfortunate is that these unedited and often unreliable translations are used as training data for artificial intelligence systems, a practice that risks entrenching errors, amplifying stylistic and semantic distortions, and giving rise to serious long-term problems in the development and deployment of machine translation technologies.

To improve the performance of machine translation systems, particularly those targeting low-resource languages, such as Macedonian, several key steps are necessary.

1. Expand and enrich training data
   There is a pressing need for high-quality, annotated training data in Macedonian and similar low-resource languages. Data collection efforts should prioritize stylistically diverse and domain-rich corpora, including literary texts, formal documents, and colloquial speech. The inclusion of varied linguistic registers ensures broader system generalizability and cultural relevance. The primary problem for the Macedonian language is the lack of a national corpus, which has been initiated for years by several official institutions, but unfortunately, it has never been created. There are collections of texts on the web, most of them are copyright protected, so they are unusable for training LLMs. A coordinated, state funded initiative to establish an open, legally compliant national corpus would therefore be a crucial step toward enabling effective and equitable development of language technologies for Macedonian.
2. Improve access to bilingual and interpretative resources
   Greater access to paper-based bilingual and interpretative dictionaries is essential. In the case of Macedonian, many of these valuable resources remain protected by copyright and are underutilized in digital contexts. Digitizing and licensing these materials for research and development could significantly enhance language visibility and resource accessibility. For Macedonian, this particularly applies to dictionaries by private publishers that do not allow the digitization of their dictionaries. Interestingly, there are two digital dictionaries of the Macedonian language: an interpretative dictionary published by the government (https://makedonski.gov.mk/) and a private one (http://drmj.eu/). The second one illegally digitized the existing bilingual dictionaries from English, Albanian and Turkish. Unfortunately, both dictionaries are not accessible for automated data extraction, so they have no impact on LLMs. Some digital repositories already exist, most of them are not in a fully machine-readable format, thus they cannot be used as training data for machine learning models.
3. Modernize and digitize phraseological dictionaries
   Existing phraseological dictionaries should be updated and digitized to reflect contemporary usage. This includes the incorporation of newly coined expressions and frequently used idioms, which are often underrepresented in traditional MT training data yet crucial for fluent, culturally accurate translations. The two official phraseological dictionaries of the Macedonian language were published in 2003 [61] and in 2008 [62]. Both were sold out a long time ago and they are not available in electronic format. In the meantime, the language has evolved and thousands of words and phrases have entered it, which are visible in the interpretive dictionary.
4. Advance evaluation methods
   MT evaluation must go beyond surface-level similarity. A more diverse set of evaluation metrics is needed to capture the nuanced behavior of translation systems. Refining current methods or incorporating reference-free approaches, such as quality estimation models, and human-in-the-loop evaluation protocols will lead to more accurate and reliable assessments of translation quality. This applies to all languages, not only to low-resourced ones. Our proposal to introduce mutual similarity can be the initial basis for establishing some new metric. For now, it is primarily assessed by existing

metrics that do not relate to mass translation. In our next research, we are thinking about establishing some combined metric that would unite both directions.

5. Link error patterns to linguistic phenomena
   Future research should investigate whether observed translation divergences correspond to specific linguistic phenomena or error taxonomies. Establishing such links can offer deeper insights into model limitations and guide more linguistically informed improvements. In our work, we have already pointed out four types and causes of translation errors: linguistic inaccuracies, stylistic errors, translation errors, and fluency and naturalness. They are widely accepted and are applied for manual translation evaluation. What is specific is that most of these errors are related to the grammar and spelling of the Macedonian language, so it is not possible to map them into other languages.
   The joint effort of researchers of mutually similar languages would contribute to the establishment of common patterns of errors and the discovery of interesting linguistic phenomena. For example, in the Macedonian language, but probably in other languages as well, there are translation errors that are calques from English, although there are original words for the same concepts that, unfortunately, under the pressure of the mass use of new words, become extinct over time.
6. Enhance low-resource support in open-source models
   Developers of open-source models, such as NLLB-600M, should prioritize the robust handling of low-resource languages. This involves fine-tuning on carefully curated datasets and minimizing common issues like hallucinations or the generation of non-existent words. Special attention should be given to ensuring linguistic fidelity, especially in morphologically rich and syntactically complex languages. Unfortunately, this requires synchronized action between linguists, who often lack technical training, and computer scientists, who may have limited expertise in the grammatical structure of the language. A true symbiosis between these two expert profiles would yield excellent results, provided they develop a shared conceptual and terminological framework. The lack of common framework is probably the main reason why the initiative for interdisciplinary studies in computational linguistics, which was started several times, was never established in our country.
   Achieving equal translation quality in all languages requires a coordinated, multidisciplinary effort. Stakeholders, primarily researchers, developers, linguists, and policy makers, should work together to: support data collection and annotation initiatives; supported community-led and open access projects; enabled responsible digitization and licensing of key language resources; promoted evaluation frameworks that are inclusive of low-resource contexts, balancing automated metrics with human expertise; and encouraged language diversity in machine translation benchmarks and collaborative tasks, thereby encouraging innovation beyond high-resource language pairs.

6. Limitations and Future Work

Despite the depth of the manual analysis and the expertise involved, this study has the following three limitations.

The error annotation was performed by a single primary annotator, and no formal inter-annotator agreement metrics were computed. Future studies will explore simplifying

or consolidating low-frequency error categories and expanding the corpus to other genres and language pairs, thereby enabling more robust quantitative analysis and facilitating broader comparison with established evaluation frameworks.

A further limitation concerns the size and scope of the evaluation dataset. The manual analysis is based on 100 sentences, which reflects a common trade-off in fine-grained MT error analysis between analytical depth and dataset scale. While this sample cannot be claimed to be fully representative of all English–Macedonian translation phenomena, it was selected to include a range of syntactic structures, lexical choices, and stylistic features characteristic of literary prose. Consequently, the findings should be interpreted as indicative of systematic tendencies in system behavior rather than as exhaustive performance measurements. Future work will aim to expand the dataset and validate the observed trends across additional texts and genres.

An additional consideration relates to the choice of Orwell's "1984" as the source text for evaluation and the potential risk of prior exposure by MT systems, particularly LLM-based models. Although the Macedonian translation has existed for many years, its electronic version is not publicly available and, to our knowledge, was not included in MT training corpora. The digital version used here is part of MULTEXT-East and was created for linguistic research. While an alternative translation is now partially accessible online, it differs substantially from our reference and was not used in this study. Moreover, it is subject to copyright protection by both the publisher and the hosting platform, which makes its inclusion in large-scale training datasets less likely. Future work will further enhance robustness by including additional control texts of similar literary and stylistic complexity, as well as newly produced or unpublished reference translations.

Acknowledgement. This work is partially financed by the Ministry of Education and Science of the Republic of North Macedonia through the project "Utilising AI and National Large Language Models to Advance Macedonian Language Capabilities".

References

1. Al-Awawdeh, N.: Translation between creativity and reproducing an equivalent original text. Psychology and Education Journal 58(1), 2559–2564 (2021)
2. Levỳ, J.: The art of translation. John Benjamins Publishing Company (2011)
3. House, J.: Translation quality assessment: Past and present. Routledge (2014)
4. Lefevere, A.: Translation/history/culture: A sourcebook. Routledge (2002)
5. Koby, G.S., Fields, P., Hague, D.R., Lommel, A., Melby, A.: Defining translation quality. Tradumàtica (12), 0413–420 (2014)
6. Scott, C.: Literary translation and the rediscovery of reading. Cambridge University Press (2012)
7. Singh, S.P., Kumar, A., Darbari, H., Singh, L., Rastogi, A., Jain, S.: Machine translation using deep learning: An overview. In: 2017 international conference on computer, communications and electronics (comptelix). pp. 162–167. IEEE (2017)
8. Ide, N., Véronis, J.: MULTEXT: Multilingual text tools and corpora. In: COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics (1994)
9. Erjavec, T.: Multext-east. In: Handbook of linguistic annotation, pp. 441–462. Springer (2017)
10. Bonchanoski, M., Zdravkova, K.: Machine Learning-based approach to automatic POS tagging of Macedonian language. In: Proceedings of the 8th Balkan Conference in Informatics. pp. 1–8 (2017)

11. Bonchanoski, M., Zdravkova, K.: Learning syntactic tagging of Macedonian language. Computer Science and Information Systems 15(3), 799–820 (2018)
12. Ljubešić, N., Zdravkova, K., Stojanoska, S., Erjavec, T., Krsnik, L.: The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Macedonian 1.1 (2021)
13. Newmark, P.: A textbook of translation, vol. 66. Prentice Hall New York (1988)
14. Stapleton, P., Kin, B.L.K.: Assessing the accuracy and teachers' impressions of Google Translate: A study of primary L2 writers in Hong Kong. English for Specific Purposes (2019), https://api.semanticscholar.org/CorpusID:201394427
15. Reinhart, C.M., Rogoff, K.S.: Is the 2007 US sub-prime financial crisis so different? An international historical comparison. American Economic Review 98(2), 339–344 (2008)
16. Marie, B.: An automatic evaluation of the WMT22 general machine translation task. arXiv preprint arXiv:2209.14172 (2022)
17. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the tenth workshop on statistical machine translation. pp. 392–395 (2015)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
19. Rei, R., De Souza, J.G., Alves, D., Zerva, C., Farinha, A.C., Glushkova, T., Lavie, A., Coheur, L., Martins, A.F.: COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 578–585 (2022)
20. Koehn, P.: Statistical Significance Tests for Machine Translation Evaluation. In: Lin, D., Wu, D. (eds.) Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. pp. 388–395. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-3250/
21. Kocmi, T., Zouhar, V., Federmann, C., Post, M.: Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1999–2014. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024), https://aclanthology.org/2024.acl-long.110/
22. Jiao, W., Wang, W., Huang, J.t., Wang, X., Shi, S., Tu, Z.: Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv preprint arXiv:2301.08745 (2023)
23. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. pp. 223–231 (2006)
24. Van Egdom, G.W., Kosters, O., Declercq, C.: The riddle of (literary) machine translation quality. Revista Tradumàtica: Traduccio i Tecnologies de la Informacio i la Comunicacio (21), 129–159 (2023)
25. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y.J., Afify, M., Awadalla, H.H.: How good are gpt models at machine translation? A comprehensive evaluation. arXiv preprint arXiv:2302.09210 (2023)
26. Sizov, F., España-Bonet, C., Van Genabith, J., Xie, R., Dutta Chowdhury, K.: Analysing Translation Artifacts: A Comparative Study of LLMs, NMTs, and Human Translations. In: Haddow, B., Kocmi, T., Koehn, P., Monz, C. (eds.) Proceedings of the Ninth Conference on Machine Translation. pp. 1183–1199. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024), https://aclanthology.org/2024.wmt-1.116/
27. Luo, J., Cherry, C., Foster, G.: To Diverge or Not to Diverge: A Morphosyntactic Perspective on Machine Translation vs Human Translation. Transactions of the Association for Computational Linguistics 12, 355–371 (04 2024), https://doi.org/10.1162/tacl_a_00645
28. Raunak, V., Menezes, A., Post, M., Hassan, H.: Do GPTs Produce Less Literal Translations? In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 1041–

1050. Association for Computational Linguistics, Toronto, Canada (Jul 2023), https://aclanthology.org/2023.acl-short.90/
29. Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., et al.: Preliminary WMT24 ranking of general MT systems and LLMs. arXiv preprint arXiv:2407.19884 (2024)
30. Lommel, A.R., Burchardt, A., Uszkoreit, H.: Multidimensional quality metrics: a flexible system for assessing translation quality. In: Proceedings of Translating and the Computer 35. Aslib, London, UK (Nov 28-29 2013), https://aclanthology.org/2013.tc-1.6/
31. Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., Macherey, W.: Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. Transactions of the Association for Computational Linguistics 9, 1460–1474 (12 2021), https://doi.org/10.1162/tacl_a_00437
32. Klubička, F., Toral, A., Sánchez-Cartagena, V.M.: Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. Machine Translation 32(3), 195–215 (2018)
33. Kocmi, T., Zouhar, V., Avramidis, E., Grundkiewicz, R., Karpinska, M., Popović, M., Sachan, M., Shmatova, M.: Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. In: Haddow, B., Kocmi, T., Koehn, P., Monz, C. (eds.) Proceedings of the Ninth Conference on Machine Translation. pp. 1440–1453. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024), https://aclanthology.org/2024.wmt-1.131/
34. Karpinska, M., Iyyer, M.: Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist. In: Koehn, P., Haddow, B., Kocmi, T., Monz, C. (eds.) Proceedings of the Eighth Conference on Machine Translation. pp. 419–451. Association for Computational Linguistics, Singapore (Dec 2023), https://aclanthology.org/2023.wmt-1.41/
35. Manakhimova, S., Macketanz, V., Avramidis, E., Lapshinova-Koltunski, E., Bagdasarov, S., Möller, S.: Investigating the Linguistic Performance of Large Language Models in Machine Translation. In: Haddow, B., Kocmi, T., Koehn, P., Monz, C. (eds.) Proceedings of the Ninth Conference on Machine Translation. pp. 355–371. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024), https://aclanthology.org/2024.wmt-1.28/
36. Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popović, M., Shmatova, M., Steingrímsson, S., Zouhar, V.: Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In: Haddow, B., Kocmi, T., Koehn, P., Monz, C. (eds.) Proceedings of the Ninth Conference on Machine Translation. pp. 1–46. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024), https://aclanthology.org/2024.wmt-1.1/
37. Costa-Jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al.: No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672 (2022)
38. Polio, C.G.: Measures of linguistic accuracy in second language writing research. Language learning 47(1), 101–143 (1997)
39. Corbett, G.G.: Morphology and agreement. The handbook of morphology pp. 191–205 (2017)
40. Benson, S., DeKeyser, R.: Effects of written corrective feedback and language aptitude on verb tense accuracy. Language Teaching Research 23(6), 702–726 (2019)
41. Friedman, V.A.: Macedonian. Lincom Europa Munich (2002)
42. Steedman, M.: The syntactic process. MIT press (2001)
43. Harris, R.A.: Writing with clarity and style: A guide to rhetorical devices for contemporary writers. Routledge (2017)

44. Chironova, I.I.: Literalism in translation: Evil to be avoided or unavoidable reality. Journal of Translation and Interpretation 7, 1–28 (2014)
45. Fitria, T.N.: Performance of google translate, microsoft translator, and deepL translator: Error analysis of translation result. Al-Lisan: Jurnal Bahasa (e-Journal) 8(2), 115–138 (2023)
46. Bruton, A.: Vocabulary learning from dictionary referencing and language feedback in EFL translational writing. Language Teaching Research 11(4), 413–431 (2007)
47. Popović, M.: Error classification and analysis for machine translation quality assessment. In: Translation quality assessment: From principles to practice, pp. 129–158. Springer (2018)
48. Rahmatillah, K.: Translation errors in the process of translation. Journal of English and Education (JEE) (2013)
49. Ping, K.: Translatability vs. untranslatability: A sociosemiotic perspective. Babel 45(4), 289–300 (1999)
50. Linlin, L.: Artificial intelligence translator DeepL translation quality control. Procedia Computer Science 247, 710–717 (2024)
51. Post, M.: A call for clarity in reporting BLEU scores. arXiv preprint arXiv:1804.08771 (2018)
52. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
53. Guerreiro, N.M., Rei, R., Stigt, D.v., Coheur, L., Colombo, P., Martins, A.F.: XCOMET: Transparent machine translation evaluation through fine-grained error detection. Transactions of the Association for Computational Linguistics 12, 979–995 (2024)
54. Rei, R., Treviso, M., Guerreiro, N.M., Zerva, C., Farinha, A.C., Maroti, C., De Souza, J.G., Glushkova, T., Alves, D.M., Lavie, A., et al.: CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. arXiv preprint arXiv:2209.06243 (2022)
55. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
56. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
57. Zdravkova, K., Kuzmanova, J.: Sonority Based Syllabification of Macedonian and Serbian. Technical editors p. 11 (2024)
58. Mirkulovska, M.: Definiteness in Macedonian with some parallels in Bulgarian. Z. Topolinjska & M. Mirkulovska, Balkan morpho-syntactic similarities a bridge for levelling differences among people. Retrieved June 20, 2018 (2005)
59. Dimova, S.: English in Macedonia. World Englishes 24(2), 187–202 (2005)
60. Tomić, O.M.: Macedonian as an Ausbau language. Pluricentric Languages. Different Norms in Different Countries. Ed. Michael Clyne, Berlin/New York. Mouton/de Gruyter pp. 437–454 (1992)
61. Shirilov, T., Dimitrovski, T.: Фразеолошки речник на македонскиот јазик [Phraseological Dictionary of the Macedonian Language]. 3 volumes, Огледало [Ogledalo], Skopje (2003–2009), iSBNs for volumes: 9989-686-22-X; 978-9989-686-02-3; 978-9989-686-13-9
62. Velkovska, S.: Македонска фразеологија со мал фразеолошки речник [Macedonian Phraseology with a Small Phraseological Dictionary]. БПТ принт [BPT Print], Skopje (2008)

Jana Kuzmanova completed her undergraduate studies at the Technical University of Munich and subsequently earned her master's degree from the University Ss. Cyril and Methodius in Skopje. Since the 2021/2022 academic year, she has been pursuing a PhD at the same institution. Her research interests include machine learning, data mining and its scientific applications, high-performance computing, and natural language processing.

Katerina Zdravkova is a full professor at the Faculty of Computer Science and Engineering at Ss. Cyril and Methodius University in Skopje. Her research and teaching activities span areas such as artificial intelligence, machine translation, computer ethics, e-learning, and information systems. She has authored numerous scientific publications and has been actively involved in international conferences and editorial work in computer science.

Received: October 20, 2025; Accepted: April 15, 2026.

Computer Science and Information Systems 23(2):861–884 https://doi.org/10.2298/CSIS251029015C

HAQCCN: A Hybrid Quantum–Classical Convolutional Network with Asymmetric Kernels for Remote Sensing Image Classification

Lianghai Chen[1,2], Yuzhen Liu[1,3], Yi Lu[1,4], Xiaoliang Wang[1,5] and Huaning Song[6]

[1] School of Computer Science and Engineering, Hunan University of Science and Technology
Xiangtan, China
Sanya Institute of Hunan University of Science and Technology
Sanya, China
[2] chenlh@mail.hnust.edu.cn
[3] yzhenliu@126.com
[4] 24010502004@mail.hnust.edu.cn
[5] fengwxl@hnust.edu.cn (corresponding author)
[6] School of Artificial Intelligence and Manufacturing,Hechi University
Hechi,Guangxi
shn_2008@163.com (corresponding author)

Abstract. Remote sensing image classification is a fundamental task for Earth observation and environmental monitoring. However, conventional convolutional neural networks (CNNs) are limited by computational capacity and struggle to efficiently process the rapidly growing volume of remote sensing data. To address this limitation, we propose HAQCCN (Hybrid Asymmetric Quantum–Classical Convolutional Network), a novel hybrid architecture that integrates quantum computation into the classical convolutional framework through asymmetric quantum convolutional circuits. In HAQCCN, the asymmetric quantum circuits enable a limited number of qubits to process more classical data while maintaining excellent feature extraction capability. Experiments conducted on the IBM Qiskit platform using the Overhead-MNIST, PatternNet, and RSI-CB256 datasets demonstrate that HAQCCN outperforms conventional CNNs and existing quantum models. Furthermore, we systematically investigate the effects of encoding schemes, the number of quantum convolutional kernels, and the number of qubits on model performance, confirming the effectiveness and scalability of the proposed method for remote sensing image classification.

Keywords: Remote Sensing Image Classification, Quantum Computing, Convolutional Neural Networks, Quantum Convolutional Network, Feature Extraction, Deep Learning.

1. Introduction

Remote sensing image classification, as a core task in the field of Earth Observation (EO) [1], holds significant importance in multiple practical applications, including environmental monitoring [2], urban planning [3], hydrological observation [4], and disaster response [5]. Along with the remarkable progress achieved in satellite and sensor technologies, the volume of multispectral and high-resolution remote sensing data

has been growing exponentially across spatial, temporal, and spectral dimensions, marking EO's full transition into the era of big data [6]. While the availability of such rich data provides a solid foundation for fine-grained analysis, it also leads to increasingly severe challenges in computational cost, storage demand, and energy consumption [7]. Convolutional neural networks (CNNs), as the most prominent deep learning technique, have demonstrated high effectiveness in remote sensing image classification [8–10]. However, their training and inference phases typically require a large number of parameters and considerable computational resources. This poses significant limitations for deployment on resource-constrained embedded systems, satellite onboard processing, or applications requiring low-latency responses, thereby restricting their practical feasibility in real-world engineering scenarios [7, 10, 11]. To address these challenges, it is imperative to explore alternative computational paradigms. Quantum computing, with its inherent potential for high-efficiency information processing, offers a promising direction to overcome the aforementioned limitations [12–14].

Quantum computing, owing to its intrinsic properties such as parallelism and quantum entanglement, has been regarded as a promising paradigm for alleviating the computational bottlenecks of classical computing [12]. It has demonstrated potential across various domains, including data processing [15, 16], data classification [17], image segmentation [18, 19], model optimization [20], computer security [21, 22] and computational acceleration [23]. Under the current era of Noisy Intermediate-Scale Quantum (NISQ) devices, Quantum Machine Learning (QML) has emerged as an active research area [24], seeking to exploit Parameterized Quantum Circuits (PQCs) or quantum feature mappings to enhance data representation and computational efficiency.Within this context, quantum computing is increasingly viewed as a potential pathway to overcome the limitations of conventional image processing techniques [12]. In particular, QML models based on PQCs or quantum annealers have exhibited unique advantages in high-dimensional feature extraction and nonlinear transformation [25, 26]. Specifically, quantum convolution—by mapping local convolutional operations onto quantum circuits—has shown promise in reducing the number of trainable parameters while enabling efficient parallel processing [27]. In this regard, several representative studies have proposed structural optimization schemes for quantum models from different perspectives. For instance, Meng et al. [28] and Wu et al. [29] respectively investigated quantum approaches for graph data processing and image analysis. The Quantum Spatial Graph Convolutional Network (QSGCN) proposed in [28] integrates PQCs into graph neural networks and provides an in-depth analysis of trainability issues, with particular attention to the barren plateau problem, thereby offering a new approach for handling non-Euclidean structured data. Meanwhile,Yang et al. [30] focused on the development of Quantum Convolutional Neural Networks (QCNNs), employing the Multi-scale Entanglement Renormalization Ansatz (MERA) to effectively reduce computational complexity. They further incorporated quantum convolutional modules into classical convolutional architectures to construct hybrid models, demonstrating their effectiveness on the MNIST dataset.

Collectively, these studies aim to improve the trainability and efficiency of quantum neural networks from different data structure perspectives. However, existing quantum image representation and encoding methods often incur significant overhead in terms of qubit count and gate operations, making them difficult to apply directly to remote sensing imagery. Moreover, most existing hybrid quantum–classical vision models are validated

only on simple, small-scale datasets and typically rely on extensive classical preprocessing [31], thereby limiting their generalization capability and end-to-end trainability in real-world remote sensing applications.

Given the aforementioned context, a key challenge lies in achieving an effective trade-off between limited quantum computational resources and the demands of large-scale remote sensing image processing, while ensuring model effectiveness. Developing a hybrid model that can operate under Noisy Intermediate-Scale Quantum (NISQ) conditions and possess practical engineering applicability remains an urgent and unresolved problem.

To address these challenges, this study proposes a Hybrid Asymmetric Quantum-Classical Convolutional Network (HAQCCN) tailored for remote sensing image classification. At the encoding stage, we introduce an asymmetric encoding scheme in which the input image is divided into multiple local patches. Each patch is projected and embedded into a limited number of qubits via controlled rotation gates. Through a specially designed encoding module, the proposed method enables the representation of data exceeding the available qubit count while maintaining model performance. This design substantially reduces qubit requirements and preserves spatial and texture information at the local scale.

During feature extraction, the parallel quantum convolutional units employ parameterized entangling layers to perform quantum-level transformations on local features. In the subsequent stages, classical aggregation and classification modules are utilized to fuse and discriminate the extracted quantum features, forming a differentiable end-to-end hybrid training pipeline that allows joint optimization between quantum and classical components.

To evaluate the effectiveness and generalization capability of the proposed AQCN, comprehensive experiments were conducted on three publicly available remote sensing datasets Overhead-MNIST [32], PatternNet [33], and RSI-CB256 [34]. These datasets encompass varying spatial resolutions and spectral characteristics, thereby reflecting the diversity of representative remote sensing scenarios. Experimental results demonstrate that, even in the presence of quantum noise, our model outperforms both classical CNN baselines and existing quantum-based approaches in classification accuracy. Furthermore, ablation studies covering encoding strategies, qubit numbers, and convolutional kernel sensitivity analyses confirm that HAQCCN exhibits strong robustness and provides a flexible performance–cost trade-off.

The main contributions of this work are summarized as follows:

1. Asymmetric Quantum Encoding Model: We propose an asymmetric quantum encoding scheme that effectively enhances the capacity of qubits to process classical information, providing a novel perspective for quantum encoding.

2. Quantum Convolutional Layer: A new quantum convolutional layer is proposed, utilizing a kernel composed of U3 rotation gates and cross-qubit CNOT gates. This structure enhances feature interaction and entanglement representation, enabling richer feature extraction with limited qubits and achieving superior performance in remote sensing image classification.

3. Systematic Analysis of Quantum Design Factors: This study investigates the impacts of quantum encoding strategies, quantum convolutional kernels, and qubit numbers on model classification performance, offering valuable insights for future research on quantum models.

The organization of this paper is as follows: Chapter 2 presents a detailed description of the proposed model's overall framework and implementation, including data preprocessing and encoding strategies, design and operational procedures of quantum convolutional kernels, measurement protocols, and the subsequent classical dense layer architecture, thereby constructing a classification model based on AQCN. Chapter 3 introduces the datasets and data processing procedures, describes the comparative models, outlines the experimental platform and configurations, and presents the experimental results. Chapter 4 provides a detailed analysis of the results from Chapter 3, discusses implications for current research, identifies limitations, and suggests potential improvements. Chapter 5 provides a summary of the paper and discusses potential directions for future research.

Through this work, we aim to provide both methodological and engineering references for the practical application of quantum computing in remote sensing image analysis, thereby promoting the development of efficient and deployable solutions for Earth Observation tasks in the context of big data.

2. Materials and Methods

This study proposes a hybrid QC-CNN model, designed to perform remote sensing image classification using supervised deep learning methods by specifically leveraging rotational encoding techniques. As illustrated in Fig. 1, the proposed QC-CNN model consists of two sequential components. The quantum circuit component efficiently extracts salient features from the input images, while the classical component subsequently performs the final classification task.

Specifically, the model comprises four main modules: (1) data processing, (2) quantum convolution, (3) measurement layer, and (4) dense layer.

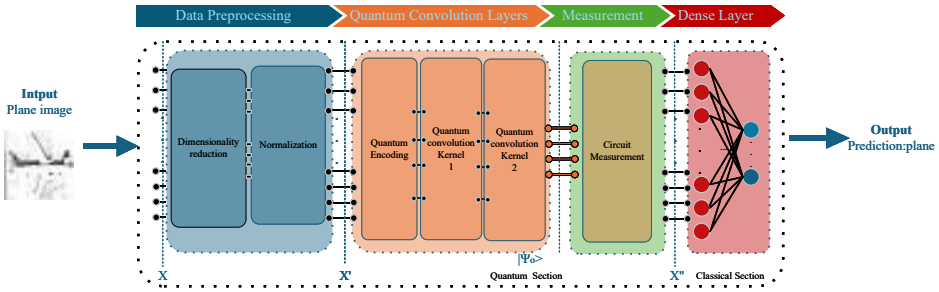


Fig. 1. Overall architecture of the HAQCCN model: 1) Input image data, perform dimensionality reduction and normalization 2) Use the quantum coding circuit to convert the image data into a quantum state, and then pass it through two quantum convolution kernels 3) The circuit measurement layer is responsible for converting the data processed by the quantum convolution kernel into classical data 4) The measured data is sent to the dense layer for classification, and the classification results of the model are output

2.1. Data Preprocessing

During the data preprocessing stage, the original image data is first subjected to dimensionality reduction to reduce computational complexity. Subsequently, the feature values are normalized to the interval $[0, \pi]$ and transformed into rotation angles $\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \ldots,$ enabling their direct use as rotation parameters for quantum RX gates. This procedure effectively performs the encoding transformation from classical data to quantum states.

2.2. Quantum Convolution Layers

Data Encoding: In this section, we will introduce the quantum coding circuit we use, the specific structure of which is shown in Fig. 2. In quantum computing, rotation gates [35] are a class of single-qubit gates that rotate quantum states around a specific axis on the Bloch sphere. we employ the RX rotation gate for encoding, which can be described as:

$$\mathrm{R_X}(\theta) = \mathrm{e}^{-\mathrm{i}\frac{\theta}{2}\mathrm{X}} \tag{1}$$

Here, X denotes the Pauli-X matrix:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{2}$$

When the RX rotation gate is applied to a qubit in the state $|0\rangle$ or $|1\rangle$, its action is as follows: Applied to $|0\rangle$:

$$R_X(\theta)|0\rangle = \cos\frac{\theta}{2}|0\rangle - i\sin\frac{\theta}{2}|1\rangle \tag{3}$$

Applied to $|1\rangle$:

$$R_X(\theta)|1\rangle = -i\sin\frac{\theta}{2}|0\rangle + \cos\frac{\theta}{2}|1\rangle \tag{4}$$

In the quantum encoding layer, CNOT gates are introduced to perform simple entanglement operations, enabling the input quantum states to the convolutional layer to represent entangled states of a convolutional block. The controlled-NOT(CNOT) gate [36], as a fundamental two-qubit operation, plays a crucial role in the transmission of information. CNOT gates can generate quantum entanglement, allowing information between different qubits to interact via quantum coherence, thereby facilitating more efficient feature mapping. Moreover, within the quantum convolutional layer, CNOT gates are employed to establish correlations between quantum channels, enabling features to propagate across qubits and enhancing the expressive capacity of the model. Compared with classical convolution, this quantum-based approach not only accelerates computation but also leverages quantum properties to improve the model's generalization capability.

The CNOT gate is a type of controlled gate that operates on two qubits: a control qubit and a target qubit. Its operational rule is as follows: if the control qubit is in the state $|0\rangle$, the target qubit remains unchanged; if the control qubit is in the state $|1\rangle$, the target qubit undergoes an X (NOT) operation, i.e., $|0\rangle \leftrightarrow |1\rangle$.

Mathematically, the action of the CNOT gate can be represented as:

$$|c, t\rangle \rightarrow |c, t \oplus c\rangle \tag{5}$$

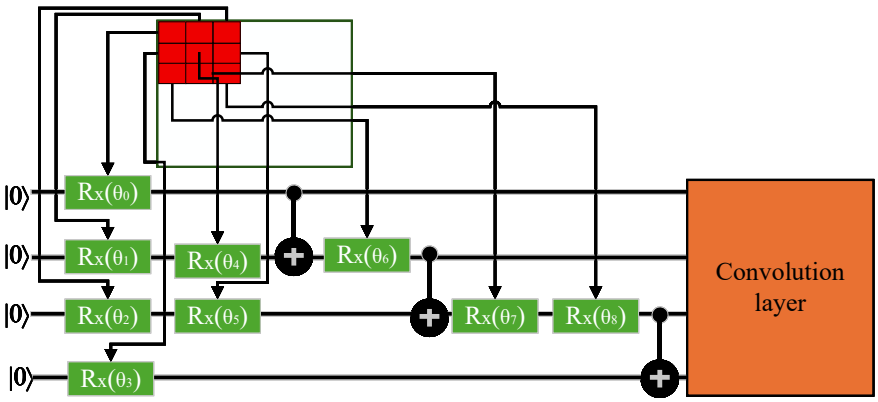


Fig. 2. Quantum coding circuit: The red block represents a 3*3 data block in classical data. Each pixel data enters the quantum circuit through the RX rotation gate (green part). The coding circuit also uses three CNOT gates (black) to entangle the data. The encoded data is finally sent to the quantum coding layer.

Here, c represents the value of the control qubit, t represents the value of the target qubit, and $\oplus$ denotes the bitwise XOR operation.

In our circuit, three CNOT gates are employed to generate entangled states among the data. These three CNOT gates connect four qubits, creating entanglement across all four qubits and enhancing the expressive capability of the quantum circuit for the input data. In our convolutional circuit, by strategically placing the RX encoding gates at key positions within the quantum circuit, four qubits are able to encode nine classical data points. In contrast, other studies on quantum convolution typically utilize a number of qubits equal to the number of convolutional data points.

Next, we provide a detailed analysis of our encoding circuit.

Let the initial state be:

$$|\psi_0\rangle = |0000\rangle \tag{6}$$

Subsequently, rotation gates are applied to the four qubits with rotation angles $\theta_0, \theta_1, \theta_2, \theta_3$. The RX gates modify the state of each qubit, resulting in the system entering the state:

$$|\psi_1\rangle = R(\theta_0)R(\theta_1)R(\theta_2)R(\theta_3)|\psi_0\rangle \tag{7}$$

The rotation operation $R_x(\theta)$ induces a corresponding rotation on each qubit, causing the qubits to enter a superposition state and thereby enabling subsequent quantum entanglement. On this basis, additional rotation gates $R_x(\theta_4)$ and $R_x(\theta_5)$ are applied to qubits q_1 and q_2, respectively, further adjusting the quantum states:

$$|\psi_2\rangle = IR(\theta_4)R(\theta_5)I|\psi_1\rangle \tag{8}$$

At this stage, the qubits are already in a more complex superposition state, laying the foundation for subsequent CNOT gate operations and quantum entanglement. In this step,

a CNOT gate is applied with q_0 as the control qubit and q_1 as the target qubit. If q_0 is in the state $|1\rangle$, the state of q_1 is flipped, thereby introducing quantum entanglement:

$$|\psi_3\rangle = \mathrm{CNOT}\,|\psi_2\rangle \tag{9}$$

Next, the rotation gate $R_x(\theta_6)$ is applied to q_1, altering its phase information and thereby affecting the coherence of the quantum state:

$$|\psi_4\rangle = \mathrm{IR}\,(\theta_6)\,II\,|\psi_3\rangle \tag{10}$$

At this stage, a CNOT gate is again applied with q_1 as the control qubit and q_2 as the target qubit. This step entangles qubits q_1 and q_2:

$$|\psi_5\rangle = \mathrm{CNOT}\,|\psi_4\rangle \tag{11}$$

The rotation gates $R_x(\theta_7)$ and $R_x(\theta_8)$ are applied to q_2 and q_3, respectively, to adjust the phases of the quantum states. Through these rotations, the coherence of the qubits can be precisely controlled:

$$\left|\psi^6\right\rangle = \mathrm{IIR}\,(\theta_7)\,I\,|\psi_5\rangle \tag{12}$$

$$|\psi_7\rangle = \mathrm{IIIR}\,(\theta_8)\,|\psi_6\rangle \tag{13}$$

Finally, a CNOT gate is applied with q_2 as the control qubit and q_3 as the target qubit. Through this operation, the quantum state of the entire system ultimately forms a complex multi-qubit entangled state:

$$|\psi_8\rangle = \mathrm{CNOT}\,|\psi_7\rangle \tag{14}$$

Through the aforementioned steps, we successfully employ a combination of rotation gates and CNOT gates to evolve the four-qubit quantum system from the initial state $|0000\rangle$ into a highly entangled quantum state $|\psi_8\rangle$, thereby completing the encoding process. The use of quantum-based encoding enables higher-dimensional representations compared with conventional machine learning methods [37, 38].

Quantum Convolution: The purpose of the quantum convolutional layer in the QC-CNN model is to perform convolution operations; this quantum layer is responsible for feature extraction and generating the corresponding feature maps.Our quantum convolution kernel is shown in Fig. 3.Within the convolutional layer, we introduce the U3 gate [39]. The U3 gate is a single-qubit gate and is considered the most general and powerful single-qubit operation. Its name derives from the fact that it is a unitary matrix parameterized by three angles. Conceptually, it can be regarded as a "universal" operation that performs arbitrary rotations on the Bloch sphere [40]. Almost any operation on a single qubit can be implemented using a single U3 gate. It has powerful expressive power, and its three learnable parameters make the quantum layer rich in adaptability.The U3 gate is represented by a 2×2 unitary matrix, and its standard matrix form is as follows:

$$U3(\theta, \phi, \lambda) = \begin{bmatrix} \cos(\theta/2) & -e^{i\lambda}\sin(\theta/2) \\ e^{i\phi}\sin(\theta/2) & e^{i(\phi+\lambda)}\cos(\theta/2) \end{bmatrix} \tag{15}$$

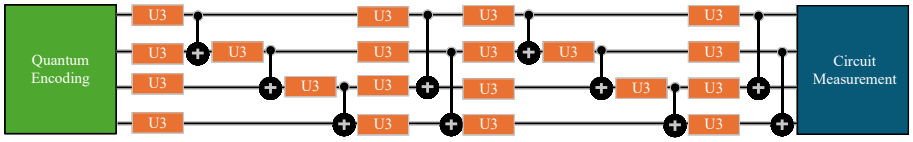


Fig. 3. Quantum convolution circuit: Processes data encoded by the quantum coding layer through two consecutive quantum convolution kernels. The orange one represents the U3 gate and the black one represents the CNOT gate. The processed data is captured and decoded by the measurement layer.

The U3 gate has three real parameters: θ (theta): Controls the rotation angle around the Y-axis. It determines the probability amplitude of the quantum state transitioning between the $|0\rangle$ and $|1\rangle$ basis states. ϕ (phi): The phase angle associated with the $|1\rangle$ state. λ (lambda): The phase angle associated with the $|0\rangle$ state. When the U3 gate is applied to a qubit:

$$\mathrm{U3}(\theta,\phi,\lambda)|0\rangle = \cos\frac{\theta}{2}|0\rangle + e^{i\phi}\sin\frac{\theta}{2}|1\rangle \tag{16}$$

$$\mathrm{U3}(\theta,\phi,\lambda)|1\rangle = -e^{i\lambda}\sin\frac{\theta}{2}|0\rangle + e^{i(\phi+\lambda)}\cos\frac{\theta}{2}|1\rangle \tag{17}$$

Our convolutional kernel, as illustrated in the Fig. 3, receives the output $|\psi_8\rangle$ from the encoding layer. We define:

$$|\psi_{\text{in}}\rangle = |\psi_8\rangle = \sum_{a,b,c,d\in\{0,1\}} \alpha_{a,b,c,d}|abcd\rangle \tag{18}$$

At the initial stage of the convolutional layer, four U3 gates are applied to the quantum states. These four U3 gates primarily perform a basic transformation on the encoded data:

$$|\psi_{\text{c1}}\rangle = \left[\mathrm{U3}_1(\theta_1,\phi_1,\lambda_1)\otimes\mathrm{U3}_2(\theta_2,\phi_2,\lambda_2)\otimes\mathrm{U3}_3(\theta_3,\phi_3,\lambda_3)\otimes\mathrm{U3}_4(\theta_4,\phi_4,\lambda_4)\right]|\psi_{\text{in}}\rangle \tag{19}$$

Subsequently, a CNOT gate is created between qubits q_0 and q_1 to enhance entanglement between them, and a U3 gate is applied to q_1 to further adjust its quantum state:

$$|\psi_{\text{c2}}\rangle = \left[\mathrm{I}\otimes\mathrm{U3}_2(\theta_6,\phi_6,\lambda_6)\otimes\mathrm{I}\otimes\mathrm{I}\right]\mathrm{CNOT}_{0,1}|\psi_{\text{c1}}\rangle \tag{20}$$

$$|\psi_{\text{c3}}\rangle = \mathrm{CNOT}_{2,3}\left[(\mathrm{I}\otimes\mathrm{I}\otimes\mathrm{U3}_2(\theta_6,\phi_6,\lambda_6)\otimes\mathrm{I})\,\mathrm{CNOT}_{1,2}|\psi_{\text{c2}}\rangle\right] \tag{21}$$

The above operations complete the propagation of quantum features. Finally, a set of U3 gates is applied to make fine adjustments to the entire quantum state after feature extraction, and two CNOT gates are used to strengthen the correlations between the qubit pairs (q_0, q_2) and (q_1, q_3):

$$|\psi_{c4}\rangle = [\mathrm{U3}_0(\theta_7, \phi_7, \lambda_7) \otimes \mathrm{U3}_1(\theta_8, \phi_8, \lambda_8) \otimes \mathrm{U3}_2(\theta_9, \phi_9, \lambda_9) \otimes \mathrm{U3}_3(\theta_{10}, \phi_{10}, \lambda_{10})] |\psi_{c3}\rangle \tag{22}$$

$$|\psi_{\mathrm{out}}\rangle = \mathrm{CNOT}_{1,3}\mathrm{CNOT}_{0,2} |\psi_{c4}\rangle \tag{23}$$

At this stage, a quantum convolutional kernel is formed. In this study, the proposed model employs two of the aforementioned convolutional kernels to enhance the feature extraction capability of the model.

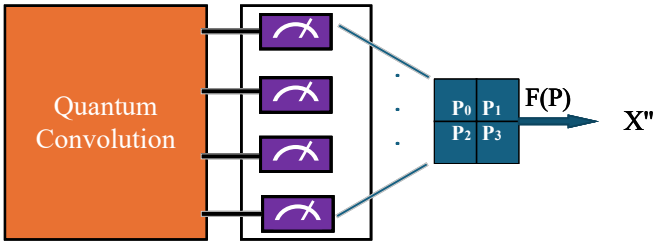


Fig. 4. In the measurement layer, we repeatedly run and measure the states of the four qubits, converting the measured value from F(P) to the value after this convolution

Measurement Layer: The quantum measurement layer, a core module of the Quantum Convolutional Neural Network (QCNN) whose structure is shown in Fig. 4, serves as a bridge between the quantum state space and the classical information space. It efficiently reduces the dimensionality of quantum data while preserving critical features. This layer transforms the quantum features extracted by the quantum convolutional layer into classically readable probability distributions through projective measurements, generalized measurements, or weak measurements, thereby supporting downstream tasks such as classification and regression. During this process, quantum measurement not only determines the efficiency of information extraction but also directly influences the expressive power and computational efficiency of the model through the combined effects of quantum state collapse and entanglement.

A measurable observable A, which is a Hermitian operator [41], can be expressed as:

$$A = \sum_i a_i P_i \tag{24}$$

Let the quantum state of the system be:

$$|\psi\rangle = \sum_i c_i |\varphi_i\rangle \tag{25}$$

The probability of obtaining the eigenvalue a_i upon measurement is given by:

$$P(a_i) = \langle\psi|P_i|\psi\rangle = |\langle\varphi_i|\psi\rangle|^2 \tag{26}$$

If the measurement outcome is a_i, the system collapses to the corresponding eigenstate:

$$|\psi'\rangle = \frac{\mathrm{P_i}|\psi\rangle}{\sqrt{\mathrm{P}\left(\mathrm{a_i}\right)}} = |\varphi_{\mathrm{i}}\rangle \tag{27}$$

After performing measurements on the four qubits, the expectation values of the four two-qubit pairs are obtained, as shown in Fig. 4. These four expectation values are computed as:

$$F(P) = \sum_{i=0}^{3} w_i P_i + b_i \tag{28}$$

and then fed into the subsequent classification layer, where w_i denotes the weights and b_i denotes the biases.

2.3. Dense Layer

After measuring the quantum layer, the results are fed into the dense layer of a classical deep learning model [42]. In a dense layer, each input node is connected to every output node, performing feature mapping through a linear transformation (i.e., a weighted sum plus a bias) followed by a nonlinear activation function. Dense layers are commonly employed to extract global features and accomplish classification tasks, particularly in the later stages of the network, where they map high-dimensional features to the target output dimensions, such as the probability distribution of class labels.

3. Experiment and Results

To evaluate the performance of our model in remote sensing image classification, we conducted comparative experiments with different models and designed additional experiments focused on our model to investigate the effects of various factors on its performance.

3.1. Data Preparation

In this study, we conducted experiments on five different Earth Observation (EO) datasets, namely Overhead-MNIST [32], PatternNet [33], and RSI-CB256 [34]. Due to the computational limitations of contemporary quantum simulators, we restricted our experiments to a subset of categories in each benchmark dataset. Additionally, we employed the Lanczos algorithm [43] to resize all labeled images in the datasets to 8×8 pixels.In the research of quantum convolutional networks, similar methods are widely employed for processing datasets [44–47].

The Overhead-MNIST [32] dataset consists of grayscale aerial images of 10 land-cover classes (e.g., ”car”, ”ship”, ”airplane”, etc.), each with a resolution of 28×28

pixels. The dataset contains 8,519 training samples and 1,065 testing samples. In our experiments, we selected six representative classes—"car", "ship", "airplane", "port", "helicopter", and "oilfield"—including all labeled images, totaling 5,098 samples for training. Additionally, 637 samples from the same classes were selected as an independent test set for performance evaluation. During the training phase, approximately 15% of the training samples were randomly allocated as a validation set, while the remaining samples were used for model training.

The PatternNet [33] dataset contains high-resolution remote sensing images from 38 categories, with approximately 800 samples per category, each of size 256×256 pixels. In our experiments, three representative classes—"coastal mansion", "parking lot", and "swimming pool"—were selected, and all corresponding samples were used. To evaluate the model, the samples were randomly divided into three non-overlapping sets: 70% for training, 15% for validation, and 15% for testing.

RSI-CB256 [34] is a global-scale remote sensing image dataset comprising 35 categories with over 24,000 images. In this study, five representative classes—"dry farm", "mangrove", "residential area", "snow mountain", and "storage room"—were selected for experiments, using all available samples in each class. All images have a resolution of 256×256 pixels. For experimental purposes, the dataset was randomly segmented into three exclusive subsets: 70% for training, 15% for validation, and 15% for testing.

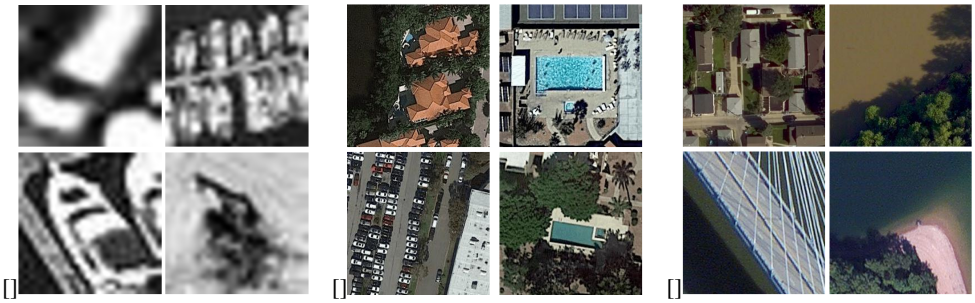


Fig. 5. The three datasets used in our experiments are: a) ovhead-mnist example b) patternNet example c) RSI-CB256 example

3.2. Model preparation

In our experiments, our model employs two quantum convolutional kernels for feature extraction. Additionally, we selected two classical convolutional schemes and four quantum convolutional schemes for comparison. The two classical models are CNN3 and DenseNet [48], where CNN3 corresponds to the classical counterpart of our model, utilizing average convolution. DenseNet includes an initial convolutional layer, two consecutive Dense blocks, a Transition Layer, and a final fully connected layer for classification.

In hybrid quantum–classical convolutional frameworks, classical data are encoded via single-qubit rotation gates and entangled through fixed circuits or CNOT gates [49]. While such methods exploit quantum superposition and entanglement for high-dimensional feature representation, limited qubits or deep circuits increase resource demands, training

complexity, and noise accumulation. The proposed method optimizes quantum kernel design and data encoding, enabling effective feature extraction with few qubits and shallow circuits. Compared to conventional approaches, it achieves higher convolutional efficiency and classification accuracy with low quantum overhead, while offering greater scalability across diverse inputs and convolutional structures.

From the perspectives of encoding, entanglement, and qubit count, we selected four existing Hybrid Quantum-Classical convolutional models for comparison.The four comparative quantum models are QCNN1 [45], QCNN2 [31], QCNN3 [30] and QCNN4 [50]. The model proposed by Fanfan [45] uses ten qubits encoded through H and RY gates and employs U3 gates to convolve the four input channels. It was evaluated on the Overhead-MNIST dataset [32]. The following three models all use four quantum bits to convolve four classical data.The model in [31] applies Hadamard (H) gates to establish entanglement and then uses RY gates to encode classical data into quantum states. The quantum convolutional model in [30] employs four qubits and integrates three types of encoding schemes. The model presented in [50] constructs the quantum circuit primarily by combining RY and CNOT gates to configure the circuit and generate entanglement. For a fair comparison across models and datasets in this work, we restrict all experiments to single-channel images.These selected models provide a comprehensive benchmark covering different qubit configurations, encoding strategies, and entanglement mechanisms, allowing a systematic evaluation of hybrid quantum-classical convolutional architectures and highlighting the potential advantages and limitations of each design in practical image processing tasks.

3.3. Experimental setup

Hardware and Software Environment: In our study, IBM's Qiskit platform [51] was used for model implementation and training. Qiskit, a major framework in quantum deep learning, enables the use of multiple simulator variants.Interested readers can refer to [52] for an in-depth discussion of different frameworks. Additionally, we used an Intel(R) Xeon(R) Gold 6458Q CPU and Nvidia A800 GPU to build a Qiskit simulation platform based on Ubuntu 22.04.5 LTS.

Experimental Setup: To ensure fairness in comparative experiments, all datasets were reduced to 8×8 using the Lanczos algorithm [43], and single-channel images were used for training. The number of epochs was set to 200, employing the cross-entropy loss function and the Adam optimizer [53]. Each training session was repeated three times, and the model that achieved the highest validation accuracy during training was employed for testing. The average test accuracy and its corresponding standard deviation were calculated for comparison and discussion. It is worth noting that, unlike [45, 54, 55], which use noiseless quantum simulators, our experiments introduce quantum noise during training to better simulate real quantum environments. Consequently, all quantum model experiments additionally incorporate the noise effects associated with running quantum models, compared with classical models.

Since the experiments are conducted on a quantum noise simulation platform, to better characterize the noise carried by the platform, we use four simple quantum circuits for the experiments. The circuits are shown in Figure 6 and mainly consist of H gates, RX gates, and CNOT gates. The H gates are used to test the noise in quantum superposition, the RX gates test the noise introduced by single-qubit rotation operations (such as RX, RY,

RZ, and U3 gates), and the CNOT gates test the noise in entangling operations. These circuits enable testing of both quantum entanglement and superposition, recording the noisy measurement outcomes on the experimental platform. Each experiment is repeated three times, and the mean and standard deviation of the results are calculated.

The experimental results are shown in Table 1, illustrating the impact of the experimental noise on quantum operations in our setup.Four simple quantum circuits based on the H gate, RX gate, and CNOT gate were designed for the experimental evaluation. The results indicate that, for circuits involving only single-qubit superposition operations, the measured outcomes agree well with the theoretical values, and the impact of noise remains minimal. In contrast, when CNOT-induced entanglement and RX-based rotations are introduced, the cumulative noise effects become significantly more pronounced, leading to larger deviations between measured and ideal results, as well as an expansion in the fluctuation range of the measurement data. This phenomenon highlights the platform's sensitivity to noise during two-qubit entanglement and dynamic rotational operations. Furthermore, a quantitative analysis of the platform's quantum noise was conducted to assess its impact on the execution of quantum circuits.

Based on the above settings, we conducted two main categories of experiments: comparative experiments and ablation studies. Specifically, five types of experiments were designed to evaluate model performance:

1) Comparative experiments across different models: This experiment aims to compare the performance of our model with six other models.

2) Performance comparison under different quantum encoding schemes: This experiment investigates how different encoding strategies affect the performance of our model.

3) Performance comparison under varying numbers of convolutional kernels: Since our model employs two quantum convolutional kernels, this experiment examines the influence of each kernel on model performance. 4) Performance comparison with different numbers of qubits: One of the primary goals of our proposed model is to reduce quantum resource consumption. Accordingly, we modified our model to a three-qubit quantum convolutional version to evaluate the effect of qubit quantity on model performance. 5) Assessment of the proposed quantum circuit's contribution to classification: In this experiment, the quantum component of our model is removed, and the preprocessed data is directly fed into the dense layer, effectively degrading the model to a fully connected neural network.

Among these five experiments, the first experiment utilizes the full dataset, whereas the subsequent experiments are conducted using the Overhead-MNIST and PatternNet datasets.

3.4. Experimental results

Model comparison performance analysis: We present the classification performance of different models in Table 2. As shown in Table 2, it is important to note that in Experiment 1, the configuration of the dataset, the preprocessing procedure and the model architecture are consistent with those in [45]. Therefore, the experimental data labeled as No. 1 are taken directly from that reference. For all other comparative experiments, to ensure fairness in all tests, we used single-channel input data, whereas [45] used randomly generated datasets. Consequently, those results are not adopted from the original paper but are instead based on our reproduced experiments under unified experimental settings.

The results demonstrate that, across the remote sensing datasets used in this study, our model achieves superior performance while using the same or even fewer qubits compared to other quantum models. compared with its classical convolutional counterpart, our proposed model shows a significant improvement in classification accuracy. Furthermore, even compared to the more complex DenseNet model, our model achieves comparable or better classification performance. Hence, relative to the classical baseline, our model exhibits better generalization ability, and when compared with quantum models that employ the same or greater number of qubits, it consistently outperforms them. Against more complex classical models, our method still achieves competitive or superior results.

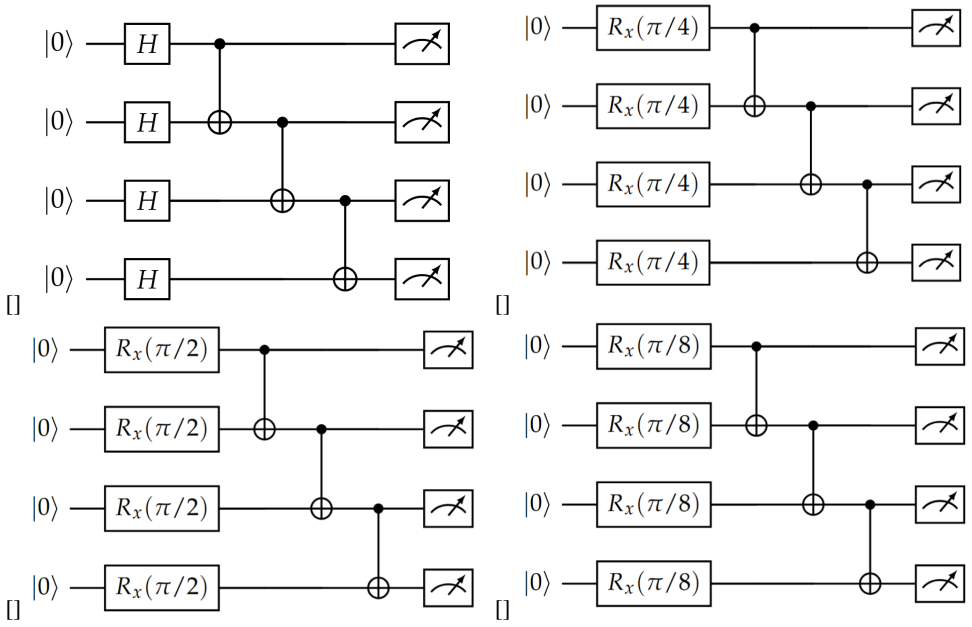


Fig. 6. Four quantum circuits for experimental environment noise evaluation

Table 1. Experimental environment noise assessment

qbit	1		2		3		4	
	Noisy	Ideal	Noisy	Ideal	Noisy	Ideal	Noisy	Ideal
1	0.495 ± 0.0308	0.5	0.1504 ± 0.0181	0.1464	0.5104 ± 0.0268	0.5	0.0547 ± 0.0099	0.38
2	0.4850 ± 0.0343	0.5	0.2630 ± 0.0245	0.25	0.4987 ± 0.0031	0.5	0.0872 ± 0.0124	0.073
3	0.5020 ± 0.0059	0.5	0.3463 ± 0.0490	0.3232	0.5117 ± 0.0326	0.5	0.1237 ± 0.0178	0.106
4	0.5098 ± 0.0174	0.5	0.3893 ± 0.0169	0.375	0.5332 ± 0.0260	0.5	0.1491 ± 0.0250	0.136

Moreover, as shown in Table 2, our model not only attains higher test accuracy but also demonstrates substantially lower standard deviation under quantum noise conditions. This indicates that our model has stronger stability, which is particularly meaningful given that the purity of current quantum systems has not yet reached ideal levels [56].

Coding analysis: Our model is capable of employing multiple types of quantum gates, such as R_X, R_Y, and R_Z, for quantum state encoding. To evaluate the impact of different encoding schemes on classification performance, we conducted a series of experiments in which the R_X, R_Y, and R_Z gates were respectively applied to the quantum encoding circuit. Each experiment was repeated three times, and the mean classification accuracy and standard deviation were computed. The experiments were performed on the OVhead and Pattern datasets described in Section 3.1, using the corresponding dataset configurations, while the experimental setup followed the description in Section 3.3. The results are summarized in Table 3.

As shown in experiments 1–3 and 13–15, the performance of the model when using R_X and R_Y gates for quantum encoding is comparable, indicating that the proposed model does not rely on a specific encoding gate and exhibits robustness with respect to the choice of encoding operation. However, when the R_Z gate is used for encoding, the classification accuracy approaches random guessing, suggesting that the model fails to perform effective learning. Through an analysis of the quantum encoding data, we found that this issue originates from a limitation in the Qiskit framework developed by IBM [51]. Specifically, during the measurement process $qc.measure$, Qiskit cannot directly observe the phase rotation induced by the R_Z operation, resulting in an inability to correctly measure the quantum states that carry convolutional information. Although this limitation falls beyond the primary scope of this study, we nevertheless present the corresponding experimental results to ensure the completeness of our analysis.

The impact of convolution kernel on performance: To evaluate the impact of the number of quantum convolutional kernels on the overall model performance, we conducted experiments using one, two, three, and four quantum convolutional kernels. The experiments were performed on the OVhead and Pattern datasets described in Section 3.1, following the corresponding dataset configurations and the experimental settings outlined in Section 3.3. The results are presented in Table 3, corresponding to experiments 4–7 and 16–19.

When only a single quantum convolutional kernel was employed, the test accuracy on the two datasets decreased by approximately 3% and 2%, respectively, compared with the model using two kernels. In contrast, when three and four convolutional kernels were used, a slight performance degradation of about 1–2% was observed. These results suggest that the performance improvement exhibits diminishing marginal returns with the increasing number of quantum convolutional kernels. Once the model reaches two quantum convolutional kernels, further increasing the kernel count yields negligible or even negative performance gains. Therefore, in the design of quantum convolutional models, blindly increasing the number of convolutional kernels does not necessarily lead to enhanced model performance.

The impact of the number of qubits on classification performance: To investigate the effect of the number of qubits on classification performance, we conducted experiments using 1, 2, 3, and 4 qubits to perform convolution on 9 classical data inputs. When 0 qubits were used, the model degenerated into a classical neural network, specifically a fully connected network. The experiments were performed on the OVhead and Pattern datasets described in Section 3.1, using the corresponding dataset configurations and following the experimental settings in Section 3.3. The results are reported in Table 3, corresponding to experiments 8–12 and 20–24.

Table 2. Model Performance Comparison on Different Datasets.

Dataset	Model	Number	Qubit	Train Acc (%)	Val Acc (%)	Test Acc (%)
Overhead-MNIST [32]	QCNN1 [45]	1	10	69.70±0.40	68.20±0.80	67.20±0.40
	QCNN2 [31]	2	4	82.88±0.67	79.32±0.99	76.56±0.86
	QCNN3 [30]	3	4	68.99±1.88	67.80±1.04	66.35±0.65
	QCNN4 [50]	4	4	95.75±2.16	80.06±0.98	76.19±0.95
	cnn3	5	–	88.90±2.75	77.97±0.72	77.29±1.34
	densnet [48]	6	–	84.42±2.36	83.29±0.20	83.15±0.59
	ours	7	4	96.35±1.81	84.34±0.80	84.46±0.42
Patter-Net [33]	QCNN1 [45]	8	10	91.63±5.02	77.59±1.37	78.15±1.12
	QCNN2 [31]	9	4	86.61±1.50	86.85±1.40	83.98±0.42
	QCNN3 [30]	10	4	68.12±2.24	68.24±4.40	66.57±1.85
	QCNN4 [50]	11	4	97.32±2.49	86.57±1.16	86.20±0.85
	cnn3	12	–	88.23±2.40	83.70±1.53	79.54±1.12
	densnet [48]	13	–	87.50±1.00	87.87±0.58	86.94±0.73
	ours	14	4	93.27±1.40	88.61±0.56	90.56±0.28
RSI-CB256 [34]	QCNN1 [45]	15	10	93.96±2.96	73.06±0.76	73.80±0.47
	QCNN2 [31]	16	4	75.71±1.46	74.06±0.56	71.79±0.93
	QCNN3 [30]	17	4	84.02±3.55	78.70±1.22	79.42±1.62
	QCNN4 [50]	18	4	92.03±1.54	81.64±0.78	79.74±0.68
	cnn3	19	–	89.52±3.61	77.32±1.25	79.98±0.95
	densnet [48]	20	–	89.25±0.48	87.66±0.47	84.59±0.19
	ours	21	4	95.65±0.85	82.27±0.47	84.24±0.33

* Acc stands for accuracy.

It can be observed that as the number of qubits decreases, the information load per qubit increases, leading to a sharp decline in classification performance. Experiments 8 and 20 in Table 3 correspond to the scenario where the proposed quantum model is entirely removed and the preprocessed data is fed directly into the dense layer. In this case, the contribution of our quantum model to classification performance is approximately 5%, demonstrating the effectiveness of the proposed model under current technological conditions.

Table 3. The impact of different modules on model performance

dataset	number	encoding			kernel				qbit					test Acc
		rx	ry	rz	1con	2con	3con	4con	0bit	1bit	2bit	3bit	4bit	
Overhead-MNIST [32]	1	✓				✓							✓	84.4584 ± 0.4153
	2		✓			✓							✓	83.7258 ± 0.7419
	3			✓		✓							✓	17.4254 ± 0.0000
	4	✓			✓								✓	81.2664 ± 0.5943
	5	✓				✓							✓	84.4584 ± 0.4153
	6	✓					✓						✓	83.4642 ± 0.0906
	7	✓						✓					✓	82.5746 ± 1.1320
	8								✓					78.8592 ± 0.3951
	9	✓				✓				✓				42.7525 ± 1.1138
	10	✓				✓					✓			52.7473 ± 2.2146
	11	✓				✓						✓		70.0157 ± 0.8307
	12	✓				✓							✓	84.4584 ± 0.4153
Pattern-Net [33]	13	✓				✓							✓	90.5556 ± 0.2778
	14		✓			✓							✓	90.0000 ± 0.4811
	15			✓		✓							✓	31.6667 ± 0.0000
	16	✓			✓								✓	88.1481 ± 0.4243
	17	✓				✓							✓	90.5556 ± 0.2778
	18	✓					✓						✓	88.9815 ± 0.1604
	19	✓						✓					✓	89.4444 ± 0.4811
	20								✓					83.7037 ± 0.8929
	21	✓				✓				✓				66.4815 ± 1.3981
	22	✓				✓					✓			74.3519 ± 1.1226
	23	✓				✓						✓		82.5000 ± 0.7349
	24	✓				✓							✓	90.5556 ± 0.2778

* Acc stands for accuracy.

4. Discussion

Through comparative experiments in Table 2, we demonstrate that the proposed HAQCCN model exhibits superior classification performance compared with existing quantum and hybrid architectures. Specifically, HAQCCN shows strong adaptability to remote sensing data, effectively capturing spectral–spatial correlations through its hierarchical quantum convolutional kernels. Furthermore, under varying levels of quantum noise, HAQCCN maintains more stable accuracy across different datasets, indicating enhanced robustness of its quantum layers and noise-tolerant encoding strategy.

In contrast, existing hybrid quantum–classical models often suffer from degraded performance when applied to remote sensing data, primarily due to insufficient feature entanglement or limited encoding capacity. They are also more sensitive to hardware-induced quantum noise, leading to fluctuations in measurement outcomes and reduced generalization. These limitations highlight the advantage of our model's design, where the adaptive quantum convolution mechanism and optimized hybrid structure jointly contribute to im-

proved noise resilience and overall classification performance.Moreover, several noteworthy phenomena emerged during the experiments, and these are analyzed in detail below.

As observed in Table 2, experiments 5, 12, and 19 show that the performance of classical convolutional models corresponding to our quantum convolution model is substantially outperformed by our proposed model. However, in experiments 6, 13, and 20, the performance of Densenet is comparable to that of our model. These results indicate that while our model significantly surpasses basic classical schemes, classical deep learning methods remain highly competitive, reflecting decades of development. On the other hand, even at this early stage, deep learning models integrating quantum techniques are already capable of challenging classical approaches. Notably, our model achieves high classification performance using only four qubits. As quantum technology continues to advance rapidly, increasing the number of available qubits will likely enhance the impact of our model across deep learning applications, including remote sensing image classification.

Additionally, we observe that in experiments 11, 17, and 18 of Table 2, our quantum model demonstrates superior performance compared to other quantum models, though its performance is not consistently outstanding across all datasets. Moreover, experiments 1, 8, and 15 show that the Fanfan model exhibits unstable performance, likely because it was originally designed for noiseless environments. The introduction of quantum noise significantly degrades its performance, as noted by the original authors. Another contributing factor may be the low utilization of qubits in its circuit, where quantum gates act on only a subset of qubits while the remaining qubits remain underutilized. Furthermore, after the quantum circuit operates, the system enters a complex entangled state [57], whose potential characteristics are not yet fully understood. Therefore, the observed anomalies in these experiments warrant further investigation, offering insights into the intrinsic properties of quantum circuits and deepening our understanding of quantum computation.

In Table 3, experiments 4–7 and 16–19, we tested the effect of varying the number of quantum convolutional kernels on model performance. When the number of kernels increases, performance does not necessarily improve and may even slightly decline. We attribute this to the increased complexity in feature processing, making it more difficult for the dense layers to capture the relevant information. Similar phenomena have been observed in classical deep learning methods [58–60], but due to quantum entanglement and superposition, the added complexity in quantum convolution can make feature extraction more challenging. Therefore, when using quantum convolutional kernels for feature extraction, excessive quantum operations may obscure features, reducing the effectiveness of downstream layers.

In Table 3, experiments 8–12 and 20–24, we investigated the impact of qubit quantity on classification performance. The results show that decreasing the number of qubits significantly reduces the accuracy of model. However, this reduction does not necessarily imply a deficiency in the feature extraction capability of the quantum circuit; rather, it indicates that a fully connected neural network struggles to utilize these features effectively for classification. These experiments highlight that the number of qubits critically affects quantum convolution-based image classification. Simultaneously, the limited availability of quantum resources restricts the strategy of simply increasing qubit count to improve performance. Accordingly, moving beyond the conventional approach of using a qubit per data element and instead processing more data with fewer qubits becomes a pivotal strat-

egy. Furthermore, this approach is essential for enhancing the computational efficiency of future quantum computers, which represents one of the primary motivations for our proposed model.

5. Conclusions

We present a quantum convolutional architecture capable of encoding a greater volume of classical information while operating with a limited qubit count, resulting in substantially enhanced qubit utilization. In experiments on remote sensing image classification, the approach outperforms existing methodologies. Furthermore, we investigated the effects of quantum encoding techniques, quantum convolutional kernels, and qubit quantity on model performance, providing a more in-depth exploration of the model's internal mechanisms. The proposed model offers a valuable reference for employing quantum computing in remote sensing image recognition and contributes to the advancement of quantum convolutional models.

The results of this research also suggest several avenues for future investigation: (1) developing image encoding and convolution techniques better suited for remote sensing image recognition, (2) exploring strategies to mitigate the problem of qubit information overload under constrained quantum conditions, and (3) designing feature extraction methods specifically tailored for scenarios with qubit information overload.

Acknowledgement. This work is supported by the Scientific Research Fund of National Natural Science Foundation of China (Grant 62372168), the Hunan Provincial Natural Science Foundation of China (Grants 2023JJ30266 and 2025JJ50399), the Research Project on teaching reform in Hunan province (Grant HNJG-2022-0791), the Hunan University of Science and Technology (Grant 2022-448),Key Laboratory of AI and Information Processing, Education Department of Guangxi Zhuang Autonomous Region (Hechi University))(grant 2024GXZDSY010), and the National Social Science Funds of China (Grant 19BZX044).

References

1. Anderson, K., Ryan, B., Sonntag, W., Kavvada, A., Friedl, L.: Earth observation in service of the 2030 agenda for sustainable development. Geo-spatial Information Science 20(2), 77–96 (2017)
2. Ustin, S.L., Middleton, E.M.: Current and near-term advances in earth observation for ecological applications. Ecological Processes 10(1), 1 (2021)
3. Gerasopoulos, E., Bailey, J., Athanasopoulou, E., Speyer, O., Kocman, D., Raudner, A., Tsouni, A., Kontoes, H., Johansson, C., Georgiadis, C., et al.: Earth observation: An integral part of a smart and sustainable city. Environmental Science & Policy 132, 296–307 (2022)
4. McCabe, M.F., Rodell, M., Alsdorf, D.E., Miralles, D.G., Uijlenhoet, R., Wagner, W., Lucieer, A., Houborg, R., Verhoest, N.E., Franz, T.E., et al.: The future of earth observation in hydrology. Hydrology and earth system sciences 21(7), 3879–3914 (2017)
5. Wang, Z., Ma, Y., Zhang, Y., Shang, J.: Review of remote sensing applications in grassland monitoring. Remote Sensing 14(12), 2903 (2022)
6. Chi, M., Plaza, A., Benediktsson, J.A., Sun, Z., Shen, J., Zhu, Y.: Big data for remote sensing: Challenges and opportunities. Proceedings of the IEEE 104(11), 2207–2219 (2016)

7. Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F.: Deep learning in remote sensing: A comprehensive review and list of resources. IEEE geoscience and remote sensing magazine 5(4), 8–36 (2017)
8. Shi, C., Zhang, X., Sun, J., Wang, L.: Remote sensing scene image classification based on self-compensating convolution neural network. Remote Sensing 14(3), 545 (2022)
9. Xu, X., Li, W., Ran, Q., Du, Q., Gao, L., Zhang, B.: Multisource remote sensing data classification based on convolutional neural network. IEEE Transactions on Geoscience and Remote Sensing 56(2), 937–949 (2017)
10. Zhang, L., Zhang, L.: Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. IEEE Geoscience and Remote Sensing Magazine 10(2), 270–294 (2022)
11. Zhang, B., Wu, Y., Zhao, B., Chanussot, J., Hong, D., Yao, J., Gao, L.: Progress and challenges in intelligent remote sensing satellite systems. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15, 1814–1822 (2022)
12. Cong, I., Choi, S., Lukin, M.D.: Quantum convolutional neural networks. Nature Physics 15(12), 1273–1278 (2019)
13. Ristè, D., Da Silva, M.P., Ryan, C.A., Cross, A.W., Córcoles, A.D., Smolin, J.A., Gambetta, J.M., Chow, J.M., Johnson, B.R.: Demonstration of quantum advantage in machine learning. npj Quantum Information 3(1), 16 (2017)
14. Chen, L., Li, T., Chen, Y., Chen, X., Wozniak, M., Xiong, N., Liang, W.: Design and analysis of quantum machine learning: a survey. Connection Science 36(1), 2312121 (2024)
15. Schumacher, B., Nielsen, M.A.: Quantum data processing and error correction. Physical Review A 54(4), 2629 (1996)
16. Eldar, Y.C., Oppenheim, A.V.: Quantum signal processing. IEEE Signal Processing Magazine 19(6), 12–32 (2002)
17. Senokosov, A., Sedykh, A., Sagingalieva, A., Kyriacou, B., Melnikov, A.: Quantum machine learning for image classification. Machine Learning: Science and Technology 5(1), 015040 (2024)
18. Caraiman, S., Manta, V.I.: Image segmentation on a quantum computer. Quantum Information Processing 14(5), 1693–1715 (2015)
19. Youssry, A., El-Rafei, A., Elramly, S.: A quantum mechanics-based framework for image processing and its application to image segmentation. Quantum Information Processing 14(10), 3613–3638 (2015)
20. Abbas, A., Ambainis, A., Augustino, B., Bärtschi, A., Buhrman, H., Coffrin, C., Cortiana, G., Dunjko, V., Egger, D.J., Elmegreen, B.G., et al.: Challenges and opportunities in quantum optimization. Nature Reviews Physics pp. 1–18 (2024)
21. Liang, W., Liu, Y., Yang, C., Xie, S., Li, K., Susilo, W.: On identity, transaction, and smart contract privacy on permissioned and permissionless blockchain: a comprehensive survey. ACM Computing Surveys 56(12), 1–35 (2024)
22. Hu, N., Zhang, D., Xie, K., Liang, W., Diao, C., Li, K.C.: Multi-range bidirectional mask graph convolution based gru networks for traffic prediction. Journal of Systems Architecture 133, 102775 (2022)
23. Efthymiou, S., Ramos-Calderer, S., Bravo-Prieto, C., Pérez-Salinas, A., García-Martín, D., Garcia-Saez, A., Latorre, J.I., Carrazza, S.: Qibo: a framework for quantum simulation with hardware acceleration. Quantum Science and Technology 7(1), 015018 (2021)
24. Preskill, J.: Quantum computing in the nisq era and beyond. Quantum 2, 79 (2018)
25. Brady, L.T., Baldwin, C.L., Bapat, A., Kharkov, Y., Gorshkov, A.V.: Optimal protocols in quantum annealing and quantum approximate optimization algorithm problems. Physical Review Letters 126(7), 070505 (2021)
26. Herrmann, J., Llima, S.M., Remm, A., Zapletal, P., McMahon, N.A., Scarato, C., Swiadek, F., Andersen, C.K., Hellings, C., Krinner, S., et al.: Realizing quantum convolutional neural

networks on a superconducting quantum processor to recognize quantum phases. Nature communications 13(1), 4144 (2022)
27. Oh, S., Choi, J., Kim, J.: A tutorial on quantum convolutional neural networks (qcnn). In: 2020 International Conference on Information and Communication Technology Convergence (ICTC). pp. 236–239. IEEE (2020)
28. Meng, Y.M., Zhang, J., Zhang, P., Gao, C., Ran, S.J.: Residual matrix product state for machine learning. SciPost Physics 14(6), 142 (2023)
29. Wu, S., Zhang, Y., Li, J.: Quantum data parallelism in quantum neural networks. Physical Review Research 7(1), 013177 (2025)
30. Yang, C.H.H., Qi, J., Chen, S.Y.C., Chen, P.Y., Siniscalchi, S.M., Ma, X., Lee, C.H.: Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6523–6527. IEEE (2021)
31. Ovalle-Magallanes, E., Avina-Cervantes, J.G., Cruz-Aceves, I., Ruiz-Pinales, J.: Hybrid classical–quantum convolutional neural network for stenosis detection in x-ray coronary angiography. Expert Systems with Applications 189, 116112 (2022)
32. Noever, D., Noever, S.E.M.: Overhead mnist: A benchmark satellite dataset. arXiv preprint arXiv:2102.04266 (2021)
33. Zhou, W., Newsam, S., Li, C., Shao, Z.: Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS journal of photogrammetry and remote sensing 145, 197–209 (2018)
34. Li, H., Dou, X., Tao, C., Wu, Z., Chen, J., Peng, J., Deng, M., Zhao, L.: Rsi-cb: A large-scale remote sensing image classification benchmark using crowdsourced data. Sensors 20(6), 1594 (2020)
35. Xiong, H., Wu, Z., Fan, H., Li, G., Jiang, G.: Quantum rotation gate in quantum-inspired evolutionary algorithm: A review, analysis and comparison study. Swarm and Evolutionary Computation 42, 43–57 (2018)
36. Bataille, M.: Quantum circuits of cnot gates: optimization and entanglement. Quantum Information Processing 21(7), 269 (2022)
37. Liang, W., Li, Y., Xie, K., Zhang, D., Li, K.C., Souri, A., Li, K.: Spatial-temporal aware inductive graph neural network for c-its data recovery. IEEE Transactions on Intelligent Transportation Systems 24(8), 8431–8442 (2022)
38. Liao, J., Guo, L., Jiang, L., Yu, C., Liang, W., Li, K., Pop, F.: A machine learning-based feature extraction method for image classification using resnet architecture. Digital Signal Processing 160, 105036 (2025)
39. Bhat, H.A., Khanday, F.A., Shah, K.A.: Optimal circuit decomposition of reversible quantum gates on ibm quantum computers. In: Handbook of Research on Quantum Computing for Smart Environments, pp. 149–164. IGI Global (2023)
40. Eltschka, C., Huber, M., Morelli, S., Siewert, J.: The shape of higher-dimensional state space: Bloch-ball analog for a qutrit. Quantum 5, 485 (2021)
41. Scholtz, F.G., Geyer, H.B., Hahne, F.: Quasi-hermitian operators in quantum mechanics and the variational principle. Annals of Physics 213(1), 74–101 (1992)
42. Liao, J., Yu, C., Jiang, L., Guo, L., Liang, W., Li, K., Pathan, A.S.K.: A method for composite activation functions in deep learning for object detection. Signal, Image and Video Processing 19(5), 362 (2025)
43. Duchon, C.E.: Lanczos filtering in one and two dimensions. Journal of Applied Meteorology (1962-1982) pp. 1016–1022 (1979)
44. Cerezo, M., Verdon, G., Huang, H.Y., Cincio, L., Coles, P.J.: Challenges and opportunities in quantum machine learning. Nature computational science 2(9), 567–576 (2022)
45. Fan, F., Shi, Y., Guggemos, T., Zhu, X.X.: Hybrid quantum-classical convolutional neural network model for image classification. IEEE transactions on neural networks and learning systems (2023)

46. Fan, F., Shi, Y., Zhu, X.X.: Land cover classification from sentinel-2 images with quantum-classical convolutional neural networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2024)
47. Sebastianelli, A., Zaidenberg, D.A., Spiller, D., Le Saux, B., Ullo, S.L.: On circuit-based hybrid quantum neural networks for remote sensing imagery classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15, 565–580 (2021)
48. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
49. Hur, T., Kim, L., Park, D.K.: Quantum convolutional neural network for classical data classification. Quantum Machine Intelligence 4(1), 3 (2022)
50. Liu, J., Lim, K.H., Wood, K.L., Huang, W., Guo, C., Huang, H.L.: Hybrid quantum-classical convolutional neural networks. Science China Physics, Mechanics & Astronomy 64(9), 290311 (2021)
51. Javadi-Abhari, A., Treinish, M., Krsulich, K., Wood, C.J., Lishman, J., Gacon, J., Martiel, S., Nation, P.D., Bishop, L.S., Cross, A.W., Johnson, B.R., Gambetta, J.M.: Quantum computing with Qiskit (2024)
52. Ramos Ferreira, F., Fernandes, J.P., Abreu, R.: Quantum software frameworks for deep learning. In: Quantum Software Engineering, pp. 281–302. Springer (2022)
53. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
54. Jones, M.A., Vallury, H.J., Hill, C.D., Hollenberg, L.C.: Chemistry beyond the hartree–fock energy via quantum computed moments. Scientific Reports 12(1), 8985 (2022)
55. Nation, P.D., Kang, H., Sundaresan, N., Gambetta, J.M.: Scalable mitigation of measurement errors on quantum computers. PRX Quantum 2(4), 040326 (2021)
56. Chen, L., Jia, Z.: On optimum entanglement purification scheduling in quantum networks. IEEE Journal on Selected Areas in Communications 42(7), 1779–1792 (2024)
57. Liu, T.: The applications and challenges of quantum teleportation. In: Journal of Physics: Conference Series. vol. 1634, p. 012089. IOP Publishing (2020)
58. Liu, C.T., Wu, Y.H., Lin, Y.S., Chien, S.Y.: Computation-performance optimization of convolutional neural networks with redundant kernel removal. In: 2018 IEEE international symposium on circuits and systems (ISCAS). pp. 1–5. IEEE (2018)
59. Hu, N., Zhang, D., Xie, K., Liang, W., Li, K., Zomaya, A.: Multi-graph fusion based graph convolutional networks for traffic prediction. Computer Communications 210, 194–204 (2023)
60. Wang, X., Xu, L., Zhou, L., Liu, Y., Xiong, N., Li, K.C.: Large language model-driven probabilistic trajectory prediction in the internet of things using spatio-temporal encoding and normalizing flows. Digital Communications and Networks (2025)

Lianghai Chen is currently pursuing a Master's degree in Software Engineering at Hunan University of Science and Technology. He received his Bachelor's degree in Computer Science and Technology in 2023. His research interests include Quantum Computing and Computer Vision (CV).

Yuzhen Liu is a Lecturer at the School of Computer Science and Engineering, Hunan University of Science and Technology. He received a Ph.D. degree in computational mathematics from Xiangtan University, Xiangtan, Hunan, China, in 2012. He has authored or co-authored about 20 journal/conference papers. His research interests include network security protection and information security.

Yi Lu is currently pursuing a Master's degree in Software Engineering at Hu- nan University of Science and Technology. She received her Bachelor's degree in Com- puter Science and Technology in 2024.Her research interests include Emotional Brain- Computer Interfaces (EBCI)and Computer Vision (CV).

Xiaoliang Wang is a professor of information technology and chair of the Department of Internet of Things Engineering,Hunan University of Science and Technology.He leads a team of researchers and students in Information Security and the Internet of Things,such as VANET security and Anonymous Authentication in Ad Hoc Networks.He received a B.E. in computer engineering from Xiangtan University,China,and a M.S. in computer science from the joint education of Xiangtan University and the Institute of Computing Technology of the Chinese Academy of Sciences,China.He received his Ph.D. from Hunan University. He had worked at Xiangtan University and the Nanjing Government of China and had also worked as a postdoctoral researcher at the University of Alabama.

Huaning Song is an associate professor at Hechi University and the vice dean of the School of Artificial Intelligence and Manufacturing. He is also one of the main members of the Key Laboratory of AI and Information Processing under the Education Department of the Guangxi Zhuang Autonomous Region. He specializes in the application of artificial intelligence models and the development of embedded technologies. He graduated from Guilin University of Electronic Technology with a master's degree in circuits and systems.

Received: October 29, 2025; Accepted: January 13, 2026.

Computer Science and Information Systems 23(2):885–916 https://doi.org/10.2298/CSIS260208016Z

FPCA: A Fully Constant-Length and Policy Updating Cross Data Domain Access Control for Cloud-Edge Collaborative Environment

Shiwen Zhang[1,2], Siwei Wen[1,2], and Wei Liang[1,2]

[1] School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China
[2] Sanya Research Institute, Hunan University of Science and Technology, Sanya 572024, China
shiwenzhang@hnu.edu.cn (corresponding author)
siweiwen@mail.hnust.edu.cn
wliang@hnust.edu.cn

Abstract. The secure data sharing across data domains in cloud-edge collaborative environment faces challenges of low data access control efficiency and a lack of dynamism. In this work, we proposed FPCA, a fully constant-length and policy updating cross data domain access control for cloud-edge collaborative environment. FPCA address the above challenges by proposing three algorithms: the Multi Data Domains Key Generation (MDKG) algorithm maintains constant secret key length and enables cross data domain access without attribute conversion, the Constant-Length Ciphertext Encryption (CLCE) algorithm maintains constant ciphertext length, reducing decryption overhead, and the Access Policy Update (APU) algorithm updates the access policy with constant-length update message and low computational complexity. Security analysis demonstrates FPCA resist key forgery, collusion, and chosen-key attacks while ensuring backward security. Simulation experiment shows that FPCA outperforms existing schemes. For FPCA, the storage costs are only 50% to 80% of other constant-length ABE schemes, the communication costs remains at 0.38–1.27 KB between entities, the time costs of access policy update and cross data domain access remain around 8.4 ms and 30.4 ms, respectively, and are lower than other similar schemes. These results confirm the efficiency of FPCA in achieving secure and dynamic cross data domain access in cloud-edge collaborative environment.

Keywords: Cross Data Domain Access Control, Constant-length Secret Key, Constant-length Ciphertext, Access Policy Update, Cloud-edge Collaborative Environment.

1. Introduction

Cloud-edge collaboration technology has promoted the development of the Internet of Things (IoT), improving the scale and efficiency of data sharing across different data domains. By 2030, the total number of IoT devices worldwide is expected to reach 50 billion [13]. Cloud servers provide powerful data storage services, edge servers can preload data from cloud servers, and users interact with nearby edge servers to share or retrieve data of interest from any data domain [19]. Cloud-edge collaboration technology provides low-cost and low-latency data sharing services for IoT users across multiple data domains [26]. Unfortunately, data uploaded directly to cloud or edge servers may pose

security risks of privacy leakage [3, 29]. For this, access control technology is essential to ensure secure data sharing between multiple data domains [14]. Based on the issue of privacy leakage, many works on data access control have been proposed [15–18]. However, these schemes have a huge key management burden and do not provide fine-grained access control, which is unsuitable for a cloud-edge collaborative environment with many IoT users.

To realize fine-grained data access control, Attribute-Based Encryption (ABE) [21,22] has been proposed. ABE provides fine-grained access control by setting attributes and formulating access policies [5, 22]. In cloud-edge collaborative environments, the data that users need to access usually comes from different data domains (Multiple data domains belong to the same trust domain, with different attribute sets between them) [11,19]. However, traditional ABE does not support data access across multiple data domains, and the length of the ciphertext and secret key increases as the number of attributes increases [21], generating massive consumption of computing resources for both servers and users, and reducing the efficiency of data access control [3]. In addition, in the cloud-edge collaborative environment, the Internet of Things must support dynamic user management to adapt to flexible access control permission change requirements, while traditional ABE cannot dynamically adjust user access control permissions. Thus, developing an efficient and dynamic cross data domain access control mechanism for the cloud-edge collaborative environment is essential.

To realize data access control across multiple data domains, several schemes use attribute conversion technology to achieve cross data domain access control [4, 20]. For example, [20] establishes conversion relationships between different attribute sets in different data domains that can convert its own secret key to a valid secret key in the target data domain. Nevertheless, there are different conversion methods between different attributes, which require the storage of multiple conversion keys and bring a high cost of attribute conversion calculation to IoT users. In addition, some attributes may not have a conversion relationship between them, resulting in inflexible key conversion across data domains. [4] consolidates the attributes of multiple data domains into a secret key and [10] embeds the attribute information from multiple data domains in ciphertext to eliminate the attribute conversion operation. However, multiple data domains contain a large number of attributes that significantly increase the length of ciphertexts and secret keys, which also burdens IoT users. [9] can keep the length of the secret key and the ciphertext constant. Unfortunately, this scheme cannot support cross data domain access control. Therefore, maintaining constant ciphertext and secret key length while reducing computation across multiple data domains in cross data domain access control to improve access control efficiency remains a challenge.

To achieve dynamic user access control, many schemes have been designed to update access policies [12, 25]. For example, [25] sets the version number for ciphertext and secret key, and uses a re-encryption key to update the access policy. This scheme requires updating both the ciphertext and the secret keys of all unreleased users simultaneously. [12] uses the Attribute Cuckoo Filter (ACF) to dynamically adjust access policies and only needs to update the ciphertext. However, this scheme needs to update the version number of each ciphertext component and conduct secret sharing of the ciphertext again, equivalent to recalculating the ciphertext. [6] and [30] divided the attributes in the access policies into different types, which can reduce the calculation of the old attribute

part in the ciphertext. However, in order to prevent adversaries from using old secret values to decrypt ciphertext, the secret value of the ciphertext needs to be reselected and the secret shared again, which still requires a lot of computation and makes the efficiency of dynamic access control remain low. Therefore, designing an efficient dynamic access control mechanism with low computational complexity to update the access policy remains a pressing challenge.

In this work, we propose an efficient and dynamic cross data domain access control scheme with policy update for a cloud-edge collaborative environment. To minimize the consumption of computing resources for cross data domain access control, we designed the Multi Data Domains Key Generation (MDKG) algorithm and the Constant-Length Ciphertext Encryption (CLCE) algorithm to achieve efficient cross data domain access control. In MDKG and CLCE, we encode and polymerize attributes across multiple data domains by attribute encoding summation and group element multiplication that can reduce and maintain the length of the secret key and the ciphertext, respectively, which can reduce the computational resource consumption for IoT users. To achieve dynamic user access control, we also designed an Access Policy Update (APU) algorithm that updates the access policy with constant-length update message and low computational complexity. In summary, the contributions of this work are as follows:

1. We propose FPCA, a fully constant-length and policy updating cross data domain access control for cloud-edge collaborative environment, which achieves efficient, fully constant-length cross data domain access control and can update the access policy with low computational complexity.
2. We propose the MDKG algorithm to improve the efficiency of cross data domain access control. The MDKG keeps the secret key length constant, allowing IoT users to efficiently access data from other data domains without any conversion operation.
3. We propose the CLCE algorithm to reduce the computational resource costs of IoT users and servers, keeping the ciphertext length constant and reducing the decryption calculations for IoT users.
4. We propose the APU algorithm to achieve dynamic access control, updating the access policy with constant-length update message and low computational complexity.

The remainder of this article is organized as follows. Section 2 discusses related work relevant to our proposed research, while the preliminary, system model, and threat model are introduced in Section 3. The multiple data domain model, the proposed FPCA, and the correctness analysis are presented in Section 4. We perform a security analysis in Section 5. We analyze and evaluate the performance of our FPCA in Section 6, and finally, concluding remarks and future directions are presented in Section 7.

2. Related Work

2.1. Cross Data Domain Access Control

To address the problem of data access across multiple data domains, Huai *et al.* [8] propose a data sharing scheme across data domains. However, this scheme has system security issues and may be threatened by various attacks during data sharing. For better security, many schemes [20, 23] proposed blockchain-based data access schemes across data

domains. Using the immutability of blockchain to ensure security. Sun *et al.* [20] establishes a conversion relationship between the attributes of each data domain and converts the attributes in the secret key of the IoT user to other attributes in the corresponding data domain through the conversion relationship. However, these schemes need to perform a large number of attribute conversion operations for cross data domain access, which imposes a heavy burden on IoT users and reduces the efficiency of access control. Fan *et al.* [6] consolidates the attributes of different data domains into a single secret key, allowing users to access data from different data domains with a single secret key. In addition, Li *et al.* [10] consolidates the attributes of different data domains into one ciphertext, allowing the ciphertext of a certain data domain to be decrypted by users of other different data domains without the need for additional attribute conversion operations by users. However, the length of the secret key and the ciphertext in these schemes will increase with the number of attributes and data domains, reducing the efficiency of access control. Yannis *et al.* [27] proposed an attribute compression scheme that reduces the number of attribute-related components in both the ciphertext and secret key, thus reducing their lengths. However, the lengths of the ciphertext and secret key still increase with increasing number of attributes. Thus, some constant-size ABE schemes [1,2,7] have been proposed. Allison *et al.* [1] and Wu *et al.* [2] can achieve a constant-length ciphertext, but the length of the secret key still increases with the number of attributes. Fan *et al.* [7] can achieve both the length of the ciphertext and the secret key constant, improving the efficiency of access control. However, these schemes cannot achieve cross data domain access control. Therefore, we design an efficient cross data domain access control scheme that removes the attribute conversion operation and keeps both the secret key length and the ciphertext length constant to improve the efficiency of data access control across multiple data domains.

2.2. Access Policy Update

To improve the efficiency of access policy update, many schemes [10, 28] use the re-encryption key as the update key to update the ciphertext. Xue *et al.* [24] proposed an access policy update scheme for hidden access policies, improving the security of the access policy update. However, the computational cost of the update calculation in these schemes increases with increasing number of attributes in the ciphertext. Fan *et al.* [6] divided the attributes in access policies into three types that can reduce the computation of parts of the ciphertext that do not require updates. Thus, reduces the computational complexity of the update of the access policy. Zuo *et al.* [30] proposed a blockchain-based access policy update scheme that also divided attributes of access policies into different types to improve the efficiency of policy update and use the blockchain to improve security. However, these schemes cannot maintain a constant length of the ciphertext, and the computational cost of updating the old ciphertext increases as the number of attributes in the ciphertext increases. Additionally, when updating the access policy for ciphertext, to prevent adversaries from decrypting the updated ciphertext with old decryption results, it is necessary to set a new secret value and perform the secret sharing calculations again, resulting in low update efficiency of the access policy and reduced flexibility of access control. Fan *et al.* [7] add a timestamp to user's attribute set and access policy. By updating the timestamps in the access policy, [7] can significantly reduce the overhead of

updating the access policy. However, it cannot modify attribute related content in the access policy except for timestamps. Additionally, it fails to update the secret value in the ciphertext, thus failing to meet the backward security. Therefore, we propose an efficient and secure policy update access control scheme that implements the update calculation with complexity $O(1)$ and keeps the length of the update message constant to improve the efficiency and flexibility of the data access control.

In Table 1, we compared the features between FPCA and existing related works.

Table 1. Feature Comparison on FPCA and Other Related Works

Scheme	DAS	CDD	AC	CCL	CSKL	APU	BS	CUM
[1]	Y	N	N	Y	N	N	N	N
[2]	Y	N	N	Y	N	N	N	N
[6]	Y	Y	N	N	N	Y	Y	N
[7]	Y	N	N	Y	Y	Y	N	Y
[8]	N	Y	N	N	N	N	N	N
[10]	Y	Y	N	N	N	Y	Y	N
[20]	Y	Y	N	N	N	N	N	N
[23]	Y	Y	N	N	N	N	N	N
[24]	Y	N	N	N	N	Y	Y	N
[27]	Y	N	Y	N	N	N	N	N
[28]	Y	N	N	N	N	Y	Y	N
[30]	Y	N	N	N	N	Y	Y	N
FPCA	Y	Y	Y	Y	Y	Y	Y	Y

Notes: DAS: Data Secrurity. CDD: Cross Data Domain Access Control. AC: Attribute Compression. CCL: Constant Ciphertext Length. CSKL: Constant Secret Key Length. APU: Access Policy Update. BS: Backward Security. CUM: Constant Update Message Length.

3. Problem Formulation

3.1. Preliminary

Selective game for ABE: The selective game described in the following involves an adversary A and a challenger CH.

1. ***Init***: A chooses an access policy $\mathfrak{R}_1$ and sends it to CH.
2. ***Setup***: CH runs the **System Initialization** algorithm to generate the public parameters PP and the master secret key MK, CH sends PP to A.
3. ***Key Query 1***: A submits the attribute set U_A to CH to request the secret key, U_A cannot satisfy $\mathfrak{R}_1$. CH runs the MDKG algorithm to generate the secret key SK_A based on U_A and sends SK_A to A.
4. ***Challenge***: A sends two messages of equal length $\{M_0, M_1\}$ to CH. CH randomly chooses $\beta \in \{0, 1\}$ and runs the CLCE algorithm to encrypt M_β based on $\mathfrak{R}_1$ and returns the challenge ciphertext CT_c to A.

5. ***Update Ciphertext***: A generates a new access policy $\mathfrak{R}_2$, $\mathfrak{R}_2$ cannot be satisfied by any U_A submitted by A. A sends $\{CT_c, \mathfrak{R}_2\}$ to CH. CH runs the APU algorithm and returns the updated ciphertext CT_{new} based on $\mathfrak{R}_2$ to A.
6. ***Key Query 2***: Repeat the process of **Key Query 1**, but none of the aforementioned secret keys can decrypt CT_c or CT_{new}.
7. ***Guess***: A outputs $\beta^{'} \in \{0, 1\}$ as a guess of β. If $\beta^{'} = \beta$, A wins the security game. The advantage of A to win the security game is defined as $\Omega = Pr[A\ win] - 1/2$.

3.2. System Model

The system model of our FPCA is illustrated in Fig. 1, which involves six participants: 1) **Center Authority**, 2) **Attribute Authority**, 3) **Edge Server**, 4) **Cloud Server**, 5) **Data Owner**, and 6) **Data User**. Their main functions are depicted below.

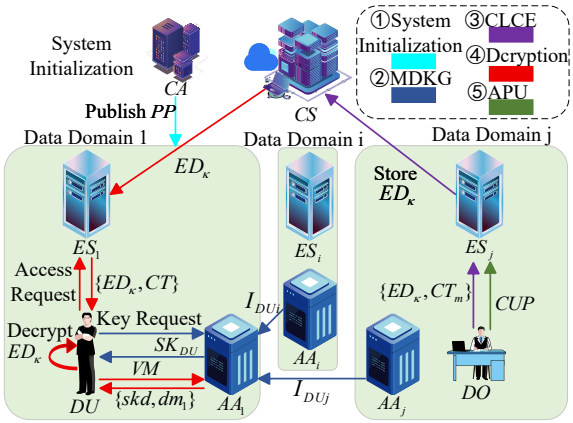


Fig. 1. The System Model of FPCA

1. **Center Authority** (CA)**:** CA is responsible for performing the System Initialization algorithm to generate the public parameters and the master secret key. CA publishes the public parameters for each entity. CA assigns the attribute set to each attribute authority. CA is trusted.
2. **Attribute Authority** (AA)**:** Each AA controls a disjoint attribute subset and is responsible for generating the secret key for the data user. AAs are also responsible for checking if the data user satisfies the access policy. AAs are trusted.
3. **Edge Server** (ES)**:** ES is responsible for the storage of the ciphertext. ES is semi-trusted.
4. **Cloud Server** (CS)**:** CS is only responsible for storing encrypted data.
5. **Data Owner** (DO)**:** DO is an IoT user. DO is responsible for developing and updating access policies. DO is also responsible for encrypting the data to be shared and generating the ciphertext according to the access policy.

6. **Data User (DU):** DU is an IoT user. DU gets the secret key from AA according to its attributes set. DU can obtain any ciphertext of interest from ES. Only DU that comply with the access policy can correctly decrypt the ciphertext using its secret key.

There are multiple data domains in the system, each data domain contains an AA and an ES. After CA performs the system initialization, AA aggregates attributes from all data domains to generate a secret key for DU. DO generates ciphertext based on the access policy. When DU wants to decrypt the ciphertext, AA first checks whether DU meets the access policy of the ciphertext. If it meets the access policy, DU can use the secret key to decrypt the ciphertext. When DO wants to update the access authority to the ciphertext, DO uses the new attribute set to generate a new access policy and an update message accordingly.

3.3. Threat Model

In FPCA, CA and AA are trusted, ES is semi-trusted that will honestly execute algorithms, but it may passively collude with adversary A, A can tamper with the access policy of the ciphertext in ES, leading to data leakage or tampering. DO is trusted but DU is semi-trusted, DU performs algorithms honestly but DU can be an adversary that attempts to access the data of other IoT users. Assume that adversary A can obtain any ciphertext in which it is interested but A cannot get a valid secret key that can decrypt the ciphertext correctly. A may perform the following attacks to break FPCA.

1. **Key Forgery Attack:** The adversary A has an invalid secret key that cannot decrypt the ciphertext, A tries to perform the key forgery attack to forge a valid secret key that can decrypt the ciphertext.
2. **Collusion Attack:** There are two possible ways for adversaries to carry out collusion attacks. First, multiple adversaries with no valid secret key may share their invalid secret key with each other and try to generate a valid secret key. Second, because ES is semi-trusted, the adversary A can temper the ciphertext stored in ES by conspiring with ES, so that A can use its invalid secret key to decrypt the tempered ciphertext.
3. **Chosen-Key Attack:** Adversary A can get multiple secret keys with different attribute sets, but none of these can correctly decrypt the ciphertext. A attempts to derive a valid secret key based on the invalid secret keys it acquired to decrypt the ciphertext.
4. **Attack the Updated Ciphertext:** DO revoked the access authority of the adversary A, A tries to decrypt the updated ciphertext based on its old secret key.

4. The Proposed FPCA Scheme

In this section, we first introduce the notation required for FPCA in Table 2. Then, we introduce the multiple data domain model of FPCA, and the scheme definitions of FPCA are defined. Subsequently, we introduce the construction of FPCA.

Table 2. Notations

Parameter	Description
N	Number of data domains
n	Total number of attributes in U_U
At	Attribute
Dm	Data domain
$\hbar$	Attribute flag, indicating whether the attribute At is in the attribute set
CoE	Attribute location public key, utilized by DO to compute the components related to the attributes within the ciphertext.
I_{DU}	The access structure of DU
$f(x, I)$	The attribute polynomial based on access structure I to convert I into element in Z_p to participate in the computation of ciphertext and secret key
s_1, s_2	The secret sharing
dm_1	The auxiliary decryption component

4.1. Multiple Data Domain Model

Here we first introduce the construction of the multiple data domain model. Then, we introduce attribute encoding and attribute aggregation.

The multiple data domain model of FPCA is illustrated in Fig. 2, which can achieve efficient data access across multiple data domains by combining attribute encoding and aggregation. In FPCA, all data domains belong to the same trust domain. Different data domains can manage non-disjoint attribute sets. Each data domain has an AA, AAi from the data domain Dm_i aggregates the attribute sets sent by AA from all other data domains.

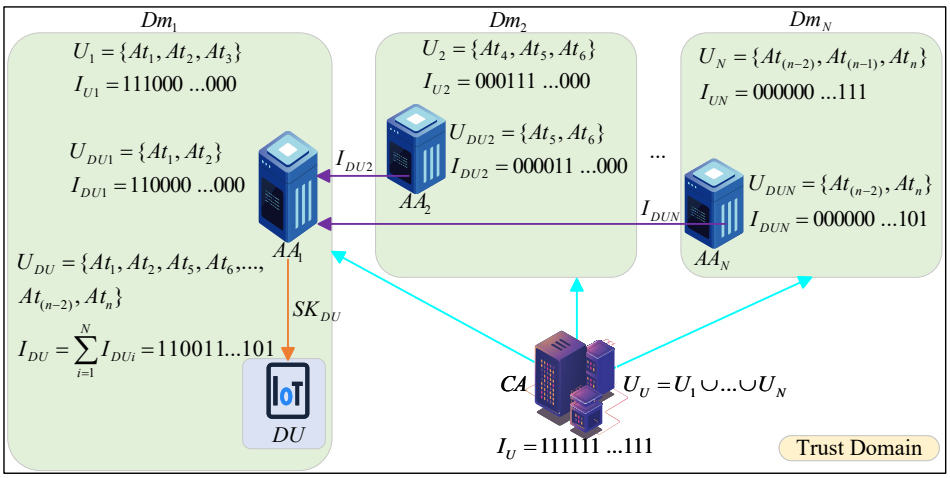


Fig. 2. The Multiple Data Domain Model of FPCA

By encoding attributes, we can quickly aggregate attribute sets from multiple data domains. Assume that U_U is the union of the attribute sets of all data domains, including duplicate attributes, set the attribute set of the data domain Dm_x as $U_x = \{At_i | i \in [1, n]\}$, which is controlled by AA_x, n is the number of attributes in U_U, $U_x \subseteq U_U$. The order of each attribute At_i in U_x and U_U is fixed. We utilize attribute encoding to generate the access structure corresponding to the attribute set. The access structure of U_x is $I_x = \hbar_{x1}..\hbar_{xn}$, which is an n-bit string. If $At_i \in U_x, \hbar_{xi} = 1$, otherwise $\hbar_{xi} = 0$, I_U is the access structure of U_U which is a bit string that entirely consists of 1, $I_U = \sum_{x=1}^{N} I_x$. In this way, the attribute set is converted into a bit string by encoding the attributes as 0 and 1. We select a subset $U_{sj} = \{At_i | i \in [1, n]\}$ for each $U_j, j \in [1, N]$. The access structure of U_{sj} is $I_{sj} = \hbar_{sj1}..\hbar_{sjn}$, if $At_i \in U_{si}, \hbar_{sji} = 1$, otherwise $\hbar_{sji} = 0$. Then, after AA_x in the data domain Dm_x receiving $\{I_{sj} | j \in [1, N], j \neq x\}$ from every other AAs, AA_x performs attribute aggregation through the following operation:

$$Aggregated\ Results = \sum_{j=1}^{N} I_{sj}. \tag{1}$$

In this way, AA_x aggregates all attributes of $\{U_{sj} | j \in [1, N]\}$.

AAs in different data domains can manage a non-disjoint attribute set, and the attribute codes of the attributes in the attribute set managed by each AA correspond to different consecutive bit strings in I_U. Even for the same attributes, their attribute codes are located in different positions in the access structure when they are in different data domains. For At_i and At_j in U_U, they can represent the same attribute when $i \neq j$, At_i and At_j belong to different data domains.

For example, in Fig. 2, CA assigns attribute sets $U_1, ..., U_N$ to $Dm_1, ..., Dm_N$, respectively. $\{I_{Ui} | i \in [1.N]\}$ is the access structure of $\{U_i | i \in [1.N]\}$. DU is in Dm_1. $AA_1, ..., AA_N$ assigns the attribute set $U_{DU1}, ..., U_{DUN}$ to DU, respectively. U_{DU} is the union of $\{U_{DUi} | i \in [1, N]\}$, and U_{DU} is the attribute set of DU. AA_1 in Dm_1 generates I_{DU} by aggregating the access structure $\{I_{DUi} | i \in [1.N]\}$ of $\{U_{DUi} | i \in [1, N]\}$. I_{DU} is the access structure of DU. AA_1 generates the secret key SK_{DU} for DU based on I_{DU}.

4.2. Scheme Definition

The algorithms of the proposed FPCA are defined as follows:

1. **System Initialization** $(1^{\epsilon}) \rightarrow (\boldsymbol{PP}, \boldsymbol{MK})$: CA inputs the security parameter ε, outputs the public parameters PP and the master secret key MK. CA assigns disjoint attribute subsets to each AAi.
2. **MDKG** $(MK, PP) \rightarrow (\boldsymbol{SK_{DU}})$: AA of each data domain defines the access structure I_{DU} for DU together and generates the secret key SK_{DU} for DU based on I_{DU}.
3. **CLCE** $(M, PP) \rightarrow (\boldsymbol{CT})$: DO defines the access policy $\mathfrak{R}$ with the access structure $I_{\mathfrak{R}}$. DO inputs the plaintext M, PP, $I_{\mathfrak{R}}$ and outputs the ciphertext CT.
4. **Decryption** $(CT, ED_{\kappa}, SK_{DU}) \rightarrow (\boldsymbol{M})$: DU gets the ciphertext CT and the encrypted data ED_{κ} from ES. AA checks if DU satisfies the access policy. DU inputs CT, ED_{κ} and the secret key SK_{DU} and outputs the plaintext M.

5. **APU** $(CT, PP, MK) \rightarrow (\boldsymbol{CT_{new}})$: DO inputs the ciphertext CT, PP and MK. DO defines the new access policy $\mathfrak{R}_{new}$, outputs the new ciphertext CT_{new}.

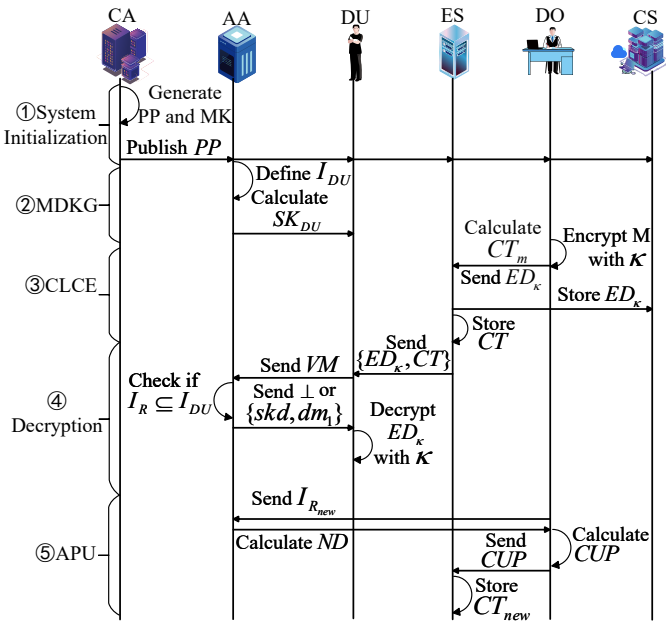


Fig. 3. The Workflow of FPCA

4.3. Construction of FPCA

In this section, we present the detailed construction of the proposed FPCA scheme, made up of five parts: 1) **System Initialization**, 2) **MDKG**, 3) **CLCE**, 4) **Decryption**, and 5) **APU**. The workflow of FPCA is shown in Fig. 3.

1) System Initialization

The System Initialization algorithm is performed by CA and each AA.

- **Step 1**: CA input Security Parameter ε and chooses a bilinear pairing group $BPG = \{G, G^* | G \times G \rightarrow G^*\}$ with prime order p and chooses g as the generator of G. The bilinear pairing is defined as $e(g^x, g^y) = e(g, g)^{xy}$, where $\{x, y \in Z_p\}$.
- **Step 2**: CA defines a hash functions $H : \{0,1\}^* \rightarrow Z_p^*$. H is one-way collision-resistant.
- **Step 3**: CA sets non-disjoint attribute set $\{U_j, j \in [1, N]\}$ for each data domain $Dm_j, j \in [1, N]$, N is the number of data domains. CA defines $U_U = \{At_i | i \in [1, n]\}$, which is the union of the attribute sets of all data domains, $U_j \subseteq U_U$, n is the number of attributes in U_U, and n is invariant. CA assigns $\gamma \in Z_p$ and U_j to AA_j of

each data domain $Dm_j, j \in [1, N]$. We define I_j as the access structure of U_j, I_j is represented as an n-bit string $\{\hbar_1 ... \hbar_n\}$.

$$\begin{cases} \hbar_i = 1, At_i \in U_j, \\ \hbar_i = 0, At_i \notin U_j. \end{cases} \tag{2}$$

- **Step 4**: CA randomly chooses $\{\alpha, q \in Z_p\}$ and sends to each AA. CA computes $CoE = \{coe_i = g^{\gamma^i} | i \in [1, n]\}$ and $BI = \{e(g, g)^\alpha, g^q\}$.
- **Step 5**: CA publishes the public parameters:

$$PP = (BPG, e(\cdot), H, g, p, CoE, BI), \tag{3}$$

and the master secret key is: $MK = (\gamma, q, \alpha)$. For the security of the system, MK is kept by CA and each AA.

2) Multiple Data Domains Key Generation Algorithm

The MDKG algorithm is performed by AA based on the multiple data domain model. The Algorithm 1 shows the details of MDKG.

Algorithm 1: MDKG Algorithm

Input: MK, PP
Output: SK_{DU}
Set n-bit string $I_{DU} = 0...0$;
Set $f(\gamma, I_{DU}) = 1$;
Randomly set $d_{DU_1}, d_{DU_2} \in Z_p$;
for $i \in [1, N]$ **do**
  Generate $I_{DUi} = \hbar_{1DUi} ... \hbar_{nDUi}$;
  Compute $I_{DU} = I_{DU} + I_{DUi}$;
end
for $y \in [1, n]$ **do**
  Compute $f(\gamma, I_{DU}) = f(\gamma, I_{DU}) \cdot (\gamma + H(y))^{1-\hbar_{yDU}}$;
end
Computes
$sk_{DU1} = g^{\alpha + q d_{DU_2}}, sk_{DU2} = g^{d_{DU_2} + d_{DU_1} f(\gamma, I_{DU})}, sk_{DU3} = g^{q d_{DU_1}}$;
Return $SK_{DU} = \{I_{DU}, sk_{DU1}, sk_{DU2}, sk_{DU3}\}$

- **Step 1**: The DU in the data domain Dm_j sends a key generation request to AA_j.
- **Step 2**: AA_j randomly chooses $\{d_{DU_1}, d_{DU_2}\} \in Z_p$ for DU.
- **Step 3**: Each attribute authority $\{AAi | i \in [1, N]\}$ defines an access structure $I_{DUi} = \hbar_{1DUi} ... \hbar_{nDUi}$ for DU based on attributes subset U_{DUi} of DU defined by each AA_i. $\{AA_i | i \in [1, N], i \neq j\}$ sends I_{DUi} to AA_j.
- **Step 4**: AA_j computes:

$$\begin{cases} I_{DU} = \sum_{i=1}^{N} I_{DUi} = \hbar_{1DU}...\hbar_{nDU}, \\ f(\gamma, I_{DU}) = \prod_{y=1}^{n} (\gamma + H(y))^{1-\hbar_{yDU}}, \\ sk_{DU1} = g^{\alpha + qd_{DU_2}}, \\ sk_{DU2} = g^{d_{DU_2} + d_{DU_1} f(\gamma, I_{DU})}, \\ sk_{DU3} = g^{qd_{DU_1}}. \end{cases} \tag{4}$$

AA_j sends the secret key $SK_{DU} = \{I_{DU}, sk_{DU1}, sk_{DU2}, sk_{DU3}\}$ to DU.

SK_{DU} contains only four elements, because n is invariant, the length of I_{DU} is n, which is constant. The attributes assigned to DU by various data domains are aggregated in I_{DU} through the attribute aggregation. The aggregation result I_{DU} is converted into an element in Z_p through $f(\gamma, I_{DU})$ and used to calculate sk_{DU2}. The lengths of sk_{DU1}, sk_{DU2} and sk_{DU3} are $|G|$, $|G|$ is the length of the element in G. Thus, the length of the secret key SK_{DU} is independent of the number of attributes and is constant.

Compared to the standard composite-key ABE, MDKG does not require the secret distribution computation, MDKG can reduce the computational costs and can maintain the length of the secret key constant.

3) Constant-Length Ciphertext Encryption Algorithm

The CLCE algorithm is performed by DO. The Algorithm 2 shows the details of CLCE.

- **Step 1**: The DO of the data domain Dm_j chooses a random number $\kappa \in G^*$ as a symmetric key to encrypt plaintext M with the symmetric encryption algorithm and gets the encrypted data ED_κ. DO sends ED_κ to ES, ES stores ED_κ in CS, CS sends the storage address $Sd(ED_\kappa)$ of ED_κ to ES.
- **Step 2**: DO design an AND-gate access policy $\mathfrak{R}$, $I_{\mathfrak{R}} = \hbar_{1\mathfrak{R}}...\hbar_{n\mathfrak{R}}$ is the access structure of $\mathfrak{R}$, $\mathfrak{R}$ contains only the attributes in U_j, thus the number of "1" in $I_{\mathfrak{R}}$ will not exceed the number of attributes in U_j. Then DO defines:

$$f(x, I_{\mathfrak{R}}) = \prod_{i=1}^{n} (x + H(i))^{1-\hbar_{i\mathfrak{R}}} = e_0 + e_1 x + ... + e_n x^n. \tag{5}$$

  We denote the coefficient of x^i by e_i.
- **Step 3**: DO chooses random numbers $\{s_v, s_1\} \in Z_p$ and sets s_v as the secret value.
- **Step 4**: DO computes:

$$\begin{cases} C_0 = \kappa \cdot e(g, g)^{\alpha s_v}, \\ C_1 = g^{s_v}, \\ C_2 = g^{qs_1}, \\ C_3 = (\prod_{i=0}^{n} coe_i^{e_i})^{s_v} = g^{s_v f(\gamma, I_{\mathfrak{R}})}. \end{cases} \tag{6}$$

Algorithm 2: CLCE algorithm

Input: M, PP
Output: CT
Set $C_3 = 1$;
DO randomly chooses $\kappa \in G^*$;
DO generates ED_κ by encrypts M with κ;
DO sends ED_κ to ES, ES stores ED_κ in CS;
CS returns $Sd(ED_\kappa)$ to ES;
DO generates $I_\mathfrak{R} = \hbar_{1\mathfrak{R}}...\hbar_{n\mathfrak{R}}$;
DO generates $f(x, I_\mathfrak{R}) = \prod_{i=1}^{n}(x + H(i))^{1-\hbar_{i\mathfrak{R}}}$;
DO sets e_i as the coefficient of x^i in $f(x, I_\mathfrak{R})$;
DO randomly chooses $\{s_v, s_1 \in Z_p\}$;
for $i \in [0, n]$ **do**
  DO computes $C_3 = C_3 \cdot (coe_i^{'})^{e_i}$;
end
DO computes $C_3 = C_3^{s_v}$;
DO computes $C_0 = \kappa \cdot e(g, g)^{\alpha s_v}, C_1 = g^{s_v}, C_2 = g^{qs_1}$;
DO sends $CT_m = \{I_\mathfrak{R}, C_0, C_1, C_2, C_3\}$ to ES;
ES stores $CT = \{Sd(ES_\kappa), CT_m\}$;
Return CT

DO stores the middle ciphertext $CT_m = \{I_\mathfrak{R}, C_0, C_1, C_2, C_3\}$ in ES. Finally, ES obtains the complete ciphertext $CT = \{Sd(ED_\kappa), CT_m\}$.

Except for $Sd(ED_\kappa)$, CT contains only five elements, because n is invariant, the length of $I_\mathfrak{R}$ is n, which is constant. As discussed above, $f(\gamma, I_\mathfrak{R})$ is an element in Z_p, the length of C_0 is $|G^*|$, $|G^*|$ is the length of element in G^*. The lengths of C_1, C_2, and C_3 are $|G|$. Thus, the length of the ciphertext CT is independent of the number of attributes and is constant.

4) Decryption

The Decryption algorithm is performed by ES, AA and DU to decrypt CT and ED_κ to obtain plaintext M.

- **Step 1**: The DU in the data domain Dm_j sends an access request to ES, requesting the encrypted data that have attracted its interest. ES finds the matching ciphertext CT and obtains ED_κ from CS according to $Sd(ED_\kappa)$. ES sends ED_κ and CT_m to DU.
- **Step 2**: DU sends the verification message $VM = \{I_\mathfrak{R}, I_{DU}, C_1, C_2, C_3, sk_{DU2}, sk_{DU3}\}$ to AA_j. AA_j checks if the following equation holds:

$$e(C_3, g^{-1})e(g, C_1^{f(\gamma, I_\mathfrak{R})}) \stackrel{?}{=} 1. \tag{7}$$

If Eq. 7 does not hold, it means that CT has been tampered with, if $I_\mathfrak{R} \not\subseteq I_{DU}$, it means that DU does not satisfy the access policy $\mathfrak{R}$ of CT, AA_j return $\perp$ to DU. If Eq. 7 and $I_\mathfrak{R} \subseteq I_{DU}$ hold, AA_j computes the transformed key skd:

$$skd = sk_{DU2} \cdot sk_{DU3}^{(f(\gamma, I_{\mathfrak{R}}) - f(\gamma, I_{DU}))/q}, \tag{8}$$

and AA_j implicitly defines $s_2 = s_v - s_1$ and computes:

$$dm_1 = e(skd, C_1^q / C_2) = e(g, g)^{qs_2 d_{DU_2} + qs_2 d_{DU_1} \cdot f(\gamma, I_{\mathfrak{R}})}. \tag{9}$$

AA_j sends skd and dm_1 to DU.

– **Step 3**: DU computes the intermediate decryption result:

$$dm = \frac{e(skd, C_2) \cdot dm_1}{e(sk_{DU3}, C_3)} = e(g, g)^{qs_v d_{DU_2}}. \tag{10}$$

Then, DU computes:

$$\kappa = \frac{C_0 \cdot dm}{e(C_1, sk_{DU1})}. \tag{11}$$

– **Step 4**: DU decrypt ED_κ with κ to obtain plaintext M.

5) Access Policy Update

The APU algorithm is performed by DO, AA, and ES when DO wants to change the access policy. For example, DO will perform the APU algorithm if DO intends to revoke or add access control privileges for a class of DU. The Algorithm 3 shows the details of the APU algorithm.

Algorithm 3: APU algorithm

Input: CT, PP, MK
Output: CT_{new}
Sets $f(\gamma, I_{\mathfrak{R}_{new}}) = 1$;
DO generates $I_{\mathfrak{R},new} = \hbar_{1\mathfrak{R},new} ... \hbar_{n\mathfrak{R},new}$;
DO sends $I_{\mathfrak{R},new}$ to AA;
for $i \in [1, n]$ **do**
  AA computes $f(\gamma, I_{\mathfrak{R}_{new}}) = f(\gamma, I_{\mathfrak{R}_{new}}) \cdot (\gamma + H(i))^{1 - \hbar_{i\mathfrak{R},new}}$;
end
AA computes $ND = g^{f(\gamma, I_{\mathfrak{R}_{new}})}$ and sends ND to DO;
DO chooses $s_{v,new} \in Z_p$;
DO computes
$C_{0,new} = \kappa \cdot e(g, g)^{\alpha s_{v,new}}, C_{1,new} = g^{s_{v,new}}, C_{3,new} = ND^{s_{v,new}}$;
DO sends $CUP = \{I_{\mathfrak{R}_{new}}, C_{0,new}, C_{1,new}, C_{3,new}\}$ to ES;
ES updates CT_m to $CT_{m,new} = \{CUP, C_2\}$;
Return $CT_{new} = \{Sd(ED_\kappa), CT_{m,new}\}$

– **Step 1**: The DO in the data domain Dm_j sets the new AND-gate access policy $\mathfrak{R}_{new}$, $I_{\mathfrak{R}_{new}} = \hbar_{1\mathfrak{R}_{new}}...\hbar_{n\mathfrak{R}_{new}}$ is the access structure of $\mathfrak{R}_{new}$. DO sends $I_{\mathfrak{R}_{new}}$ to AA_j. AA_j computes:

$$f(\gamma, I_{\mathfrak{R}_{new}}) = \prod_{i=1}^{n} (\gamma + H(i))^{1-\hbar_{i\mathfrak{R}_{new}}}, \tag{12}$$

and sends the attribute update component $ND = g^{f(\gamma, I_{\mathfrak{R}_{new}})}$ to DO.

– **Step 2**: DO chooses a new secret value $s_{v,new} \in Z_p$ and computes:

$$\begin{cases} C_{0,new} = \kappa \cdot e(g,g)^{\alpha s_{v,new}}, \\ C_{1,new} = g^{s_{v,new}}, \\ C_{3,new} = ND^{s_{v,new}} = g^{s_{v,new} f(\gamma, I_{\mathfrak{R}_{new}})}. \end{cases} \tag{13}$$

– **Step 3**: DO sends the update message $CUP = \{I_{\mathfrak{R}_{new}}, C_{0,new}, C_{1,new}, C_{3,new}\}$ to ES. ES obtains the new middle ciphertext $CT_{m,new} = \{CUP, C_2\}$. Finally, ES stores the updated ciphertext $CT_{new} = \{Sd(ED_\kappa), CT_{m,new}\}$.

CUP contains only four elements, because n is invariant, the length of $I_{\mathfrak{R}_{new}}$is n, which is constant. As discussed above, the lengths of $C_{0,new}, C_{1,new}, C_{3,new}$ are the same as in CT. Thus, the length of the update message CUP is independent of the number of attributes and is constant. We only need to update four elements in ciphertext. In this way, APU can significantly reduce the overhead of updating access policies.

4.4. Correctness Analysis

We first analyze the correctness of the decryption. Set the access policy of the ciphertext CT as $\mathfrak{R}$, set the attributes structure of DU as I_{DU}. Let $I_{\mathfrak{R}} \subseteq I_{DU}$. AA checks:

$$\begin{aligned} e(C_3, g^{-1}) e(g, C_1^{f(\gamma, I_{\mathfrak{R}})}) &= e((\prod_{i=0}^{n} coe_i^{e_i})^{s_v}, g^{-1}) e(g, g^{s_v f(\gamma, I_{\mathfrak{R}})}) \\ &= e(g^{s_v f(\gamma, I_{\mathfrak{R}})}, g^{-1}) e(g, g^{r f(\gamma, I_{\mathfrak{R}})}) \\ &= e(g,g)^{-s_v f(\gamma, I_{\mathfrak{R}})} e(g,g)^{s_v f(\gamma, I_{\mathfrak{R}})} \\ &= e(g,g)^{s_v f(\gamma, I_{\mathfrak{R}}) - s_v f(\gamma, I_{\mathfrak{R}})} \\ &= 1. \end{aligned}$$

The Eq. 7 holds, AA computes:

$$\begin{aligned} skd &= sk_{DU2} \cdot sk_{DU3}^{(f(\gamma, I_{\mathfrak{R}}) - f(\gamma, I_{DU}))/q} \\ &= g^{d_{DU_2} + d_{DU_1} f(\gamma, I_{DU})} \cdot g^{q d_{DU_1} \cdot (f(\gamma, I_{\mathfrak{R}}) - f(\gamma, I_{DU}))/q} \\ &= g^{d_{DU_2} + d_{DU_1} f(\gamma, I_{DU})} \cdot g^{d_{DU_1} \cdot f(\gamma, I_{\mathfrak{R}}) - d_{DU_1} \cdot f(\gamma, I_{DU})} \\ &= g^{d_{DU_2} + d_{DU_1} f(\gamma, I_{DU}) + d_{DU_1} \cdot f(\gamma, I_{\mathfrak{R}}) - d_{DU_1} \cdot f(\gamma, I_{DU})} \\ &= g^{d_{DU_2} + d_{DU_1} \cdot f(\gamma, I_{\mathfrak{R}})}, \end{aligned}$$

and

$$\begin{aligned} dm_1 &= e(skd, C_1^q / C_2) \\ &= e(g^{d_{DU_2}+d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})}, g^{qs_v}/g^{qs_1}) \\ &= e(g^{d_{DU_2}+d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})}, g^{q(s_v-s_1)}) \\ &= e(g,g)^{q(s_v-s_1)d_{DU_2}+q(s_v-s_1)d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})} \\ &= e(g,g)^{qs_2 d_{DU_2}+qs_2 d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})}. \end{aligned}$$

Then, DU computes:

$$\begin{aligned} dm &= \frac{e(skd, C_2)\cdot dm_1}{e(sk_{DU3}, C_3)} \\ &= \frac{e(g^{d_{DU_2}+d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})}, g^{qs_1})\cdot e(g,g)^{qs_2 d_{DU_2}+qs_2 d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})}}{e(g^{qd_{DU_1}}, (\prod_{i=0}^{n} coe_i^{e_i})^{s_v})} \\ &= \frac{e(g,g)^{qs_1 d_{DU_2}+qs_1 d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})}\cdot e(g,g)^{qs_2 d_{DU_2}+qs_2 d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})}}{e(g^{qd_{DU_1}}, g^{s_v f(\gamma, I_{\mathfrak{R}})})} \\ &= \frac{e(g,g)^{q(s_1+s_2)d_{DU_2}+q(s_1+s_2)d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})}}{e(g,g)^{qs_v d_{DU_1} f(\gamma, I_{\mathfrak{R}})}} \\ &= \frac{e(g,g)^{qs_v d_{DU_2}+qs_v d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})}}{e(g,g)^{qs_v d_{DU_1} f(\gamma, I_{\mathfrak{R}})}} \\ &= e(g,g)^{qs_v d_{DU_2}}, \end{aligned}$$

and

$$\begin{aligned} \kappa &= \frac{C_0 \cdot dm}{e(C_1, sk_{DU1})} \\ &= \frac{\kappa \cdot e(g,g)^{\alpha s_v}\cdot e(g,g)^{qs_v d_{DU_2}}}{e(g^{s_v}, g^{\alpha+qd_{DU_2}})} \\ &= \frac{\kappa \cdot e(g,g)^{\alpha s_v+qs_v d_{DU_2}}}{e(g,g)^{\alpha s_v+qs_v d_{DU_2}}} \\ &= \kappa. \end{aligned}$$

Thus, DU that satisfy the access policy of the ciphertext can get the symmetric key κ and use it to decrypt ED_κ to get the plaintext M.

Second, we analyze the correctness of the APU algorithm. Assume that DO generates a new access policy $\mathfrak{R}_{new}$ for CT and generates a new secret value $s_{v,new}$. AA computes $ND = g^{f(\gamma, I_{\mathfrak{R}_{new}})}$, DO computes $\{C_{0,new} = \kappa \cdot e(g,g)^{\alpha s_{v,new}}, C_{1,new} = g^{s_{v,new}}, C_{3,new} = ND^{s_{v,new}}\}$. Thus, the new ciphertext $CT_{new} = \{Sd(ED_\kappa), I_{\mathfrak{R}_{new}}, C_{0,new}, C_{1,new}, C_2, C_{3,new}\}$. If DU meets $\mathfrak{R}$ but does not satisfy the new access policy $\mathfrak{R}_{new}$, AA returns $\perp$ to DU. If DU uses SK_{DU} and the old $\{skd, dm_1\}$ to decrypt the

updated ciphertext CT_{new}, then

$$\begin{aligned}
&\frac{e(skd, C_2) \cdot dm_1}{e(sk_{DU3}, C_{3,new})} \\
&= \frac{e(g^{d_{DU_2}+d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})}, g^{qs_1}) \cdot e(g,g)^{qs_2 d_{DU_2}+qs_2 d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})}}{e(g^{qd_{DU_1}}, ND^{s_{v,new}})} \\
&= \frac{e(g,g)^{qs_1 d_{DU_2}+qs_1 d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})} \cdot e(g,g)^{qs_2 d_{DU_2}+qs_2 d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})}}{e(g^{qd_{DU_1}}, g^{s_{v,new} f(\gamma, I_{\mathfrak{R}_{new}})})} \\
&= \frac{e(g,g)^{q(s_1+s_2)d_{DU_2}+q(s_1+s_2)d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})}}{e(g,g)^{qs_{v,new} d_{DU_1} f(\gamma, I_{\mathfrak{R}_{new}})}} \\
&= e(g,g)^{qs_v d_{DU_2}+qs_v d_{DU_1}\cdot f(\gamma, I_{\mathfrak{R}})-qs_{v,new} d_{DU_1} f(\gamma, I_{\mathfrak{R}_{new}})} \neq e(g,g)^{qs_{v,new} d_{DU_2}},
\end{aligned}$$

DU cannot obtain κ. If DU_A satisfies $\mathfrak{R}_{new}$, then $I_{\mathfrak{R}_{new}} \subseteq I_{DU_A}$. Let $SK_{DU_A} = \{I_{DU_A}, sk_{DU_{A,1}} = g^{\alpha+qd_{DU_{A,2}}}, sk_{DU_{A,2}} = g^{d_{DU_{A,2}}+d_{DU_{A,1}} f(\gamma, I_{DU_A})}, sk_{DU_{A,3}} = g^{qd_{DU_{A,1}}}\}$ be the secret key of DU_A, AA implicitly defines $s_{2,new} = s_{v,new} - s_1$ and computes $skd_{new} = sk_{DU_{A,2}} \cdot sk_{DU_{A,3}}^{(f(\gamma, I_{\mathfrak{R}_{new}})-f(\gamma, I_{DU_A}))/q} = g^{d_{DU_{A,2}}+d_{DU_{A,1}}\cdot f(\gamma, I_{\mathfrak{R}_{new}})}, dm_{1,new} = e(skd_{new}, C_{1,new}^q / C_2) = e(g,g)^{qs_{2,new} d_{DU_{A,2}}+qs_{2,new} d_{DU_{A,1}}\cdot f(\gamma, I_{\mathfrak{R}_{new}})}$. DU_A computes

$$\begin{aligned}
dm_{new} &= \frac{e(skd_{new}, C_2) \cdot dm_{1,new}}{e(sk_{DU_{A,3}}, C_{3,new})} \\
&= \frac{e(g^{d_{DU_{A,2}}+d_{DU_{A,1}}\cdot f(\gamma, I_{\mathfrak{R}_{new}})}, g^{qs_1}) \cdot e(g,g)^{qs_{2,new} d_{DU_{A,2}}+qs_{2,new} d_{DU_{A,1}}\cdot f(\gamma, I_{\mathfrak{R}_{new}})}}{e(g^{qd_{DU_{A,1}}}, ND^{s_{v,new}})} \\
&= \frac{e(g,g)^{q(s_1+s_{2,new})d_{DU_{A,2}}+q(s_1+s_{2,new})d_{DU_{A,1}}\cdot f(\gamma, I_{\mathfrak{R}_{new}})} \cdot}{e(g^{qd_{DU_{A,1}}}, g^{s_{v,new} f(\gamma, I_{\mathfrak{R}_{new}})})} \\
&= \frac{e(g,g)^{qs_{v,new} d_{DU_{A,2}}+qs_{v,new} d_{DU_{A,1}}\cdot f(\gamma, I_{\mathfrak{R}_{new}})} \cdot}{e(g,g)^{qs_{v,new} d_{DU_{A,1}} f(\gamma, I_{\mathfrak{R}_{new}})}} \\
&= e(g,g)^{qs_{v,new} d_{DU_{A,2}}},
\end{aligned}$$

and computes

$$\begin{aligned}
&\frac{C_{0,new} \cdot dm_{new}}{e(C_{1,new}, sk_{DU_{A,1}})} \\
&= \frac{\kappa \cdot e(g,g)^{\alpha s_{v,new}} \cdot e(g,g)^{qs_{v,new} d_{DU_{A,2}}}}{e(g^{s_{v,new}}, g^{\alpha+qd_{DU_{A,2}}})} \\
&= \frac{\kappa \cdot e(g,g)^{\alpha s_{v,new}+qs_{v,new} d_{DU_{A,2}}}}{e(g,g)^{\alpha s_{v,new}+qs_{v,new} d_{DU_{A,2}}}} \\
&= \kappa.
\end{aligned}$$

Thus, the APU algorithm can successfully update the ciphertext, the updated ciphertext can be decrypted normally by DU that meet the updated access policy, and can prevent DU that meet the old access policy but do not meet the updated access policy from decrypting the updated ciphertext.

5. Security Analysis

The main goal of the security analysis in our FPCA is to achieve indistinguishability of the data and to resist collusion attacks on the secret key and the ciphertext. We combined the Discrete Logarithm (DL) problem, the Decisional Diffie-Hellman (DDH) problem, the Decisional Bilinear Diffie-Hellman (DBDH) problem, and the selective game for ABE for security analysis. The hash function H is a random oracle. In security analysis, adversaries do not have valid secret keys to decrypt the challenge ciphertext or the updated ciphertext. The adversary can apply for multiple invalid keys and attempt to synthesize valid keys using these invalid keys. Adversaries can share their invalid secret keys with other adversaries and freely obtain the ciphertext from ES. As ES is semi-trusted, the adversary can also conspire with ES.

5.1. Formal Security Analysis

Theorem 1: Based on the DBDH problem, if an adversary A can break FPCA in probabilistic polynomial time (PPT) with an advantage Ω that is non-negligible, then a PPT simulator $\mathfrak{B}$ can be constructed to break the DBDH problem with an advantage $\Omega/2$.

Proof: The challenger CH is trusted, CH can correctly execute FPCA, CH is responsible for responding to adversary A's query and generating challenge ciphertext, the purpose of CH is to maintain the security of the selective game. Given the bilinear pairing group $\{p, g, G, G^*\}$, CH randomly chooses $\{a, b, c \in Z_p\}$ and generates a DBDH instance $DI = \{g, g^a, g^b, g^c, Z\}$, CH randomly chooses $\mu \in \{0, 1\}$, if $\mu = 0$, then $Z = e(g, g)^{abc}$, otherwise, Z is a random number in G^*. CH sends DI to the simulator $\mathfrak{B}$. $\mathfrak{B}$ can only use the DBDH instance to simulate FPCA. $\mathfrak{B}$ plays the role of CH in the interaction with A, but $\mathfrak{B}$ cannot obtain $\{a, b, c\}$. $\mathfrak{B}$ must be able to respond to all valid key queries, and the keys to which it responds must match the real key distribution. The challenge ciphertext generated by $\mathfrak{B}$ must be indistinguishable from the real encryption. The purpose of $\mathfrak{B}$ is to solve the DBDH problem. A submits two messages of equal length and challenge access policy, A can continuously request secret keys from $\mathfrak{B}$, but none of these secret keys can satisfy the challenge access policy, the purpose of A is to break the security of FPCA.

1. ***Init***: A submits a challenge access policy $\mathfrak{R}1$ to $\mathfrak{B}$, $I_{\mathfrak{R}1} = \hbar_{1\mathfrak{R}1}...\hbar_{n\mathfrak{R}1}$ is the access structure of $\mathfrak{R}1$.
2. ***Setup***: $\mathfrak{B}$ randomly chooses $\{\alpha_1, \gamma_c\} \in Z_p$ and computes $CoE_c = \{coe_{ci} = g^{\gamma_c^i} | i \in [1, n]\}$. $\mathfrak{B}$ computes $Y = e(g, g)^{\alpha_1} e(g, g)^{ab} = e(g, g)^{\alpha_1 + ab} = e(g, g)^{\alpha}$. $\mathfrak{B}$ sets $g^{q_c} = g^b$ and sets $BI_c = \{Y, g^{q_c}\}$ and hash function $H : \{0, 1\}^* \to Z_p^*$. $\mathfrak{B}$ outputs the public parameters to A:

$$PP = (BPG, e(\cdot), H, g, p, CoE_c, BI_c).$$

3. ***Key Query 1***: After A sends the key request and an attributes set U_A to $\mathfrak{B}$, $\mathfrak{R}1 \not\subseteq U_A$, $I_A = \hbar_{1A}...\hbar_{nA}$ is the access structure of U_A. $\mathfrak{B}$ randomly chooses $d_{c1} \in Z_p$ and

chooses $d_{c2}^{'} \in Z_p$, then $\mathfrak{B}$ implicitly defines $d_{c2} = d_{c2}^{'} - a$. $\mathfrak{B}$ computes

$$\begin{cases} sk_{s1} = g^{\alpha_1 + bd_{c2}^{'}} = g^{\alpha_1 + ab + b(d_{c2}^{'} - a)} = g^{\alpha + q_c d_{c2}}, \\ sk_{s2} = \dfrac{g^{d_{c2}^{'}}}{g^a} \cdot g^{d_{c1} f(\gamma_c, I_A)} = g^{d_{c2} + d_{c1} f(\gamma_c, I_A)}, \\ sk_{s3} = g^{bd_{c1}} = g^{q_c d_{c1}}. \end{cases}$$

$\mathfrak{B}$ sends the simulated secret key $SK_A = \{I_A, sk_{s1}, sk_{s2}, sk_{s3}\}$ to A, $\mathfrak{R}\mathfrak{1}$ cannot be satisfied by any secret key requested by A from $\mathfrak{B}$.
In the real secret key, $\{d_{A1}, d_{A2}\} \in_R Z_p$ are chosen by AA, $\{sk_1 = g^{\alpha + q_c d_{A1}}, sk_2 = g^{d_{A2} + d_{A1} f(\gamma_c, I_A)}, sk_3 = g^{q_c d_{A1}}\} \in G$. In SK_A, because $\{d_{c1}, d_{c2}\} \in_R Z_p$, $\{sk_{s1}, sk_{s2}, sk_{s3}\} \in G$ and the distributions of $\{d_{A1}, d_{A2}\}$ and $\{d_{c1}, d_{c2}\}$ in Z_p are identical, therefore, the distributions of $\{sk_1, sk_2, sk_3\}$ and $\{sk_{s1}, sk_{s2}, sk_{s3}\}$ in G are identical. A does not know $\{\alpha, \gamma_c, q_c, d_{c1}, d_{c2}\}$, therefore A cannot distinguish between the simulated secret key SK_A and the real secret key, in the view of A, SK_A is real.

4. ***Challenge***: A sends two plaintexts $\{M_0, M_1\}$ to $\mathfrak{B}$ as the challenge query, the lengths of $\{M_0, M_1\}$ are the same. $\mathfrak{B}$ randomly chooses $\beta \in \{0, 1\}$ and $s_1 \in Z_p$. $\mathfrak{B}$ sets $g^{s_c} = g^c$, s_c is the secret value. $\mathfrak{B}$ computes

$$\begin{cases} C_{c0} = M_\beta \cdot Z \cdot e(g, g)^{\alpha_1 c}, \\ C_{c1} = g^c = g^{s_c}, \\ C_{c2} = g^{bs_1} = g^{q_c s_1}, \\ C_{c3} = g^{cf(\gamma_c, I_{\mathfrak{R}\mathfrak{1}})} = g^{s_c f(\gamma_c, I_{\mathfrak{R}\mathfrak{1}})}, \end{cases}$$

and sends the challenge ciphertext $CT_c = \{I_{\mathfrak{R}\mathfrak{1}}, C_{c0}, C_{c1}, C_{c2}, C_{c3}\}$ to A.
In addition, $\mathfrak{B}$ randomly chooses $s_2 \in Z_p$ and computes

$$\begin{aligned} dm_{1c} &= e(\frac{g^{d_{c2}^{'}}}{g^a} \cdot g^{d_{c1} f(\gamma_c, I_{\mathfrak{R}\mathfrak{1}})}, g^b)^{s_2} \\ &= e(g^{d_{c2} + d_{c1} f(\gamma_c, I_{\mathfrak{R}\mathfrak{1}})}, g^b)^{s_2} \\ &= e(g, g)^{bs_2 d_{c2} + bs_2 d_{c1} f(\gamma_c, I_{\mathfrak{R}\mathfrak{1}})} \\ &= e(g, g)^{q_c s_2 d_{c2} + q_c s_2 d_{c1} f(\gamma_c, I_{\mathfrak{R}\mathfrak{1}})}, \end{aligned}$$

$\mathfrak{B}$ sends dm_{1c} to A.
In real ciphertext, $\{s_v, s_1\} \in_R Z_p$, $C_0 = M_\beta \cdot e(g, g)^{\alpha s_v} \in G^*$ and $\{C_1 = g^{s_v}, C_2 = g^{q_c s_v}, C_3 = g^{s_v f(\gamma, I_{\mathfrak{R}\mathfrak{1}})}\} \in G$, in CT_c, $\{s_c, s_1\} \in_R Z_p$, $\{Z, C_{c0}\} \in G^*$ and $\{C_{c1}, C_{c2}, C_{c3}\} \in G$, because the distributions of $\{s_v, s_1\}$ in real ciphertext and $\{s_c, s_1\}$ in CT_c are identical in Z_p, thus, the distributions of $\{C_1, C_2, C_3\}$ and $\{C_{c1}, C_{c2}, C_{c3}\}$ in G are identical. The distribution of $Z \cdot e(g, g)^{\alpha_1 c}$ and $e(g, g)^{\alpha s_v}$ is identical in G^*, then the distribution of C_0 and C_{c0} is identical in G^*.
The real auxiliary decryption component $dm_1 = e(g, g)^{q_c s_2 d_{A_2} + q_c s_2 d_{A_1} \cdot f(\gamma_c, I_{\mathfrak{R}\mathfrak{1}})}$, as discussed above, $\{s_2, d_{A_1}, d_{A_2}, d_{c1}, d_{c2}\} \in_R Z_p$, $\{dm_1, dm_{1c}\} \in G$, the distributions of dm_1 and dm_{1c} are identical in G. $\{s_c, s_1, s_2, \alpha, \gamma_c, q_c, d_{c1}, d_{c2}, Z\}$ are not

disclosed to A and because $\mathfrak{R}1 \not\subseteq U_A$, A cannot decrypt CT_c correctly by using dm_{1c} and A cannot determine whether $s_1 + s_2$ equals S_c. Therefore, from the view of A, CT_c and dm_{1c} are both real and not simulated.

5. ***Update Ciphertext***: A generates a new challenge access policy $\mathfrak{R}2$, $I_{\mathfrak{R}2} = \hbar_{1\mathfrak{R}2}...\hbar_{n\mathfrak{R}2}$ is the access structure of $\mathfrak{R}2$, $\mathfrak{R}2 \not\subseteq U_A$, $\mathfrak{R}2$ cannot be satisfied by any secret key that A requests from $\mathfrak{B}$. A submits $\{CT_c, \mathfrak{R}2\}$ to $\mathfrak{B}$ as the update ciphertext query. $\mathfrak{B}$ randomly chooses $s_n \in Z_p$ and computes $g^{s_{c,new}} = g^{cs_n}$, $s_{c,new}$ is the new secret value. $\mathfrak{B}$ computes

$$\begin{cases} C_{c0,new} = M_\beta \cdot (Z \cdot e(g,g)^{\alpha_1 c})^{s_n}, \\ C_{c1,new} = g^{cs_n} = g^{s_{c,new}}, \\ C_{c3,new} = g^{cs_n f(\gamma_c, I_{\mathfrak{R}2})} = g^{s_{c,new} f(\gamma_c, I_{\mathfrak{R}2})}, \end{cases}$$

and sends the updated challenge ciphertext $CT_{c,new} = \{I_{\mathfrak{R}2}, C_{c0,new}, C_{c1,new}, C_{c2}, C_{c3,new}\}$ to A.
In real updated ciphertext, $s_{v,new} \in_R Z_p$, $C_{0,new} \in G^*$, $\{C_{1,new}, C_{3,new}\} \in G$, in the updated challenge ciphertext, $Z \in G^*$, $s_n \in_R Z_p$, and $c = s_c \in_R Z_p$, then $s_{c,new} = cs_n \in_R Z_p$, $C_{c0,new} \in G^*$, $\{C_{c1,new}, C_{c3,new}\} \in G$. Because A does not obtain $\{s_c, s_n, \gamma_c, \alpha_1, Z\}$, the distributions of $\{C_{c1,new}, C_{c3,new}\}$ and $\{C_{1,new}, C_{3,new}\}$ are identical in G, and the distribution of $\{C_{c0,new}\}$ and $\{C_{0,new}\}$ are identical in G^*, thus, in the view of A, $CT_{c,new}$ is real.
6. ***Key Query 2***: Repeat **Key Query 1**.
7. ***Guess***: A outputs $\beta_1 \in \{0,1\}$ as a guess of β, $\mathfrak{B}$ outputs $\mu_1 \in \{0,1\}$ as a guess of μ. If $\beta_1 = \beta$, then $\mathfrak{B}$ outputs $\mu_1 = 0$ to guess $Z = e(g,g)^{abc}$, otherwise, $\mathfrak{B}$ outputs $\mu_1 = 1$ to guess Z is a random number in G^*.

If $\mu = 0$, then $C_{c0} = M_\beta \cdot e(g,g)^{abc+\alpha_1 c} = M_\beta \cdot e(g,g)^{\alpha s_c}$ and $C_{c0,new} = M_\beta \cdot e(g,g)^{(abc+\alpha_1 c)s_n} = M_\beta \cdot e(g,g)^{\alpha s_{c,new}}$ are valid ciphertext components. A has the advantage Ω of guessing β correctly. In this case, the probability of $\mathfrak{B}$ guessing correctly is $Adv^0_{\mathfrak{B}} = P[\beta_1 = \beta | \mu = 0] = P[\mu_1 = \mu | \mu = 0] = \Omega + 1/2$.

If $\mu = 1$, then Z is a random number in G^*. Thus, C_{c0} and $C_{c0,new}$ are random numbers in G^* in the view of A. A has no advantage in guessing β correctly. In this case, the probability of $\mathfrak{B}$ guessing correctly is $Adv^1_{\mathfrak{B}} = P[\beta_1 \neq \beta | \mu = 1] = P[\mu_1 = \mu | \mu = 1] = 1/2$.

The advantage of $\mathfrak{B}$ in breaking the DBDH problem is

$$\begin{aligned} Adv^{DBDH}_{\mathfrak{B}} &= (Adv^1_{\mathfrak{B}} + Adv^0_{\mathfrak{B}} - 1)/2 \\ &= (1/2 + \Omega + 1/2 - 1)/2 \\ &= \Omega/2. \end{aligned}$$

Thus, the **Theorem 1** is proved.

5.2. Informal Security Analysis

1) Resist Key Forgery Attack:

Assume that there is a ciphertext $CT_{I_{\mathfrak{R}}}$ under the access policy $\mathfrak{R}$, $I_{\mathfrak{R}}$ is the access structure of $\mathfrak{R}$. The secret key of A is $SK_A = \{sk_{A1} = g^{\alpha + qd_{A2}}, sk_{A2} = g^{d_{A2} + d_{A1} f(\gamma, I_A)}$,

$sk_{A3} = g^{qd_{A1}}\}$, $I_{\mathfrak{R}} \not\subseteq I_A$. A attempts to decrypt $CT_{I_{\mathfrak{R}}}$ by forging a valid secret key that meets the access policy $\mathfrak{R}$. To achieve this, A must calculate $sk_{\mathfrak{R},2} = g^{d_{A2}+d_{A1}f(\gamma,I_{\mathfrak{R}})}$. Because, $\{\gamma, q, d_{A1}, d_{A2}\}$ are unknown to A. A cannot calculate $f(\gamma, I_{\mathfrak{R}})$ without γ and cannot calculate $\{g^{d_{A1}}, g^{d_{A2}}\}$ without q. According to the DDH problem, even if A can obtain $g^{d_{A1}f(\gamma,I_{\mathfrak{R}})}$ and calculate $g^{f(\gamma,I_{\mathfrak{R}})}$, but A cannot calculate $g^{d_{A1}f(\gamma,I_{\mathfrak{R}})}$ by $g^{d_{A1}}$ and $g^{f(\gamma,I_{\mathfrak{R}})}$. Thus, A cannot calculate $sk_{\mathfrak{R},2} = g^{d_{A2}+d_{A1}f(\gamma,I_{\mathfrak{R}})}$, our FPCA can resist key forgery attacks.

2) Resist Collusion Attack:

Between multiple adversaries: Assume that adversaries A_1 and A_2 have secret keys $SK_{A1} = \{sk_{A11} = g^{\alpha+qd_{A12}}, sk_{A12} = g^{d_{A12}+d_{A11}f(\gamma,I_{A1})}, sk_{A13} = g^{qd_{A11}}\}$ and $SK_{A2} = \{sk_{A21} = g^{\alpha+qd_{A22}}, sk_{A22} = g^{d_{A22}+d_{A21}f(\gamma,I_{A2})}, sk_{A23} = g^{qd_{A21}}\}$, respectively. Both I_{A_1} and I_{A_2} do not meet the access policy $\mathfrak{R}$ of the ciphertext $CT_{I_{\mathfrak{R}}}$, but $I_{\mathfrak{R}} \subseteq I_{A_1} + I_{A_2}$. A_1 and A_2 attempted to derive $sk_{A12,col} = g^{d_{A12}+d_{A11}f(\gamma,I_{A_1}+I_{A_2})}$ to replace sk_{A12} to decrypt $CT_{I_{\mathfrak{R}}}$ by collusion attack. To achieve this, A_1 and A_2 must obtain d_{A11}, $g^{d_{A11}}$ or $f(\gamma, I_{A_1} + I_{A_2})$. But $\{\gamma, q, d_{A1}, d_{A2}\}$ are unknown to A_1 and A_2. According to the analysis of resisting key forgery attack, the DL problem and the DDH problem, A_1 and A_2 cannot obtain d_{A11}, $g^{d_{A11}}$ or $f(\gamma, I_{A_1} + I_{A_2})$ and cannot calculate $g^{d_{A12}+d_{A11}f(\gamma,I_{A_1}+I_{A_2})}$. Thus, multiple adversaries cannot forge a valid secret key to decrypt the ciphertext by collusion attack.

Between ES and adversary: Because ES is semi-trusted, adversary A may temper with the ciphertext stored in ES by conspiring with ES. Assume that ciphertext $CT_{I_{\mathfrak{R}}} = \{I_{\mathfrak{R}}, C_0 = \kappa \cdot e(g,g)^{\alpha s_v}, C_1 = g^{s_v}, C_2 = g^{qs_1}, C_3 = g^{s_v f(\gamma,I_{\mathfrak{R}})}\}$, the secret key of A is $SK_A = \{sk_{A1} = g^{\alpha+qd_{A2}}, sk_{A2} = g^{d_{A2}+d_{A1}f(\gamma,I_A)}, sk_{A3} = g^{qd_{A1}}\}$, $I_{\mathfrak{R}} \not\subseteq I_A$. A attempts to replace $\{I_{\mathfrak{R}}, C_3\}$ by $\{I_A, C_{A3} = g^{s_v f(\gamma,I_A)}\}$, so that A can decrypt $CT_{I_{\mathfrak{R}}}$ by SK_A. A has g^{s_v} and $g^{f(\gamma,I_A)}$, but γ and s_v are unknown to A, A cannot calculate $f(\gamma, I_A)$. According to the DDH problem, A cannot calculate $C_{A3} = g^{s_v f(\gamma,I_A)}$ by g^{s_v} and $g^{f(\gamma,I_A)}$ and according to the DL problem A cannot calculate $\{f(\gamma, I_A), s_v\}$ by and $g^{f(\gamma,I_A)}$ and g^{s_v}. According to Eq. 7, $e(C_3, g^{-1})e(g, C_1^{f(\gamma,I_A)}) \neq 1$, AA will return $\perp$ to A. Thus, A cannot decrypt $CT_{I_{\mathfrak{R}}}$. Thus, the adversary cannot forge a valid ciphertext by conspiring with ES.

3) Resist Chosen-Key Attack

Assume that adversaries A obtained multiple secret keys with different attribute set: $\{SK_{Ai} = \{sk_{A1} = g^{\alpha+qd_{A2}}, sk_{A2} = g^{d_{A2}+d_{A1}f(\gamma,I_{Ai})}, sk_{A3} = g^{qd_{A1}}\}|i \in [1, L]\}$, L is the number of secret keys that A obtained. None of $\{I_{Ai}|i \in [1, L]\}$ meets the access policy $\mathfrak{R}$ of ciphertext $CT_{I_{\mathfrak{R}}}$, but $I_{\mathfrak{R}} \subseteq \sum_{i=1}^{L} I_{Ai}$. Thus, A attempts to derive a valid secret key by calculating $sk_{A3,val} = g^{d_{A2}+d_{A1}f(\gamma,\sum_{i=1}^{L} I_{Ai})}$ to decrypt $CT_{I_{\mathfrak{R}}}$. To obtain $sk_{A3,val}$, A first needs to calculate $g^{d_{A1}f(\gamma,\sum_{i=1}^{L} I_{Ai})}$. A has $g^{qd_{A1}}$ and $g^{f(\gamma,\sum_{i=1}^{L} I_{Ai})}$, but $\{\gamma, q, d_{A1}, d_{A2}, f(\gamma, \sum_{i=1}^{L} I_{Ai})\}$ are unknown to A. According to the DL problem, A cannot calculate $\{q, d_{A1}, f(\gamma, \sum_{i=1}^{L} I_{Ai})\}$ by g^q, $g^{qd_{A1}}$ and $g^{f(\gamma,\sum_{i=1}^{L} I_{Ai})}$, and according to the DDH problem, A cannot calculate $g^{qd_{A1}f(\gamma,\sum_{i=1}^{L} I_{Ai})}$ by $g^{qd_{A1}}$ and $g^{f(\gamma,\sum_{i=1}^{L} I_{Ai})}$, and cannot calculate $g^{d_{A1}f(\gamma,\sum_{i=1}^{L} I_{Ai})}$. Thus, A cannot obtain $sk_{A3,val}$, our FPCA can resist the chosen-key attack.

4) FPCA Satisfies Backward Security

Assume DO wants to revoke the access permission of A to the ciphertext $CT_{I_A} = \{I_A, C_0 = \kappa \cdot e(g,g)^{\alpha s_o}, C_1 = g^{s_o}, C_2 = g^{qs_1}, C_3 = g^{s_o f(\gamma, I_A)}\}$, DO changes κ to $\kappa 1$ and performs the APU algorithm to update the ciphertext CT_{I_A} to $CT_{I_\Re} = \{I_\Re, C_0 = \kappa 1 \cdot e(g,g)^{\alpha s_v}, C_1 = g^{s_v}, C_2 = g^{qs_1}, C_3 = g^{s_v f(\gamma, I_\Re)}\}$, $I_\Re \not\subseteq I_A$, A cannot use the secret key $SK_A = \{sk_{A1} = g^{\alpha + qd_{A2}}, sk_{A2} = g^{d_{A2} + d_{A1} f(\gamma, I_A)}, sk_{A3} = g^{qd_{A1}}\}$ to decrypt $CT_{I_\Re}$. From the above analysis, we can see that A cannot calculate $sk_{A2,new} = g^{d_{A2} + d_{A1} f(\gamma, I_\Re)}$ or $C_{3,new} = g^{s_v f(\gamma, I_A)}$ for decryption. To obtain $\kappa 1$, A has to calculate $e(g,g)^{\alpha s_v}$. A has $e(g,g)^{\alpha}$ and g^{s_v}, but α and s_v are unknown to A. Thus, according to the DBDH problem, A cannot calculate $e(g,g)^{\alpha s_v}$ by $e(g,g)^{\alpha}$ and g^{s_v}. A does not know g^{α}, then A cannot calculate $e(g,g)^{\alpha s_v}$ by the bilinear pairing operation. Thus, A cannot obtain $e(g,g)^{\alpha s_v}$. In addition, A may try to calculate $\kappa 1 \cdot (e(g,g)^{\alpha s_v})^{s_o / s_v} = \kappa 1 \cdot e(g,g)^{\alpha s_o}$ and use $e(g,g)^{\alpha s_o}$ to obtain $\kappa 1$. But, according to the DL problem, A cannot calculate $\{s_o, s_v\}$ by $\{g^{s_o}, g^{s_v}\}$ and A does not have $e(g,g)^{\alpha s_v}$. Thus, A cannot calculate $\kappa 1 \cdot e(g,g)^{\alpha s_o}$ to get $\kappa 1$, A cannot decrypt the updated ciphertext by using the old secret key, our FPCA satisfies the backward security.

6. Performance Analysis and Evaluation

6.1. Experiment Setting

We perform simulation experiments for the performance evaluation of FPCA on storage, communication, and computation costs. We choose Sun *et al.* [20], Fan *et al.* [6], Li *et al.* [10], and Wu *et al.* [2] as comparison schemes. We implemented servers including CA, AA, ES, and CS under the Win 11 64-bit operating system with Intel(R)core(TM) i3-10105 3.70GHz, 12GB memory. We deploy the IoT network on Win 11 64-bit operating system with 2.5 GHz AMD A10-9620P RADEON R5 and 8GB of memory. We perform the simulation experiment under the jdk1.8.0-152 environment with the JPBC library.

In this experiment, we set the Type-A curve $E(F^v) : y^2 = x^3 + x$ with an order p of 160 $bits$ and $|v| = 512$ $bits$ for [20], [6], [10], [2] and FPCA. We choose G and G^* from the subgroups of $E(F^v)$. As the lengths of $|G|$ and $|G^*|$ are 1024 $bits$, we use $|G|$ to represent the length of the element in G or G^*. Table 3 shows the parameters required for the experiment.

For the data access across data domain simulation experiment, we set that there are two data domains in the system, $NDD = 2$, DU in the data domain Dm_1 access data from the data domain Dm_2, both Dm_1 and Dm_2 have 50 attributes, $n = 100$, and $NA_{C1} = NA_{C2} = 20$.

6.2. Theoretical Analysis

Regarding the theoretical analysis for FPCA with other comparison schemes, we first perform the feature analysis in Table 1. [20], [6], [10], and FPCA support cross data domain access control. Only FPCA and [2] can achieve constant ciphertext length for better access control efficiency, but [2] cannot achieve constant secret key length, while FPCA maintains the secret key length constant. FPCA, [6], and [10] can update the access policy. Only FPCA can maintain the update message in a constant length. Thus, we choose

Table 3. Experiment Parameters

Parameter	Description
NA	Number of attributes
NA_{DU}	Number of attributes for the DU
NA_T	Number of attributes in the access policy
NA_{msa}	Minimum number of attributes that satisfy the access policy for DU
NA_{add}	Number of attributes added to access policy
NA_{rev}	Number of attributes revoked from access policy
NDD	Number of data domains
NA_{Ui}	Number of attributes of DU in data domain Dm_i
T_G	The running time of multiplication operations in group
T_{EXP}	The running time of modular exponentiation operation
T_{BP}	The running time of bilinear pairing

[20], [6], and [10] as compared schemes for cross data domain access control, we choose [6] and [10] as compared schemes for access policy updates, and we choose [2] as a compared scheme for constant-length ABE.

We have summarized the complexity of FPCA and other compared schemes in Table 6, where KG denotes key generation, EN is encryption, DE is decryption, $NA_B = NA_{DU} + NA_T$, $NA_E = NA_T + NA_T + NA_{add} - NA_{rev}$, $NA_C = NA_T + NA_{msa}$.

Table 4 shows the storage costs. Public parameters PP are stored by almost all entities in the access control system. The length of PP in FPCA is less than in other compared schemes. In addition, the ciphertext is stored by DO, DU and ES, the secret key is stored by DU and AA, and the update message is stored by DO and ES. AA in FPCA also stores VM, which is the verification message of FPCA, the storage costs of VM are $2n + 5|G|$. According to Table 6, since n is invariant, the storage complexity of the secret key, the ciphertext and the update message for FPCA are $O(1)$, while for [20], [6] and [10] are $O(NA_{DU})$, $O(NA_T)$ and $O(NA_E - NA_T)$, respectively. For [2], the storage complexity of the ciphertext is $O(1)$, but the storage complexity of the secret key is $O(NA_{DU})$, which is higher than FPCA. The length of the ciphertext in [2] is $4|G| - n$ longer than in FPCA. Therefore, the storage costs of FPCA are lower than those of other compared schemes.

In addition, we also present the communication costs in Table 4. In the system initialization phase, CA sends the public parameters PP to all other entities. In the key generation phase, AA sends the secret key to DU. In the encryption phase, DO sends the middle ciphertext to ES, and DO sends the verification message to ES in [2]. In the decryption phase, ES sends the ciphertext to DU. In FPCA, VM, skd, and dm_1 are also transmitted between AA and DU when decryption. In the access policy update phase, DO sends the update message to ES, and in FPCA, ND and the access structure $I_{\mathfrak{R}_{new}}$ of the new access policy $\mathfrak{R}_{new}$ are also transmitted between AA and DO. The lengths of messages transmitted in FPCA are constant. Thus, according to Table 6, the communication complexity in FPCA is $O(1)$ and is lower than [20], [6], and [10]. Although the communication complexity of the ciphertext in [2] is $O(1)$, the sum of the lengths of the

Table 4. The Comparison of Storage and Communication Costs

Scheme	Public Parameters Length	Secret Key Length	Ciphertext Length	Update Message Length
[20]	$\|Z_p\| + 3\|G\| + (7+3n)\|G\|$	$8\|G\| + 3\|G\| \cdot NA_{DU}$	$4\|G\| + 2\|G\| \cdot NA_T$	/
[6]	$(\|Z_p\| + \|G\|) \cdot (1+n) + 3\|G\|$	$8\|G\| \cdot NA_{DU}$	$4\|G\| + (5\|G\| + 4\|Z_p\|) \cdot NA_T$	$\|G\| + 4\|Z_p\| \cdot (NA_T - NA_{rev}) + (2\|G\| + 3\|Z_p\|) \cdot NA_{add}$
[10]	$6\|G\| + (\|Z_p\| + 2\|G\|) \cdot n$	$(NA_{DU}+2) \cdot 3\|G\|$	$2\|G\| + 5\|G\| \cdot NA_T + \|Z_p\|$	$4\|G\| \cdot NA_T + 2\|G\| \cdot (NA_T - NA_{rev} + NA_{add})$
[2]	$6\|G\| + (\|Z_p\| + \|G\|) \cdot n$	$2\|G\| + 2NA_{DU} \cdot (\|G\| + \|Z_p\|)$	$8\|G\|$	/
FPCA	$\|Z_p\| + 3\|G\| + n \cdot \|G\|$	$n + 3\|G\|$	$n + 4\|G\|$	$n + 3\|G\|$

Table 5. The Comparison of Computational Costs

Scheme	Key Generation	Encryption	Decryption	Access Policy Update
[20]	$4T_{EXP} + T_{BP} + (6T_{EXP} + 3T_G) \cdot (NA_{DU}+1)$	$4T_{EXP} + (8T_{EXP} + 5T_G) \cdot (NA_T + 1)$	$6T_{BP} + 4T_{EXP} + 4T_G + (3T_{BP} + T_{EXP} + 2T_G) \cdot NA_{msa}$	/
[6]	$(12T_{EXP} + 3T_G) \cdot NA_{DU} + T_{EXP}$	$T_{BP} + 4T_{EXP} + T_G + (8T_{EXP} + 2T_{BP} + T_G) \cdot NA_T$	$NA_T \cdot (3T_{EXP} + 5T_G) + 5T_{BP} - 4T_G + NA_{msa} \cdot (5T_{EXP} + 3T_{BP} + 6T_G)$	$T_{EXP} + T_G + (NA_T - NA_{rev} + NA_{add}) \cdot (5T_{EXP} + 3T_G + T_{BP})$
[10]	$11T_{EXP} + T_G + NA_{DU} \cdot (3T_{EXP} + T_G)$	$3T_{EXP} + 2T_G + (1 + NA_T) \cdot (4T_{EXP} + T_G) + 3 \cdot T_{BP}$	$NA_{msa} \cdot (4T_{EXP} + 3T_{BP} + 3T_G) + 2T_{BP} + 4T_{EXP} + T_G$	$4T_{EXP} + T_{BP} + T_G + 2(NA_T - NA_{rev} + NA_{add}) \cdot (T_{EXP} + T_G)$
[2]	$3T_{EXP} \cdot (1 + NA_{DU}) + T_G \cdot (2NA_{DU} + 1)$	$(2T_{EXP} + 2T_G + T_{BP}) \cdot NA_T + T_{EXP}$	$2T_{BP} + NA_{msa} \cdot (2T_{BP} + T_G)$	/
FPCA	$3T_{EXP}$	$4T_{EXP} + NA_T \cdot (T_{EXP} + T_G) + T_G$	$6T_{BP} + 4T_G + 3T_{EXP}$	$4T_{EXP} + 2T_G$

Table 6. The Comparison of Complexity

Scheme	Storage Complexity					Communication Complexity			Computational Complexity			
	CA	AA	DU	DO	ES	AA & DU	DO & ES	DU & ES	KG	EN	DE	AUP
[20]	$O(1)$	$O(NA_{DU})$	$O(NA_B)$	$O(NA_T)$	$O(NA_T)$	$O(NA_{DU})$	$O(NA_T)$	$O(NA_T)$	$O(NA_{DU})$	$O(NA_T)$	$O(NA_{msa})$	/
[6]	$O(1)$	$O(NA_{DU})$	$O(NA_B)$	$O(NA_T)$	$O(NA_E)$	$O(NA_{DU})$	$O(NA_E)$	$O(NA_T)$	$O(NA_{DU})$	$O(NA_T)$	$O(NA_C)$	$O(NA_E)$
[10]	$O(1)$	$O(NA_{DU})$	$O(NA_B)$	$O(NA_T)$	$O(NA_E)$	$O(NA_{DU})$	$O(NA_E)$	$O(NA_T)$	$O(NA_{DU})$	$O(NA_T)$	$O(NA_{msa})$	$O(NA_E)$
[2]	$O(1)$	$O(NA_{DU})$	$O(NA_{DU})$	$O(1)$	$O(1)$	$O(NA_{DU})$	$O(1)$	$O(1)$	$O(NA_{DU})$	$O(NA_T)$	$O(NA_{msa})$	/
FPCA	$O(1)$	$O(1)$	$O(1)$	$O(1)$	$O(1)$	$O(1)$	$O(1)$	$O(1)$	$O(1)$	$O(NA_T)$	$O(1)$	$O(1)$

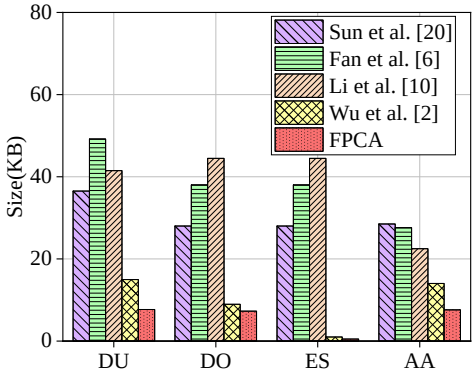


Fig. 4. The Storage Costs

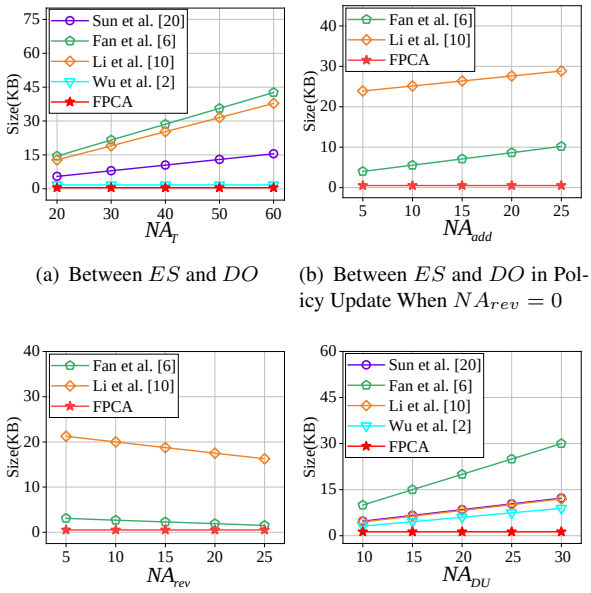
ciphertext and the update message in FPCA is lower than the ciphertext in [2]. Thus, the communication costs of FPCA are lower than other compared schemes.

In Table 5, we present the computation cost. We omit the computation costs for the hash operation and the computation in Z_p, which are negligible. We also omit the symmetric encryption/decryption operation, which has the same running time in all the compared schemes. The multiplication operations in G and G^* have similar processing times, so we use T_G uniformly to represent the multiplication operations of group elements. According to Table 6, the computational complexity of FPCA is $O(1)$ in stages other than encryption, while for other schemes compared it is $O(NA)$. In the encryption phase, the computational complexities of all schemes compared are $O(NA_T)$, only [20] and FPCA does not require bilinear pairing operations, and the computational costs of FPCA are lower than those of [20] by $NA_T \cdot (7T_{EXP} + 4T_G) + 8T_{EXP} + 4T_G$. Therefore, the computational costs of our FPCA are lower than those of other compared schemes.

The cross data domain access for FPCA and other compared schemes primarily involves the transmission and decryption of ciphertext from Dm_2. The complexities of decryption and transmission of ciphertext have been discussed above. In addition, [20] needs to perform conversion calculations on the secret key with computational complexity of $O(NA_{U1})$ before decryption. Thus, the computational complexity of FPCA is lower than other schemes compared for cross data domain access.

6.3. Experimental results

In Fig. 4, we evaluated the storage costs of each entity in FPCA and other compared schemes. We ignore the storage costs of CA and CS because CA only stores the public parameters PP, which we analyze in all other entities, CS stores the encrypted data ED_κ that have the same storage costs in FPCA and other compared schemes. As we can see from Fig. 4, FPCA have lower storage costs than other compared schemes. The storage costs of DU, DO, ES, and AA are only $7.41\ KB$, $7.27\ KB$, $0.5\ KB$, and $7.35\ KB$ in FPCA and are only 1% to 25% of [20], [6], and [10]. Comparing FPCA and [2], the storage costs of FPCA are only 50% to 80% of [2]. This is because the length of PP, the secret key, the ciphertext, and the update message in FPCA are constant.



(a) Between ES and DO

(b) Between ES and DO in Policy Update When $NA_{rev} = 0$

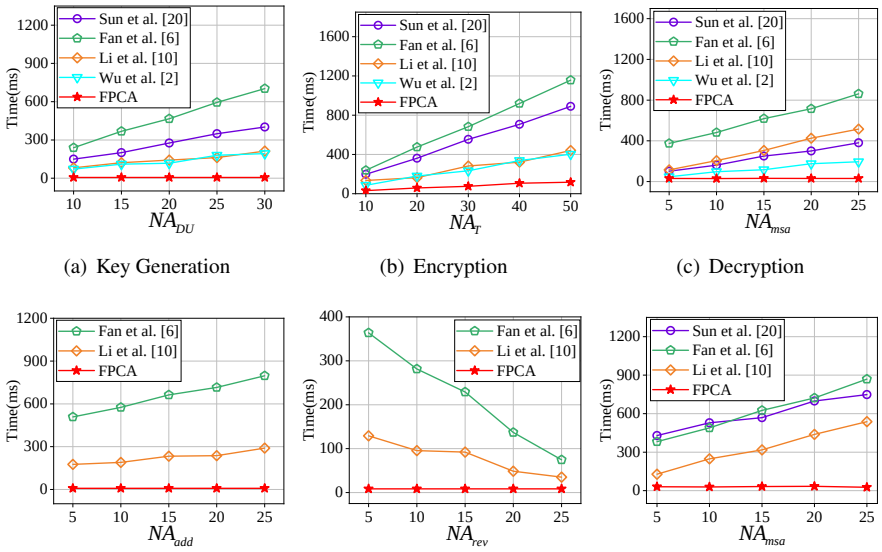
(c) Between ES and DO in Policy Update When $NA_{add} = 0$

(d) Between AA and DU

Fig. 5. The Communication Costs

In Fig. 5, we evaluated the communication costs between the entities in FPCA as well as other compared schemes. We ignore the communication with CA, because it only transmits PP, which we have already analyzed in storage costs, and the length of PP is not significantly different in the FPCA and compared schemes. Figs. 5(a), 5(b) and 5(c) show the communication costs between ES and DO, where FPCA remains at $0.5\,KB$, $0.38\ KB$, and $0.38\ KB$, respectively, and are lower than other comparative schemes. The communication costs between ES and DU are consistent with Fig. 5(a). Fig. 5(d) shows the communication costs between AA and DU, FPCA remains $1.27\ KB$ and is lower than other schemes compared. This is because in FPCA, the lengths of transmitted messages are constant and the length of the ciphertext is lower than [2].

Fig. 6 shows the time costs of various stages of access control in our FPCA and other compared schemes. Except for Figs. 6(a) and 6(b), we keep NA_{DU} and NA_T unchanged. Fig. 6(a) shows the time costs of key generation for FPCA and the compared schemes, FPCA remains around $5.55\ ms$. Fig. 6(b) shows the time costs of encryption for FPCA and the compared schemes, where the time costs of FPCA increase from $31.4\ ms$ to $117.4\ ms$ as NA_T grows. Fig. 6(c) shows the time costs of decryption for FPCA and the compared schemes, FPCA remains around $29.75\ ms$ and is lower than other compared schemes. Figs. 6(d) and 6(e) show the time costs of access policy update for FPCA and the compared schemes, the average running time of FPCA remains around $8.4\ ms$. The time costs of various stages of access control for FPCA are lower than those of other compared schemes. This is because the computational complexities of key generation, decryption, and access policy update for FPCA are $O(1)$, while for other schemes are



(a) Key Generation (b) Encryption (c) Decryption

(d) Access Policy Update When $NA_{rev} = 0$ (e) Access Policy Update When $NA_{add} = 0$ (f) Cross Data Domain Access

Fig. 6. The Time Costs

$O(NA)$. For encryption, although the computational complexity for all the schemes compared is $O(NA_T)$, but only FPCA and [20] have no bilinear pairing operation, and the computational costs of FPCA are lower than [20].

In Fig. 6(f), we evaluated the time costs of accessing data from another data domain. We can see that when NA_{msa} increases, the time costs of FPCA remain around $30.4\ ms$, while for [20], [6], and [10] are increase and higher than FPCA. This is because in FPCA, the computational complexity of decrypting ciphertext from Dm_2 is $O(1)$, and the lengths of the ciphertext and other transmitted messages are constant.

In summary, the experimental results show that the storage, communication, and computation costs of FPCA are lower than those of other compared schemes in all phases of cross data domain access control. Compared with different access control schemes, FPCA has a significant advantage in increasing the efficiency of cross data domain access control and updating the access policy.

6.4. Engineering applications

Cloud-edge collaboration technology finds widespread application in the field of new energy vehicle manufacturing, enabling large-scale data processing and real-time response services for new energy vehicle production. When combined with data access control technology, it allows new energy vehicle manufacturing enterprises to safely share production data. These production data are generated and utilized by numerous IoT devices in various business lines. However, the production of new energy vehicles involves multiple distinct business lines, each belonging to different data domains and possessing unique

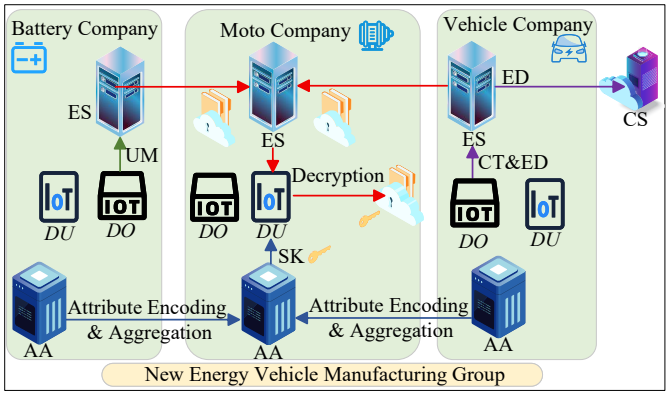


Fig. 7. Application Scenario of FPCA in New Energy Vehicle Manufacturing Industry

access control attribute sets. Among data sharing, DU needs multiple secret keys to access data from different business lines. The large number of attributes can lead to an expansion in secret key and ciphertext lengths, and unable to update data access authority flexibly. These impose a high computational resource overhead on resource-constrained IoT devices in business lines, thereby limiting production efficiency. FPCA enables data access for multi-business lines using one secret key while maintaining constant ciphertext and secret key length, and also reduces the overhead of updating data access authority, achieves dynamic data sharing across business lines, and improves collaborative production efficiency, as shown in Fig. 7.

In Fig. 7, the new energy vehicle manufacturing group comprises a battery company, a motor company, and a vehicle company, each engaged in different businesses. Each company represents a data domain and is equipped with an AA and an ES. Both DU and DO utilize IoT devices on the production line. In Fig. 7, when shared data is generated during the vehicle process, DO of the vehicle company first sets an access policy, then encrypts the vehicle production data to be shared, and uses attribute encoding to produce a constant-length ciphertext CT. DO uploads CT and encrypted data ED to ES of the vehicle company. ES stores CT and uploads ED to the central CS for storage. After registration by DU of the motor company, AA of the motor company generates a constant-length secret key SK containing the attributes of the three companies using attribute encoding and aggregation. DU can use SK to decrypt shared data from various companies, which DU obtains through ES of the motor company. When DO of the battery company wants to adjust the access policy of battery production data, it uses attribute encoding to generate an update message UM of constant-length and uploads it to ES to replace the corresponding part in the ciphertext, thus completing the update.

7. Conclusions and Future Work

In this work, we present FPCA, a fully constant-length and policy-updating cross data domain access control for a cloud-edge collaborative environment. To improve the efficiency of cross data domain access control, we design the MDKG and CLCE algorithms to keep

the length of the secret key and the ciphertext constant and allow the user to access data from other data domains without any conversion operation. We design the APU algorithm to update the access policy with a constant-length update message and low computational complexity to achieve dynamic access control. The security analysis proved the security of our FPCA. The experiment results indicate that our FPCA can achieve efficient cross data domain access control.

However, FPCA cannot hide the access policy, considering that access policies contain user privacy, which may be leaked to adversaries. Moreover, most of the encryption and decryption are performed by IoT devices of IoT users, which are resource-constrained. The low computational efficiency of IoT users will limit the improvement of access control efficiency. Thus, in the future, we will consider adding hidden access policy functionality to FPCA to address the issue of user privacy leakage. Finally, to improve the efficiency of FPCA, we will consider outsourcing most of the encryption and decryption calculations to servers in FPCA.

Acknowledgement. This work is supported by the Scientific Research Fund of Hunan Provincial Education Department (No.24A0337), and the Natural Science Foundation of Hunan Province (No.2025JJ50348 and 2025JJ50399).

References

1. Allison, L., Brent, W.: Unbounded hibe and attribute-based encryption. Lecture Notes in Computer Science 6632, 547–567 (2011)
2. Axin, W., Yinghui, Z., Jianhao, Z., Qiuxia, Z., Yu, Z.: Hierarchal bilateral access control with constant size ciphertexts for mobile cloud computing. IEEE Transactions on Cloud Computing 12(2), 659–670 (2024)
3. Bingcheng, J., Qian, H., Peng, L., Sabita, M., Yan, Z.: Blockchain empowered secure video sharing with access control for vehicular edge computing. IEEE Transactions on Intelligent Transportation Systems 24(9), 9041–905 (2023)
4. Chen, J., Lu, F., Liu, Y., Peng, S., Cai, Z., Mo, F.: Cross trust: A decentralized ma-abe mechanism for cross-border identity authentication. International Journal of Critical Infrastructure Protection 44, 100661 (2024)
5. Chen, L., Chen, Y., Liang, W., Li, X., Li, K.C., Wang, J., Xiong, N.: Mass: A multi-attribute sketch secure data sharing scheme for iot wearable medical devices based on blockchain. IEEE Internet of Things Journal 12(2), 1990–2001 (2025)
6. Fan, H., Li, Q., Xiong, J., Li, R., Chen, W., Huang, H.: Decentralized access control for privacy-preserving cloud-based personal health record with verifiable policy update. IEEE Internet of Things Journal 11(9), 16887 – 16901 (2024)
7. Fan, Y., Liu, S., Tan, G., Lin, X.: Cscac: One constant-size cpabe access control scheme in trusted execution environment. International Journal of Computational Science and Engineering 19(2), 162–168 (2019)
8. Huai, Q., Yuan, W., Wu, Y., Fan, P.: Downlink otfs-rsma cross-domain transmission scheme and sum-rate maximization. IEEE Communications Letter 29(3), 600–604 (2025)
9. Khandla, D., Shahy, H., Bz, M.K., Pais, A.R., Raj, N.: Expressive cp-abe scheme satisfying constant-size keys and ciphertexts. Cryptology ePrint Archive 2019, 1257 (2019)
10. Li, J., Zhang, E., Han, J., Zhang, Y., Shen, J.: Ph-mg-abe: A flexible policy-hidden multi-group attribute-based encryption scheme for secure cloud storage. IEEE Internet of Things Journal 12(2), 2146 – 2157 (2024)

11. Meng, X., Liang, W., Xu, Z., Li, K.C., Khan, M.K., Kui, X.: An anonymous authenticated group key agreement scheme for transfer learning edge services systems. ACM Transactions on Sensor Networks 20(3), 1–23 (2024)
12. Mosteiro-Sanchez, A., Barcelo, M., Astorga, J., Urbieta, A.: End to end secure data exchange in value chains with dynamic policy updates. Future Generation Computer Systems 158, 333–345 (2024)
13. Papatsimouli, M., Lazaridis, L., Ziouzios, D., Dasygenis, M., Fragulis, G.: Internet of things (iot) awareness in greece. SHS Web of Conferences. EDP Sciences 139, 03013 (2022)
14. S., A.S., Han, R., Rudolph, C., Grobler, M.: Dacp: Enforcing a dynamic access control policy in cross-domain environments. Computer Networks 237, 110049 (2023)
15. Shiwen, Z., Jiayi, H., Wei, L., Keqin, L.: Mmds: A secure and verifiable multimedia data search scheme for cloud-assisted edge computing. Future Generation Computer Systems 151, 32–44 (2024)
16. Shiwen, Z., Yibin, Y., Wei, L., Arthur, S.V.K., Guoqi, X., Kim-Kwang, R.: Mkss: An effective multi-authority keyword search scheme for edge-cloud collaboration. Journal of Systems Architecture 144, 102998 (2023)
17. Shiwen, Z., Ziwei, Y., Wei, L., Kuan-Ching, L., Ciprian, D.: Baka: Biometric authentication and key agreement scheme based on fuzzy extractor for wireless body area networks. IEEE Internet of Things Journal 11(3), 5118–5128 (2023)
18. Shiwen, Z., Ziwei, Y., Wei, L., Kuan-Ching, L., Di Martino, B.: Bcae: A blockchain-based cross domain authentication scheme for edge computing. IEEE Internet of Things Journal 11(13), 24035–24048 (2024)
19. Su, X., An, L., Cheng, Z., Weng, Y.: Cloud–edge collaboration-based bi-level optimal scheduling for intelligent healthcare systems. Future Generation Computer Systems 141, 28–39 (2023)
20. Sun, J., Xu, G., Li, H., Zhang, T., Wu, C., Yang, X.: Sanitizable cross-domain access control with policy-driven dynamic authorization. IEEE Transactions on Dependable and Secure Computing pp. 1–17 (2025)
21. Tianqiao, Z., Mingming, J., Fucai, L., Yuyan, G.: A lattice-based puncturable cp-abe scheme with forward security for cloud-assisted iot. IEEE Internet of Things Journal 12(14), 26538–26554 (2025)
22. Vipul, G., Omkant, P., Amit, S., Brent, W.: Attribute-based encryption for fine-grained access control of encrypted data. Proceedings of The 13th ACM Conference on Computer and Communications Security pp. 89–98 (2006)
23. Wang, P.B., Li, K.C., Shi, R.H., Shao, B.S.: Vc-dcps: Verifiable cross-domain data collection and privacy-persevering sharing scheme based on lattice in blockchain-enhanced smart grids. IEEE Internet of Things Journal 10(14), 12449–12461 (2023)
24. Xue, J., Shi, L., Zhang, W., Li, W., Zhang, X., Zhou, Y.: Poly-abe: A traceable and revocable fully hidden policy cp-abe scheme for integrated demand response in multi-energy systems. Journal of Systems Architecture 143, 102982 (2023)
25. Yang, M., Wang, H., Wan, Z.: Pul-abe: An efficient and quantum-resistant cp-abe with policy update in cloud storage. IEEE Transactions on Services Computing 17(3), 1126–1139 (2024)
26. Yang, R., He, H., Xu, Y., Xin, B., Wang, Y., Qu, Y., Zhang, W.: Efficient intrusion detection toward iot networks using cloud–edge collaboration. Computer Networks 228, 109724 (2023)
27. Yannis, R., Brent, W.: Practical constructions and new proof methods for large universe attribute-based encryption. Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security pp. 463–474 (2013)
28. Ying, Z., Jiang, W., Liu, X., Xu, S., Deng, R.H.: Reliable policy updating under efficient policy hidden fine-grained access control framework for cloud data sharing. IEEE Transactions on Services Computing 15(6), 3485–3498 (2021)
29. Zhang, S., Ren, F., Liang, W., Li, K., Ling, N.: Gpvo-fl: Grouped privacy-preserving and verification-outsourced federated learning in cloud-edge collaborative environment. IEEE Transactions on Network and Service Management 22(5), 4175 – 4191 (2025)

30. Zuo, Y., Xu, L., Li, J., Li, J., Wang, X., Piran, M.J.: Secure and efficient blockchain-based access control scheme with attribute update. IEEE Transactions on Consumer Electronics 71(1), 1539 – 1550 (2024)

Shiwen Zhang received his B.S. degree in Information and Computing Science from the University of Changsha in 2010, and received his Ph.D. degree from the College of Computer Science and Electronic Engineering, Hunan University, in 2016. He is an associate professor at the School of Computer Science and Engineering, Hunan University of Science and Technology. He is a senior member of IEEE and CCF. His research interests include security and privacy issues in cloud computing, privacy protection, and information security.

Siwei Wen received a B.S. degree from Xiangtan University and is currently pursuing an M.S. degree from the School of Computer Science and Engineering at Hunan University of Science and Technology. His research interests include access control and privacy protection in cloud and edge computing.

Wei Liang received a Ph.D. degree in computer science and technology from Hunan University in 2013. He was a Post-Doctoral Scholar at Lehigh University from 2014 to 2016. He is currently a Professor and the Dean of the School of Computer Science and Engineering at Hunan University of Science and Technology, China. He has authored or co-authored more than 140 journal/conference papers. His research interests include network security, cloud computing, and security management in edge computing.

Received: February 08, 2026; Accepted: February 22, 2026.

Computer Science and Information Systems 23(2):917–946 https://doi.org/10.2298/CSIS251029022A

A framework for the automated thematic annotation of Open Government Data

Abdul Aziz[1], Mohsan Ali[2], Dagoberto José Herrera-Murillo[1], Maria Ioanna Maratsi[2], Francisco J. Lopez-Pellicer[1], and Javier Nogueras-Iso[1]

[1] Aragon Institute of Engineering Research (I3A), Universidad de Zaragoza, Zaragoza, Spain
{abdul.aziz, dherrera, fjlopez}@unizar.es
jnog@unizar.es (corresponding author)
[2] Department of Information and Communication Systems Engineering, University of the Aegean, Samos, Greece
{mohsan, ioanna.m}@aegean.gr

Abstract. Governmental policies for transparency and reuse of public sector information have encouraged the launch of open government data portals around the world. Many of these portals are based on pyramidal structures: national open data portals are aggregators of the contents harvested from open data portals maintained by governments in charge of administrative areas with a narrower scope. Taking into account this hierarchical organization, these open data portals lack consistent and scalable mechanisms for thematic annotation, limiting dataset discoverability. This work proposes a framework for the automated thematic classification of open government data. The framework integrates (i) thematic annotation quality assessment, (ii) supervised machine learning models trained on annotated metadata corpora, and (iii) embedding-based semantic similarity methods for theme assignment in the absence of reliable annotations. The framework is evaluated using 29,793 datasets from *data.europa.eu*, the European open data portal. Experimental results show that supervised models achieve high classification performance, with Support Vector Machines reaching an accuracy of 93.65%, while unsupervised embedding-based approaches achieve substantial semantic agreement with portal-assigned themes (74.56%) using transformer-based representations. These results demonstrate that the proposed framework enables scalable, consistent, and interoperable thematic annotation, offering both theoretical contributions to automated metadata enrichment and practical value for integration into large-scale open data portal infrastructures.

Keywords: Data Annotation, Open Government Data, Open Data Portals, Automated Annotation, Thematic Annotation

1. Introduction

In the era of digital governance, governments worldwide are increasingly adopting open data initiatives to facilitate the access to open government data (OGD) [25,17,33]. OGD, as a subset of the broader concept of open data (data in open format to be shared, used and reused for any purpose), covers a wide range of topics - from budget records to environmental monitoring among other examples. OGD have the potential to enhance transparency, foster citizen participation, support evidence-based policymaking, and drive innovation through the use and reuse of data [34,32,37]. The increased deployment of open

data catalogues and portals has enabled the distribution and straightforward retrieval of substantial quantities of open data in the information-driven society and has leveraged the growth of the open data movement [53,4,7,26].

However, the challenge lies in making sense of this enormous resource. OGD portals are usually organized on the basis of pyramidal structures: national open data portals are aggregators of the contents harvested from open data portals maintained by governments in charge of administrative areas with a narrower scope. This means that open data catalogues must usually integrate heterogeneous metadata records describing datasets that have been harvested from different catalogues with a more local scope. Therefore, often open data catalogues struggle with findability due to limited and inconsistent metadata [7,3].

Taking into account these scenarios, where OGD portals at national or cross-national level must integrate the metadata contents from different sources, it is essential to provide users with classification tools in order to easily locate the information of interest [10]. Information classification schemes generally fall into two main types: exact and ambiguous [38]. Exact organizational schemes precisely divide information into clear and mutually exclusive sections, using methods such as alphabetical, chronological, or geographical sorting. Ambiguous organizational schemes, on the other hand, delimit information into categories that resist precise definition, often due to linguistic ambiguities and human subjectivity. This classification encompasses thematic organization schemas along with other variants such as task-oriented and metaphor-oriented schemas. Thematic organization, in particular, prioritizes the grouping of related items, promoting associative learning, and facilitating adaptive information retrieval strategies.

This work is focused on the study of the thematic annotation of datasets included in OGD catalogues. Rich metadata is one of the core tools to fulfill the FAIR principles [62] and improve the findability, accessibility, interoperability, and reuse of digital assets. To address the findability aspect, one of the almost mandatory recommendations is to include keywords and themes within metadata. This is typically checked by metadata quality evaluation methodologies [30,42,61]. Although national or even cross-national catalogues like *data.europa.eu*, the official European Data Portal, combines metadata compliant with different metadata vocabularies, most of them are derived from DCAT [28,11]. DCAT is the W3C's Data Catalog vocabulary for describing open data [60]. The advantage of using DCAT derived vocabularies is that the thematic annotation of datasets is encouraged thanks to the inclusion of a specific property called *dcat:theme*. Within the metadata, themes provide a higher degree of semantic structure that goes beyond individual keywords and descriptions. Users may conveniently find important information by classifying the datasets according to their applicable themes, independently of the use of exact terminology.

Nevertheless, just the use of metadata vocabularies including a property for thematic annotation is not enough. A missing or wrong thematic annotation hinders the findability. Therefore, we must ensure that the content of this property is accurate. Recognizing the potential of thematic annotation, several researchers have investigated several techniques for annotation, which involves giving significant labels to data. Manual annotation is precise and thorough but is limited in its capacity to scale up [6,64,47]. Automatic annotation exploits machine learning and natural language processing (NLP) to automatically classify information thematically [19]. Using automated annotation may have several advan-

tages. It has the ability to significantly reduce the amount of human work needed, enabling the rapid processing of large datasets. Moreover, it may promote consistency and establish criteria in the annotation process, hence improving the findability and availability of the annotated resources [51].

While there has been substantial progress in publishing OGD, many OGD portals still lack a consistent and scalable mechanism for the thematic annotation of datasets coming from heterogeneous sources, making it difficult for users to discover relevant datasets efficiently. Existing approaches often rely on manual categorization and are inconsistent across portals. This need for an automated mechanism establishes a research niche, which is covered in this work with the design of a framework that encompasses a structured and systematic set of concepts, methods, and software components to guide the process of thematic annotation of OGD. The design of this framework aims to address three main research questions:

1. What is the current state of thematic annotation in open government datasets, and how accurately do these annotations align with the content of the datasets? To answer this question, we have proposed an implementation of the method proposed by Nogueras-Iso et al. [42] for evaluating the thematic classification correctness.
2. Assuming that our collection of datasets (corpus) is properly annotated with themes, to what extent can new datasets be automatically classified? To answer this question, we have tested different machine learning algorithms and preprocessing strategies on an annotated metadata corpus.
3. In the case of not having an annotated corpus, or in the case our corpus is not properly annotated, how can relevant themes be assigned to a dataset from free text metadata? To answer this question, we have tested different strategies based on word-embeddings and sentence-embeddings of metadata to identify the closest theme from a predefined list of themes based on their definitions.

The rest of this paper is structured as follows. Section 2 provides a literature review about the thematic annotation of OGD. Then Section 3 presents our proposed framework for the automated thematic annotation of OGD, which includes the evaluation of the thematic classification correctness in the case of having an existent annotated corpus. Section 4 presents the results after applying the proposed methodology to a corpus of metadata records from the European Data portal (*data.europa.eu*), which are discussed in Section 5. Finally, this paper concludes with a summary of the contributions and some ideas for future work.

2. Related Research

Metadata is a critical component in numerous facets of data management, encompassing the integration, transmission, and transformation of data, among others [43]. As highlighted by the frameworks for assessing the quality of open data portals [30,42,61,49], missing or incomplete metadata hinders the findability of data. A wrongly assigned theme that does not accurately represent the content of a dataset reduces its discoverability and search recall for stakeholders, thereby motivating the need for efficient and accurate thematic annotation and negatively affecting the usability and usefulness of open government

data portals [5]. User-centered studies of open data portals have similarly reported difficulties in dataset discovery and navigation, often linked to metadata quality and categorization practices [41], reinforcing the need for systematic thematic annotation approaches.

Given the importance of providing correct metadata without the burden of accomplishing this task manually, different strategies have been suggested and developed over the years with the aim of achieving a fully or partially automated thematic annotation of resources. Although each strategy emphasizes a unique set of conceptual areas of knowledge and experience, artificial intelligence techniques like machine learning are acquiring an increasing role of portal curators [54,44,1].

As Semantic Web technologies are widely used as a mechanism to publish and reuse open data [15] and metadata is the core of the Semantic Web [58], many research works on automatic annotation are close related to the use of these technologies. For instance, Pavia et al. [45] applied ensemble methods to classify Web-scale datasets through their metadata using a hybrid Recurrent Neural Network composed of LSTM and Bi-directional LSTM units and Naïve Bayes models at a second phase. In a more specific context and regarding bibliographic data, Carducci et al. [9] worked on text categorization for automatic metadata annotation in order to annotate records, separating between philosophical documents and other disciplines. To facilitate this binary classification purpose, they employed NLP and other ensemble learning techniques, integrating domain knowledge and information gained through semantic networks (BabelNet) to decide whether a given document (e.g., thesis) is within the philosophical domain or not. The annotated data is then used to train the chosen supervised learning algorithms and automatically classify the metadata according to the thematic subject of the examined record. Likewise, Verberne et al. [59] investigated the processing and classification of electoral manifestos. After optimizing different parameters including passage segmentation, OCR, or formatting, the results showed that the classifier matches human experts in accuracy and recall.

There are also recent studies focused on OGD portals highlighting the role of automated keyword extraction in enhancing thematic organization and improving findability in open data portals. For instance, Ahmed et al. [2] proposed BRYT, an automated keyword extraction tool that merges and select the most prominent keywords obtained by different techniques based on the statistical distribution of words in the metadata and Large Language Models (LLM). Similarly, Kliimask and Nikiforova [29] introduced TAGIFY, a language model-powered tagging interface aimed at improving data discoverability through enriched metadata in OGD portals. Freire et al. [18] analyzed the use of LLMs in diverse data integration and data discovery tasks, synthesizing recent advancements in this rapidly growing domain. Moreover, Zhang et al. [65] introduced AutoDDG, a framework that is explicitly designed for the automated generation of dataset descriptions for tabular data. AutoDDG utilizes a data-driven methodology to provide a concise summary of dataset content. It employs LLMs to enrich these summaries with semantic information and to generate comprehensible descriptions. These approaches underscore how NLP and semantic tagging can mitigate linguistic ambiguities inherent in ambiguous organizational schemes, thereby supporting more effective associative learning and adaptive retrieval strategies. Huseynov et al. [22] also emphasized the power of NLP to propose a recommender system for datasets. Using the Word2Vec word-embedding technique to encode the free text content of different metadata properties in a vector space, their system provides the users with the possibility of selecting an input dataset and discovering the

recommended datasets with a closer embedding in the vector space. Somehow connected to recommender systems, Bogdanovich et al [8] proposed a method based on Formal Concept Analysis to create a lattice of keywords using as input source the tags for describing datasets under the same thematic category but hosted in different open data portals. This lattice of keywords (concepts) allows cross-portal search of related datasets.

Several attempts for improved annotation services using semantic approaches have also been made in specific data domains such as the biomedical domain. Sasse et al. [52] conducted a literature review on existing semantic metadata annotation services and identified their software requirements in accordance with the FAIR principles: availability as open code; compatibility with common data formats; use of FAIR terminologies; possibility of terminology search; suggestion of annotations; availability of interfaces to external terminologies; and extension of terminologies. Although they concluded that there are not metadata annotation tools that meet all the requirements, this study highlights the importance of annotation tools and the availability of functionalities for suggesting annotations. In a more specific context about the psychiatric and psychological domain, Hudon et al. [20] analyzed the literature on the potential of machine learning to assist in the thematic annotation and classification of text in a psycho-therapeutic context. Their findings demonstrated that, although the existing literature on this specific topic is limited, some techniques such as Support Vector Machine classifiers achieved sufficient accuracy in the performed text classifications, and that this type of classifier is consistently used for classification in the context of medical or clinical text data [21].

Automatic annotation has also been attempted for environmental science metadata. Tuarob et al. [56] aimed to alleviate the problem of environmental metadata harvesting from various and disparate sources with varying levels of metadata quality and curation. They gathered datasets from 4 different archives, selecting for each of them a subset of 1000 annotated documents, and the textual content and attributes of the documents were pre-processed (removal of stop-words, stemming etc.) to obtain a *tf-idf* (term frequency - inverse document frequency) representation. In order to rank automatically candidate themes for the dataset, they used different similarity measures based on cosine similarity and Latent Dirichlet Allocation. Focusing on the processing of images, Ellen et al. [14] targeted plankton image classification using context metadata (such as perimeter, symmetry, temporal and geographic information, etc.) in order to improve the performance of feature-based classifiers. They demonstrated that the inclusion of context metadata might be of substantial gain for classification accuracy in deep learning models, mainly Convolutional Neural Networks. Likewise, Peng et al. [46] proposed a unique biological data classification feature selection method to enhance feature categorization. The technique uses filter and wrapper approaches: it pre-selects feature subsets to improve search efficiency and utilizes ROC curves to assess feature and subset performance. Furthermore, on the viticulture domain, Mylonas et al. [39] proposed a platform for data annotation that includes a thesauri manager for the obtainment of Linked Data Vocabularies. These vocabularies are used in the platform for both manual and automatic annotation based on NLP techniques and supervised learning models such as *k-nearest* neighbors and linear and random forest regression.

When domain-specific research is being carried out, the specificities and domain-sensitive requirements need to be taken into account, to prevent or be aware of in-advance algorithmic biases and limitations. Wu et al. [63] presented the status for automated meta-

data annotation in the cultural heritage domain and discussed the potential of machine learning applications supporting the curating processes of digital artifacts. They provided a summary of recommendations to improve these aspects of automated metadata annotation by leveraging already existing text and images of high quality, utilizing inference of meaning for classification from simple object recognition to tackle metaphoric and symbolic representations in the digital realm, and providing quality indicators on the results to tackle non-uniform and non-consistent automated indexing. Similarly, Ibáñez et al. [23] provided a quantitative analysis of Linked Data in accessible government datasets throughout Europe. They examined the popularity of RDF as a publication format, the accuracy of connected datasets, and the prevalence of established terminologies. Furthermore, the negative effect of poor metadata description on the discoverability of digital cultural heritage artifacts was also addressed by Kaldeli et al. [27] who proposed CrowdHeritage, an ecosystem supportive of end-to-end improvement of metadata utilizing crowdsourcing, machine and human intelligence, semantic, and aggregation techniques.

The framework for thematic annotation proposed in this work integrates the existing knowledge in the state of the art of this field. First, the initial assessment of thematic classification correctness adapts the methodology proposed by Nogueras-Iso et al. [42] to establish quality controls on dataset themes. Second, the supervised classification techniques applied for new datasets in case of having a previously annotated corpus are similar to other works in the literature [45,21]. In addition, the needs for preprocessing and feature representation are similar to other works using free text metadata as input [56]. Third, the unsupervised classification techniques applied in our framework also share some similarities with respect to the works of Ahmed et al. [2], Kliimask and Nikiforova [29] and Huseynov et al. [22] as they also exploit the benefits of using word embeddings and language models. Our proposal compiles all these alternatives within a unified framework, which allows the comparison of the suggested thematic annotations for new datasets in two different scenarios: the existence of a properly annotated corpus; or the unavailability of a properly annotated corpus.

3. Methodology

This section outlines our proposed methodology of the thematic annotation of OGD. Figure 1 shows the general workflow envisioned in this framework. In the case of counting on an annotated corpus, we first need to evaluate the thematic classification correctness before building a machine learning model for the classification of datasets. In contrast, if there is not an available annotated corpus or its classification correctness is not acceptable, we opt for predicting the closest theme measuring the similarity between the word/sentence embeddings of datasets and themes.

3.1. Evaluation of thematic classification correctness

As we need an annotated corpus for the ulterior development of automatic classification models, it is necessary to evaluate first the thematic classification correctness of the corpus. For this evaluation we propose to follow the method proposed by Nogueras-Iso et al. [42]. This method is a customization of the original method proposed by Ureña-Cámara et al. [57], which adapts ISO 19157 standard for geographic information quality

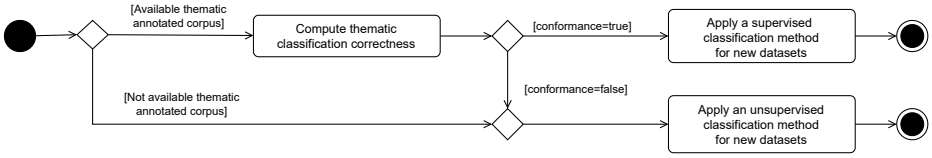


Fig. 1. Workflow for the automated thematic annotation of OGD

.

for the assessment of geographic metadata compliant with the ISO 19115 metadata standard. Nogueras-Iso et al. describe how to perform the assessment of open data metadata compliant with DCAT-based metadata models serialized in RDF. They propose a series of quality controls on six quality categories: completeness, logical consistency, temporal quality, thematic accuracy, positional correctness, and quality of free text. In particular, the thematic accuracy category includes a quality element focused on thematic classification correctness, i.e., the correctness of the thematic keywords and categories included in the metadata with respect to a universe of discourse.

The assessment of the thematic correctness must be made using a sample-based inspection and a *Limiting Quality* index, which determines the sample size (n) according to the corpus size and the maximum number of errors (Ac) that can be accepted to assure a statically equivalent percentage of errors (*Acceptance Quality Limit* or *AQL*) if the full corpus were evaluated. Therefore, the computation of the thematic classification correctness requires the compilation of two associated results: a quantitative result and a conformance result. The quantitative result consists in obtaining a numerical value for the ISO 19157 D.63 measure, which is defined as the number of incorrectly classified records. The conformance result verifies whether the number of errors in the quantitative result surpasses or not the acceptable number of errors (Ac) for the considered sample size.

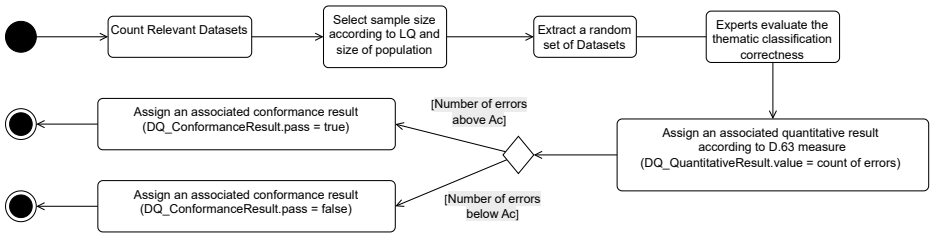


Fig. 2. The workflow for reporting the thematic classification correctness.

Figure 2 shows the workflow that must be followed to compute the thematic classification correctness. In general, the workflow of the assessment starts by considering the whole corpus of datasets of our case, including all the valid and relevant samples. Then, a selected sample size proportionate to the initial population is included to undergo the pre-assessment process, and a random sample set is chosen using a random number gener-

ator. Afterwards, the process of manual assessment of instances from the selected sample is initiated. In order to implement this assessment, we decided that the random sample had to be evaluated by different experts and that this evaluation implied the inspection of the resources associated with the datasets. Then, the experts should assign to them between one and three related themes according to the perceived content of the dataset, its title, its description, and the associated keywords. Then, a consensus should be reached by the experts to consider correct a theme classification if at least one of the assigned themes by the experts corresponded to the initial theme assigned to the dataset. The cases where the initially assigned theme of the dataset does not correspond to the themes assigned by the experts should be annotated as errors. Finally, the associated quantitative and conformance results are assigned.

3.2. Learning to automatically classify based on annotated corpus

Assuming that we count on an annotated corpus of datasets where each dataset has been properly annotated with themes, this component of our annotation framework is focused on building models for the automatic thematic annotation of datasets. For this purpose, we have tested different machine learning algorithms that are typically applied for automatic classification problems in supervised scenarios. Figure 3 shows the workflow followed to build a model for the thematic annotation of OGD thematic annotation. The proposed steps in this workflow are the selection of metadata properties, the normalization of the input text to extract terms, the transformation of the terms into an appropriate feature representation, and the generation of the classification models.

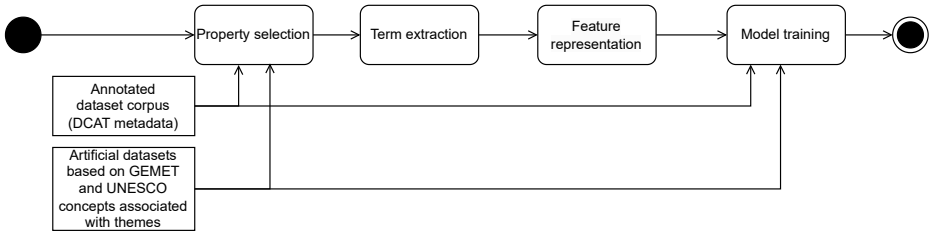


Fig. 3. Proposed method for automated classification of open datasets based on an annotated corpus.

Property selection In this framework, we assume that metadata is compliant with a metadata vocabulary derived from DCAT. As shown in Figure 4, this type of vocabularies include free-text properties for describing a dataset in terms of a title (*dct:title*), a general description (*dct:description*) and several keywords (*dcat:keyword*). In addition, datasets have also an associated theme thanks to the *dcat:theme* property, whose range is a concept from a well-known Knowledge Organization System (KOS) expressed in SKOS format [36]. Therefore, we decided to use the combination of the text provided in *dct:title*, *dct:description* and *dcat:keyword* as input text for the classification. With respect

to the themes or categories to be assigned after the classification process, we assumed in the experiments that the list of themes belonged to the KOS proposed by the European data portal [48], but this is interchangeable with any other KOS if a different corpus of metadata must be classified.

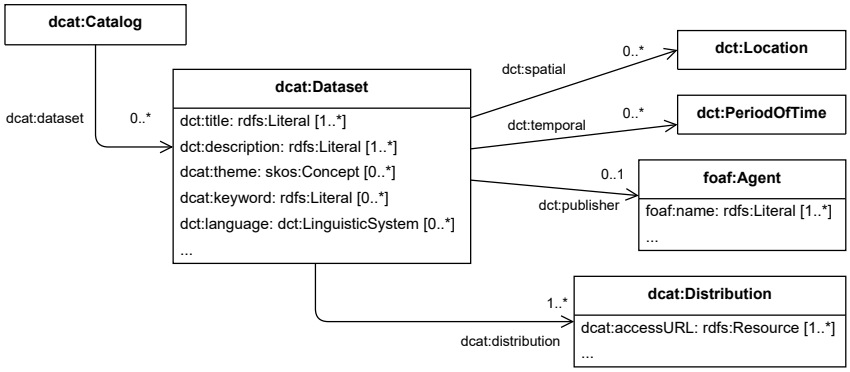


Fig. 4. An excerpt of DCAT-AP metadata model highlighting the free-text elements for describing datasets and its thematic classification (*dcat:theme*). DCAT-AP [16] is one of the main application profiles derived from DCAT for the description of public sector information.

In addition, we also considered the possibility of generating some artificial datasets for each theme to reinforce the classification models to be built. For this research study, we selected GEMET[3] and UNESCO[4] thesauri due to their thematic breadth and multilingual coverage. These thesauri have been widely used for cataloging purposes over the years to facilitate a harmonized thematic classification of datasets and reduce the gap between the vocabulary of data users and data publishers [13,35]. Using the GEMET and the UNESCO thesauri, we aligned the European data themes with the main themes in GEMET and UNESCO (a theme is defined as a micro-thesaurus in UNESCO). Using this alignment, we extracted the preferred labels of the concepts associated with each theme in GEMET and UNESCO thesauri. Dividing the list of associated concepts for each theme in groups of a fixed number of concepts, we converted each group of concepts into an artificial dataset classified with a European data theme and described with the preferred labels of these concepts.

Term extraction The next step in the workflow is the tokenization of the input text describing each dataset and the extraction of terms. For the transformation of tokens into final terms, we have considered a mandatory *basic* level of normalization, and two optional processes of normalization called *translation* and *tailored* normalization.

The mandatory *basic* normalization level incorporates the following processes: stop word removal (i.e., removing common words like ‘a’, ‘an’, ‘the’, etc.), special character

[3] https://www.eionet.europa.eu/gemet/en/themes/
[4] https://vocabularies.unesco.org/browser/thesaurus/en/groups

removal (i.e., removing characters like '$', '%', '&', etc.), link removal (i.e., removing hyperlinks), lowercasing (i.e., converting all text to lowercase) and stemming (i.e., reducing words to their root form).

In addition, we observed that although metadata from OGD catalogues can be downloaded in RDF format and the language of metadata properties can be restricted to a common language such as English, the free-text content frequently appears in other languages. To address this issue, we explored the use of a *translation* normalization approach that employs an API to detect the most likely source language in the free text values and translate them into English, thereby improving consistency.

Last, we also considered a *tailored* normalization to remove noise in the free text derived frequently from spelling mistakes and the use of non-common English words such as acronyms, the names of data provider organizations or other technical terms which only make sense within the context of the data provider organization. For this purpose, there are resources like PyEnchant,[5] which provides access to a dictionary of the English dialects spoken in different regions of the world such as American English, British English, or Australian English and can be used to discard terms not contained in this dictionary.

Feature representation Feature representation is an essential step in our workflow since our objective is to convert unprocessed text input into a vector representation acceptable for our machine learning models. Here, we will explore key features commonly used in text processing, specifically unigrams, bigrams, and trigrams (also called *n-grams*) for all our experiments. Unigrams are typically bag-of-words vector representations where each word is a distinct dimension. Bigrams are vector representations where each dimension is a biword found in the input text. Trigrams are vector representations where each dimension is a distinct trigram. Figure 5 illustrates an example of the unigrams, bigrams, and trigrams that can be generated from an input text.

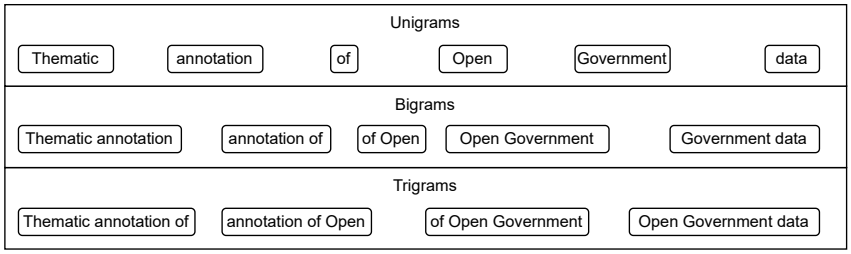


Fig. 5. An example of unigrams, bigrams and trigrams for the sentence "Thematic Annotation of Open Government Data".

N-grams do not require any typical type of calculation performed using equations or formulas. Basically, *n-gram* features are constructed by calculating the word sequences in the corpus. These number of N-grams affect the unique number of features fed into the final model training. For example, unigram features would be fewer in number than if we

[5] https://pypi.org/project/pyenchant/

combined unigram with bigram features. As the number of *n-grams* increases, the vector length (sparsity) increases, which increases the space and time complexity of the model training. On the other hand, there is not a standard way to decide what value of *n* for the *n-grams* will work optimally.

Model training The critical step of the workflow is the training of models where the system learns from the labelled cases (datasets annotated with themes). The models recognize patterns and properties that divide various classes, and this allows to generalize the problem and classify new, unlabeled datasets.

In particular, we have used the One-vs-Rest (OvR) classifier with three machine learning techniques: Logistic Regression (LR), Multinomial Naïve Bayes (MNB), and Support Vector Machines (SVM). The OvR classifier is used in machine learning for situations involving multiple class classifications, because it partitions multi-class problems with more than two classes into a series of binary classification tasks. Each class has its own binary classifier that has been trained to distinguish it from the others. However, data class imbalances may hurt its performance on under-represented groups [55].

Our underlying problem is to classify the open datasets not just into multi-class but also into multiple multi-class themes. For instance, a dataset in the European data portal could be classified into more than one theme of the 13 proposed themes. The OvR classifier can help us to train MNB, SVM, and LR for multiple, multi-class classification.

3.3. Predicting the closest theme of a dataset based on word/sentence embeddings

The objective of this component of the thematic framework is to predict the correct theme when an annotated corpus is unavailable or the datasets in this corpus are not properly annotated. Figure 6 shows the proposed method for the prediction of the closest theme of a dataset based on the similarity between the word/sentence embeddings representing a dataset and its potential associated themes.

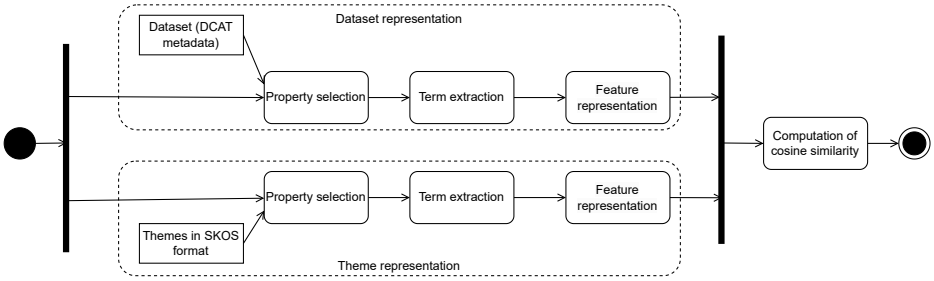


Fig. 6. Proposed method for the prediction of the closest theme of a dataset based on the similarity of word/sentence embeddings

We propose the use of word and sentence embeddings for representing datasets and themes instead of a bag-of-words representation because this allows us to represent similar input texts as close points in a vector space with a number of dimensions much lower

than the number of dimensions needed in vector spaces generated when using bag-of-words representations. Furthermore, the training data and neural networks used to generate these embeddings allow us to find similarities between two texts even in the case of not having any lexical matching between the compared texts.

Next subsections explain the process proposed for the property selection, term extraction, feature representation of both datasets and themes. In addition, we describe how we have computed the cosine similarity between the dataset and theme embeddings.

Property selection In the case of datasets, the property selection is similar to the one proposed for an annotated corpus of datasets in section 3.2. We assume that metadata is compliant with a DCAT vocabulary and we select the content of *dct:title*, *dct:description* and *dcat:keyword*. In addition, we considered two possibilities for generating the input text of a dataset: the concatenation of the three properties, and just the concatenation of title and description. We considered the second possibility because some sentence embeddings may not work properly if we create sentences including disconnected keywords.

In the case of themes, we assume that the list of themes is provided in SKOS format and that each theme is represented with an SKOS concept having an associated preferred label (*skos:prefLabel*) and a definition (*skos:definition*). Therefore, the input text to be processed for each theme is the concatenation of its preferred label and its definition in English.

Term extraction The next step in the proposed method is the extraction of tokens from the input texts for datasets and themes and the generation of terms. In this case we considered a basic normalization consisting in the removal of special characters and the transformation of text to lower case. It must be taken into account that the word/sentence embedding representation avoids implicitly the appearance of non-common English words. In addition, in some cases we also considered the removal of stop words.

Feature representation For the representation of datasets and themes, we considered different possibilities of embeddings:

- Sum of GloVe word embeddings: The terms extracted from the input text of each dataset and theme are converted into a word embedding according to the Global Vectors for Word Representation (GloVe)[6] using vectors of 200 dimensions. To represent the complete input text, this alternative computes the sum of the word embeddings.
- Average of GloVe word embeddings: This alternative is similar to the previous one, but in this case the complete input text is represented with the average of the word embeddings.
- BERT sentence embeddings: This alternative transforms the input text into a vector representation of the sentence by applying the pretrained Bidirectional Encoder Representations from Transformers (BERT) [12].
- HuggingFace sentence embeddings: This alternative transforms the input text into a sentence embedding thanks to HuggingFace representation [31,40].

[6] https://nlp.stanford.edu/projects/glove/

Cosine similarity To find the closest themes that can be associated with a dataset, we propose the use of the cosine similarity distance, which is typically applied to compute the ranking of results in information retrieval systems using a vector space model for representing documents and queries. Equation 1 shows the customization of this cosine distance to our context: the similarity between a theme T and a dataset D is equivalent to the cosine of the angle formed by the vectors $\vec{T}$ and $\vec{D}$ corresponding to their word/sentence embeddings. The similarity is therefore a real value between 0 (least similarity) and 1 (most similarity), which is computed dividing the scalar product of the embedding vectors by the product of their norms.

$$Similarity(T, D) = Cosine(\vec{T}, \vec{D}) = \frac{\vec{T} \cdot \vec{D}}{\|\vec{T}\|\|\vec{D}\|} \tag{1}$$

As the similarity is computed for all datasets that require annotation and all the candidate themes, the output of this step is a matrix where each row represents a dataset and the similarity of each theme is provided in the columns. This way we can generate a rank of associated themes for each dataset, and select, for instance, the top 3 themes.

4. Experiments and results

This section describes the applicability of the thematic annotation framework to a corpus of metadata records downloaded from *data.europa.eu*, the official portal for European OGD. The implementation of the thematic framework (Python programs and notebooks), together with the data and the associated results, are available in Zenodo.[7]

4.1. Corpus description

The metadata used in our experiments came from *data.europa.eu*. This portal serves as a centralized access point to open data published by both European Union institutions and member states. The metadata describing the datasets is compliant with the DCAT-AP vocabulary [16] and can be queried through an SPARQL end-point.[8] In July 2022 we developed a harvester program to download a corpus of 29,793 metadata records in RDF format containing title (*dct:title*), description (*dct:description*), theme (*dcat:theme*) and keyword (*dcat:keyword*) properties. One of the constraints applied to filter the corpus was to have metadata records with at least one associated theme from the list of themes proposed by the European data portal. We also restricted the download to the metadata records declaring the use of English as language, and having at least one title and one description in English.

Figure 7 shows the distribution of the datasets in the corpus among the thirteen thematic categories of the European Data Portal: 'Agriculture, fisheries, forestry and food' (AGRI), 'Economy and finance' (ECON); 'Education, culture and sport' (EDUC), 'Energy' (ENER), 'Environment' (ENVI), 'Government and public sector' (GOVE), 'Health' (HEAL), 'International issues' (INTR), 'Justice, legal system and public safety' (JUST), 'Regions and cities' (REGI), 'Population and society' (SOCI), 'Science and technology'

[7] https://doi.org/10.5281/zenodo.18317554
[8] https://data.europa.eu/sparql

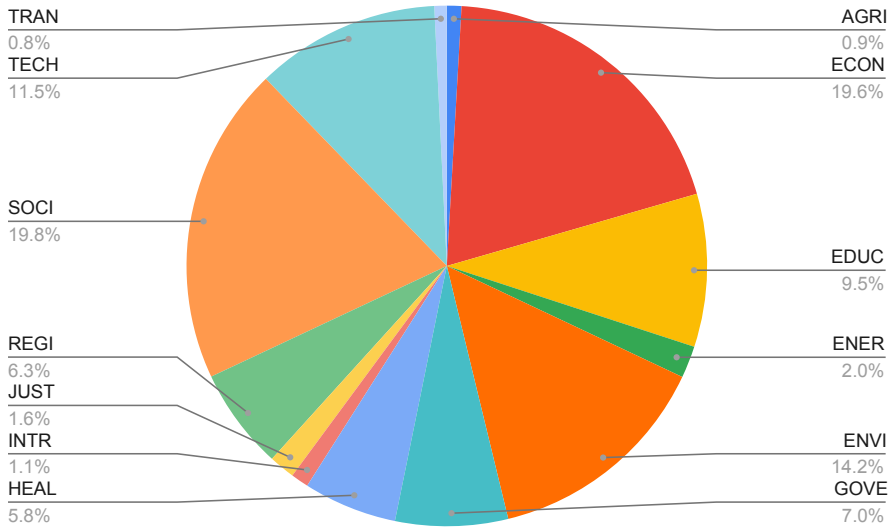


Fig. 7. Distribution of datasets across the 13 themes of the European Data Portal.

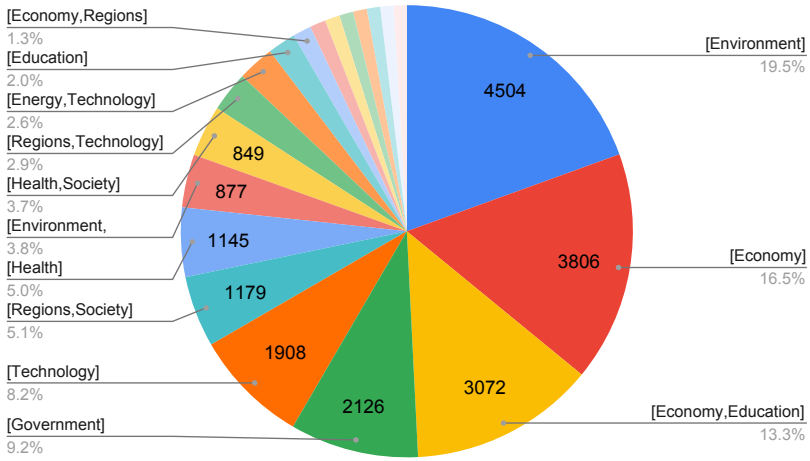


Fig. 8. Distribution of datasets with respect to the 20 most frequent combinations of themes.

(TECH), and 'Transport' (TRAN). In addition to this, as the datasets may be associated with more than one theme, the pie chart shown in Figure 8 illustrates the distribution of the datasets according to the 20 most frequent combination of themes.

Furthermore, after a manual inspection of the records we realized that a significant number of metadata records had metadata properties with text content in a different lan-

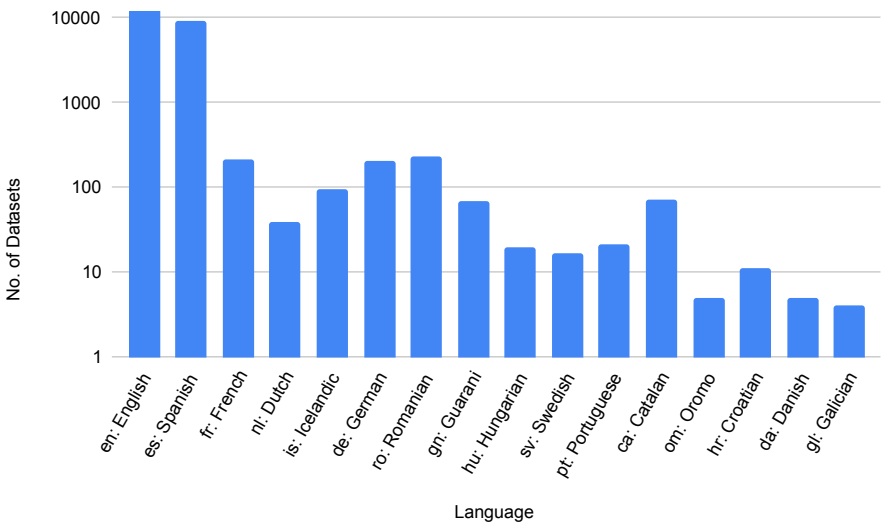


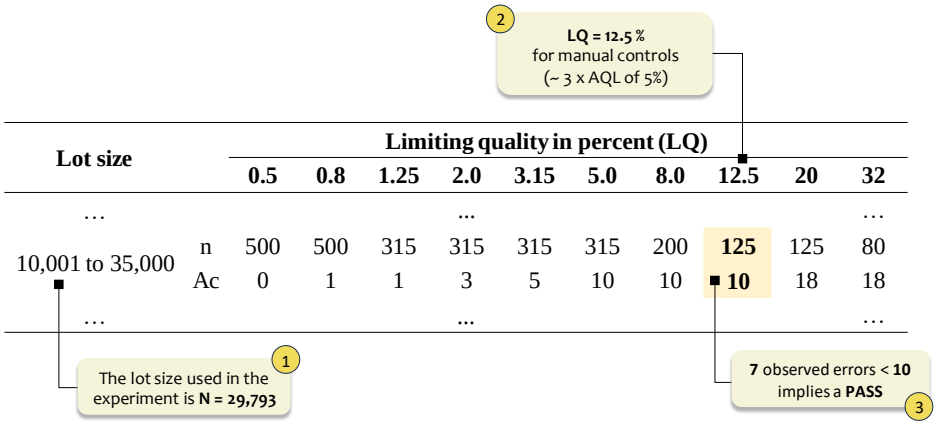
Fig. 9. The language distribution of the datasets.

guage from English. Although metadata records were specifically retrieved declaring the use of English as the language attribute for string literal values, this was not the case for many of the records. Using an API to detect the most likely source language, Figure 9 shows the distribution of languages employed in the corpus. This circumstance motivated the translation of the input text into English as a normalization process during term extraction for some experiments in Section 4.3. In a similar way, it was noticed that 18,633 words found in the input text of the corpus were not recognized as common English words, and this motivated a tailored normalization level for some experiments in Section 4.3 to remove these noise words.

4.2. Results of thematic classification correctness evaluation

Considering an AQL of 5% of errors in the thematic classification of corpus datasets, the Limiting Quality (LQ) that must be applied to a corpus that is manually inspected is thrice the AQL. As the table of ISO 2859-2 standard [24] defining the relationship between the lot size of the corpus and the selected LQ does not provide a value for 15%, the most approximate value of 12.5% must be selected.

Figure 10 describes the process followed to identify the size (n) of the sample that must be evaluated and the maximum number of errors (Ac) that can be accepted in Table A of ISO 2859-2 taking into account that our corpus consists of 29,793 datasets and we want an LQ of 12.5%. Following this, a sample of 125 records was randomly selected. The sample was then evaluated by two experts, who manually visited the dataset resources and assigned to them between one and three related themes according to the perceived content of the dataset and the text contained in title, description, and keyword properties. As indicated in section 3.1, the cases where none of the initially assigned dataset themes



Lot size		Limiting quality in percent (LQ)									
		0.5	0.8	1.25	2.0	3.15	5.0	8.0	12.5	20	32
…					…						…
10,001 to 35,000	n	500	500	315	315	315	315	200	**125**	125	80
	Ac	0	1	1	3	5	10	10	**10**	18	18
…					…						…

Fig. 10. Results of thematic classification correctness for a lot size of 29,793 records and LQ of 12.5%.

matched with one of the themes assigned by the experts were considered as errors. Upon this criterion, only 7 cases of incorrect classification were counted. As the number of errors was below the *Ac* threshold of 10 items, the quality control was passed and the thematic classification of the corpus was considered correct.

4.3. Results of automated supervised classification

This section presents the experiments performed to build models for the automated thematic annotation of datasets using the proposed approach in section 3.2 for supervised classification. Tables 1, 2 and 3 show the description of the 54 experiments that were performed considering different variants for input datasets, term extraction, feature representation and use of machine learning techniques:

- The *Input* column indicates the alternatives used for the input records. The default alternative is the use of the annotated corpus of 29,793 records (denoted as *core*). A second alternative, as proposed in section 3.2, was the incorporation of artificial datasets generated from associated themes in GEMET and UNESCO thesauri. Following this approach, we generated 686 additional records and an extended corpus of 30,479 (denoted as *extended*).
- The *Term extraction* column indicates the alternatives for term extraction: the *basic*, *translation* and *tailored* normalization levels explained in section 3.2.
- The *Feature representation* column indicates the alternatives for feature representation: the use of unigrams (*uni*); the combined use of unigrams and bigrams (*uni+bi*); and the combined use of unigrams, bigrams and trigrams (*uni+bi+tri*). The number of dimensions in the vector representation of each alternative is shown in the tables within parentheses.
- The *Classification technique* column indicates the alternatives for machine learning classification techniques (*LR*, *MNB*, or *SVM*). As indicated in section 3.2, we used the

OvR classifier to solve our multi-class classification problem. When making a prediction, all available binary classifiers are applied to the input data until one produces a confidence score high enough to be considered trustworthy. This method simplifies complex multi-class problems into binary decisions, and it enhances classification performance by focusing on differences between classes [55].

Table 1. Experiments and results for automated supervised classification: *core* input.

#	Input	Term extraction	Feature representation	Classification technique	Accuracy
1	core	basic	uni (35,891)	LR	0.8881
2				MNB	0.7710
3				SVM	0.9365
4	core	basic	uni+bi (290,600)	LR	0.84114
5				MNB	0.57377
6				SVM	0.8995
7	core	basic	uni+bi+tri (659,880)	LR	0.8073
8				MNB	0.5137
9				SVM	0.8503
10	core	basic + translation	uni (25,622)	LR	0.8854
11				MNB	0.7817
12				SVM	0.9355
13	core	basic + translation	uni+bi (265,399)	LR	0.8417
14				MNB	0.5472
15				SVM	0.8920
16	core	basic + translation	uni+bi+tri (619,227)	LR	0.8082
17				MNB	0.4969
18				SVM	0.8394

Tables 1, 2 and 3 also include a column with the accuracy obtained for each experiment. This accuracy is computed according to equation 2 taking into account the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

SVM is the machine learning technique that performed best for all the variants incorporated in the experiments related to the input, term extraction, and feature representation. We also computed the confusion matrices for each theme. For instance, Figure 11 shows the confusion matrices for each individual theme in the best experiment, i.e., experiment 3 in Table 1.

Figure 11 also includes the precision, recall and F1 evaluation metrics according to formulas in equation (3):

$$Precision = \frac{TP}{TP + FP}; Recall = \frac{TP}{TP + FN}; F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

In addition, Figure 12 shows the curve known as the receiver operating characteristic (ROC) for experiment 3. The ROC curve is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR). Whereas TPR reflects the percentage of cases that were properly labelled as positive, FPR reflects the proportion of instances that were incorrectly

Table 2. Experiments and results for automated supervised classification: *extended* input; *basic* and *translation* normalization.

#	Input	Term extraction	Feature representation	Classification technique	Accuracy
19	extended	basic	uni	LR	0.8751
20			(30,443)	MNB	0.7654
21				SVM	0.9226
22	extended	basic	uni+bi	LR	0.8288
23			(280,834)	MNB	0.5656
24				SVM	0.8827
25	extended	basic	uni+bi+tri	LR	0.7959
26			(645,748)	MNB	0.4992
27				SVM	0.8325
28	extended	basic	uni	LR	0.8721
29		+ translation	(26,497)	MNB	0.7623
30				SVM	0.9217
31	extended	basic	uni+bi	LR	0.8298
32		+ translation	(273,849)	MNB	0.5346
33				SVM	0.8754
34	extended	basic	uni+bi+tri	LR	0.7935
35		+ translation	(638,605)	MNB	0.4769
36				SVM	0.8254

Table 3. Experiments and results for automated supervised classification: *extended* input; *tailored* normalization.

#	Input	Term extraction	Feature representation	Classification technique	Accuracy
37	extended	basic	uni	LR	0.8715
38		+ tailored	(17,371)	MNB	0.7737
39				SVM	0.9152
40	extended	basic	uni+bi	LR	0.8248
41		+ tailored	(222,559)	MNB	0.5242
42				SVM	0.8720
43	extended	basic	uni+bi+tri	LR	0.7909
44		+ tailored	(529,092)	MNB	0.4647
45				SVM	0.8201
46	extended	basic	uni	LR	0.8677
47		+ translation	(12,906)	MNB	0.7696
48		+ tailored		SVM	0.9142
49	extended	basic	uni+bi	LR	0.8236
50		+ translation	(215,844)	MNB	0.4971
51		+ tailored		SVM	0.8632
52	extended	basic	uni+bi+tri	LR	0.7886
53		+ translation	(524,646)	MNB	0.4433
54		+ tailored		SVM	0.8090

classified as positive (see formulas in equation 4). It can be observed that the area under the curve (AUC) of the ROC curve is close to the maximum value for practically all of the themes, which demonstrates that the configuration of the SVM experiment has a high probability to assign correctly the theme of a dataset.

$$TPR = \frac{TP}{TP + FN}; FPR = \frac{FP}{FP + TN} \tag{4}$$

4.4. Results of theme prediction

This section presents the results of the approach proposed in section 3.3 to predict automatically the closest theme according to the similarity between the word/sentence em-

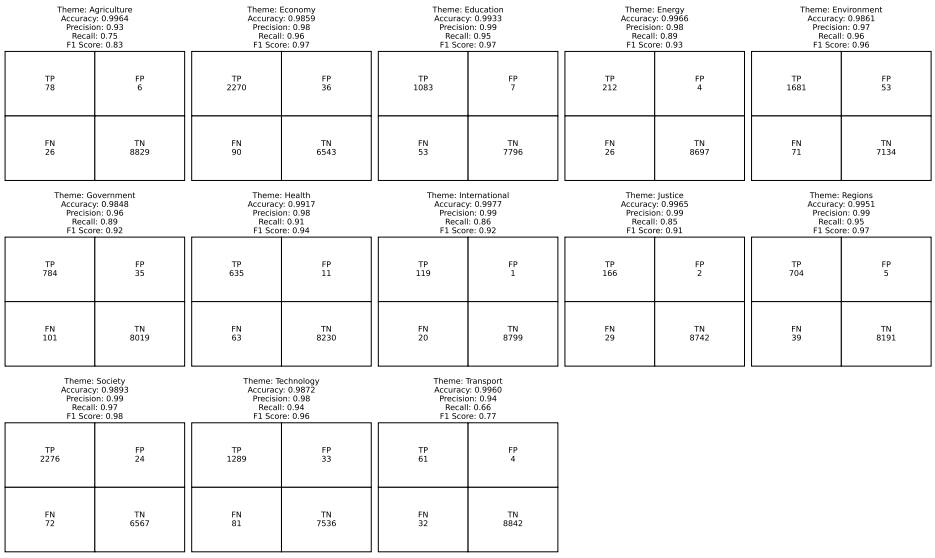


Fig. 11. Confusion matrices for all the themes in experiment 3 of Table 1 (*core* input, *basic* normalization, unigram features, SVM, overall accuracy of 93.65%)

beddings of the metadata content and the definition of the European Data themes. It is an unsupervised approach to predict themes. Table 4 shows the description of the 7 experiments that were performed to automatically assign themes to the datasets in our corpus considering different variants for property selection, term extraction, and feature representation:

- The *Dataset property selection* column indicates the alternatives for the selection of metadata properties describing the datasets as proposed in ẽsection 3.3: the concatenation of three properties (*title+description+keywords*) or just the concatenation of title and description (*title+description*). It must be noted that the property selection for themes is not detailed because it is maintained in all the experiments: the input text is the concatenation of the preferred label and definition of each theme in English.
- The *Term extraction* column indicates the alternatives for term extraction explained in section 3.3: *basic* normalization and the additional process of *stop word removal* in some cases.
- The *Feature representation* column shows the alternatives that have been used for feature representation as proposed in section 3.3: *GloVe sum* indicates the use of GloVe word embeddings and the representation of the full text as the sum of the embeddings of each word in the text; *GloVe average* indicates the use of GloVe word embeddings the representation of the full text as the sum of the embeddings of each word in the text; *BERT* indicates the use of BERT sentence embeddings; and *HuggingFace* indicates the use of HuggingFace transformers for sentence embeddings.

In order to have an orientation about the appropriateness of the predicted themes by the different experiments, we compared the top three themes (ranked by decreasing cosine

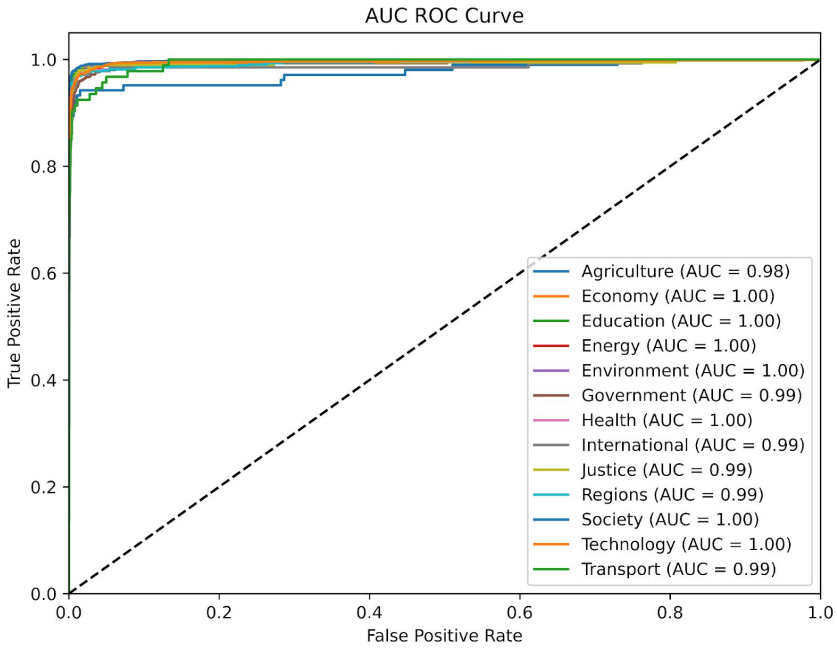


Fig. 12. ROC curve for all the themes in experiment 3 of Table 1 (*core* input, *basic* normalization, unigram features, SVM, overall accuracy of 93.65%)

Table 4. Experiments and results for theme prediction

#	Dataset property selection	Term Extraction	Feature Representation	Agreement score		
				Top 1	Top 2	Top 3
1	title + description + keywords	basic	GloVe sum	0.4127	0.5394	0.6536
2	title + description + keywords	basic	GloVe average	0.4397	0.5770	0.6865
3	title + description + keywords	basic + stop-word removal	GloVe average	0.4680	0.6186	0.7253
4	title + description	basic	BERT	0.2480	0.3360	0.3920
5	title + description + keywords	basic	BERT	0.1908	0.3132	0.4109
6	title + description	basic	HuggingFace	0.3920	0.6320	0.7120
7	title + description + keywords	basic	HuggingFace	0.5023	0.6734	0.7456

similarity distance) with the original dataset themes assigned in the annotated corpus. Table 4 includes an agreement score for the top 1, top 2 and top 3 themes. This agreement score measures the proportion of matches between the top 1/2/3 themes and the assigned themes in the corpus. For instance, if a dataset was originally annotated with the "society" theme, and "society" is the third more relevant theme assigned, this is a match for the top

3 agreement score. Equation 5 shows the formula for computing the agreement score of a $corpus$ and top n predicted themes: $themes(d)$ stands for the function that returns the themes assigned to a dataset d in the $corpus$; $predicted\_themes(d, n)$ stands for the function that returns the top n predicted themes of a dataset d; and $|corpus|$ is the number of datasets in the corpus.

$$Agreement\_score(corpus, n) = \frac{\sum_{d \in corpus} \begin{cases} 1, & \text{if } themes(d) \cap predicted\_themes(d, n) \neq \emptyset \\ 0, & \text{otherwise} \end{cases}}{|corpus|} \tag{5}$$

It can be observed that the best agreement score is obtained with the HuggingFace Transformer for sentence embeddings in experiment 7 and comparing the 3 best ranked themes with the original dataset themes.

5. Discussion

This section discusses the experimental results explicitly in relation to the research questions (RQs) formulated in the Introduction section and highlights how the proposed framework for thematic annotation of OGD addresses each of them. The feasibility of our thematic annotation framework was tested through a series of experiments using a corpus of metadata records which is a representative sample of the current metadata describing OGD in Europe. In the first part of the experiments, we evaluated the correctness of the existing thematic annotations. Although we assessed that the thematic classification correctness of the annotated corpus had an acceptable quality with less than 5% of errors, it must be acknowledged that in many cases the original theme of the inspected datasets for quality control was not the first option in the proposed list of up to three themes assigned by the experts doing the evaluation. This observation is consistent with prior studies highlighting how inconsistencies in metadata structuring and categorization negatively affect the usability and usefulness of open government data [5]. This claim motivates the need for a framework that assists in the thematic annotation relying on the training with a big corpus of annotated datasets, where the biases are minimized. Overall, the results obtained for RQ1 indicate that existing thematic annotations in large OGD portals are generally acceptable but not free from inconsistencies, thereby justifying the need for systematic and scalable support mechanisms such as the proposed framework.

The second research question (RQ2) investigates the extent to which new datasets can be automatically classified when a properly annotated corpus is available. As a second part of the experiments, we analyzed the feasibility of different machine learning techniques to build models for automatic thematic annotation of datasets. In the experiments we cleaned the DCAT-based metadata text by applying several text processing techniques, which also included the translation of false English text and the removal of non-common English terms. With respect to feature representation, we also tested the combined use of unigrams, bigrams and trigrams. Supervised classification techniques such as Logistic Regression and Naive Bayes, as well as Support Vector Machines (SVM), showed effectiveness in classifying datasets with themes using titles, descriptions and keywords, being SVM the technique having the highest accuracy of 93.65%. These results provide a

clear and positive answer to RQ2, demonstrating that supervised machine learning techniques particularly SVM can be effectively integrated into the proposed framework to support large-scale automatic thematic annotation when high-quality annotated metadata are available. Moreover, the results are consistent with similar works in related domains where SVM has also provided a high accuracy for supervised classification [21].

The third research question (RQ3) addresses the issues about the absence of a properly annotated corpus and how can relevant themes be assigned to a dataset based solely on free-text metadata. The final part of our experiments also considered the possibility of not having an annotated corpus or not counting on a perfect annotated corpus with themes. In this case we proposed a representation of texts derived from metadata and theme descriptions in terms of word or sentence embeddings. To predict the themes closer to a dataset, we computed the cosine distance between the embedding representations of the dataset and the candidate themes. After doing experiments with the same sample of datasets extracted from *data.europa.eu* and different techniques for word embeddings (GloVe) and sentence embeddings (BERT and HuggingFace Transformers), we concluded that HuggingFace Transformers were the best approach. The predicted themes have a high agreement score (74.56%) with respect to the original themes assigned in the European data portal. Although the obtained agreement score of 0.7456 is lower than the accuracy obtained with the best experiment for classification models (0.9365), these numbers are not comparable. In the component for thematic prediction of our framework for thematic annotation, we are assuming that the initial thematic annotation is not perfect (or not existent) and we try to identify the closest theme according to the similarities of the language models used to generate the embeddings of dataset metadata and theme descriptions. In some cases, the definition proposed by the European Union [48] for a theme consists of a reduced number of words (sentences) and may not encompass all the possible aspects that the datasets associated with this theme may cover. For instance, the 'Health' theme is defined in just three sentences.[9] Larger texts would generate an embedding vector representation with a better alignment with all the aspects covered by a dataset theme. The experts that evaluated the thematic classification correctness assessed that there were not more than 5% of errors in the classification, but their manual annotation of themes was not constrained by the short definitions of themes. The results obtained for RQ3 show that embedding-based approaches constitute a viable alternative within the framework when annotated corpora are missing or unreliable, thereby increasing the applicability of the proposed thematic annotation framework. The obtained results are coherent with the reported experiments of similar works such as the one proposed by Huseynov et al. [22] for a dataset recommender, which also employed embedding-based representations of metadata and cosine similarity to identify the closer datasets.

6. Conclusions

This paper has presented a framework for the thematic annotation of OGD, which has been tested against a representative sample of 29,793 datasets from *data.europa.eu*, a portal that aggregates datasets (and their associated metadata) harvested from both the

[9] *"This concept identifies datasets covering the domain of health. Health is a state of physical, mental and social well-being in which disease and infirmity are absent. Dataset examples: COVID-19 Coronavirus data; European Cancer Information System."*

member states of the European Union and the European institutions. With respect to the research questions formulated in this study, the results allow us to draw the following conclusions. First, in response to RQ1, we showed that while existing thematic annotations in large OGD portals show an overall acceptable level of correctness, they also present inconsistencies and subjectivity, thus motivating the need for systematic support mechanisms. Second, addressing RQ2, we demonstrated that supervised machine learning techniques, especially SVM, can accurately classify new datasets when a well annotated metadata corpus is available. Finally, in response to RQ3, we confirmed that embedding-based approaches using free-text metadata and theme descriptions provide an applicable solution when annotated corpora are missing. Together, these results validate the design of the proposed framework as a comprehensive and flexible solution for thematic annotation in heterogeneous OGD environments.

6.1. Theoretical Contributions

From a theoretical perspective, this work advances the understanding of thematic annotation in the context of open government data by conceptualizing it as a structured and multi-component process. The proposed framework integrates evaluation of existing annotations, supervised classification, and embedding-based thematic prediction into a unified model, offering a systematic view of how different annotation strategies can be combined depending on data availability and quality. By explicitly addressing multiple annotation scenarios, this study contributes to the literature on metadata quality, semantic enrichment, and data findability in OGD ecosystems.

6.2. Practical Contributions

From a practical standpoint, the proposed framework provides actionable guidance for practitioners and open data portal operators seeking to improve dataset discoverability and consistency of thematic categorization. The framework can be integrated into the dataset ingestion pipelines of widely used open data platforms such as CKAN, DKAN, or Socrata using metadata models based on DCAT where general properties like title, description, keywords, and themes are available. The supervised classification algorithms can be trained to facilitate the automatic annotation of new inserted metadata records. The only requirement for customizing the framework to metadata in other languages is to adjust the term extraction libraries for a satisfactory performance of tokenization, stop removal or other text pre-processing steps in specific languages. In the case of unsupervised classification for theme prediction, we would just need to select pre-trained models for word/sentence embeddings (e.g., Glove, BERT, . . .) in specific languages. By reducing reliance on manual annotation and mitigating subjectivity, the framework has the potential to improve the usability of open data portals and facilitate more efficient dataset discovery for diverse user groups.

6.3. Limitations

This study is also subject to several limitations. Although *data.europa.eu* stands as one of the largest open data government portals and serves as a hub for the national OGD portals

of the European countries, it is crucial to recognize that the categorization of datasets may heavily reflect the biases of the entities responsible for publishing them. The performance of the framework may be influenced by the selection of the vocabulary for data themes because the appropriateness of their titles (preferred labels) and definitions is essential for the assessment of the thematic classification correctness of the annotated corpus (later used in supervised classification techniques) and the approach proposed for theme prediction, which relies on the generation of an embedding-based representation of each theme definition. Exploring the relationship and compatibility of thematic classification schemes employed in OGD portals across other regions [8] could enhance the representativeness and generalization of automated thematic classification algorithms.

In addition, the quality of metadata presents another significant constraint. The prevalence of datasets nominally labelled in English but containing text in other languages exemplifies the noise inherent in the training data. Consequently, sensitivity to such noise emerges as a pertinent consideration in the algorithmic approach to the thematic classification of datasets.

Last, it must be observed that our framework has not been integrated and tested within the scope of an open data portal with end users. Our framework is not aimed at being directly executed by end users interacting with open data portals, but to be integrated during the ingestion process of datasets in a data portal. In order to simulate the thematic annotation during this ingestion process, this work reports experiments whose results have been evaluated in terms of relevance measures, which are employed in the information retrieval discipline to estimate user satisfaction. However, we acknowledge that techniques like A/B testing [50] could be used to verify with end users if an open data portal incorporating this innovation during the ingestion process is better accepted than the portal without the innovation. For instance, we could compare the number of clicks on the first hits returned by both portals with thematic searches.

6.4. Future Research Directions

Building on the findings and limitations of this study, several avenues for future research emerge. First, we would like to explore if the information related to the application schema of the different distributions of datasets can help us to improve the automatic thematic classification of datasets. Available distributions in machine readable formats such as CSV or RDF can provide in some cases meaningful names of thematic attributes of the dataset content. Even in the case of RDF (graph data), these attributes are usually selected from well-known vocabularies, and this may be used to infer links with the themes that can be assigned automatically. Second, we could also explore alternative approaches to unsupervised classification for theme prediction based on the use of keyword extraction techniques [2] and see whether the extracted keywords align with the salient keywords of theme definitions. Last, the impact of data policies on thematic annotation practices and the user experience in accessing and utilizing annotated data could be more deeply investigated to understand how regulations influence the effectiveness of open data ecosystems.

References

1. Ahmed, U.: Reimagining open data ecosystems: a practical approach using AI, CI, and knowledge graphs. In: BIR Workshops. pp. 235–249 (2023)

2. Ahmed, U., Alexopoulos, C., Piangerelli, M., Polini, A.: BRYT: Automated keyword extraction for open datasets. Intelligent Systems with Applications 23, 200421 (2024)
3. Alexopoulos, C., Loukis, E., Charalabidis, Y.: A methodology for determining the value generation mechanism and the improvement priorities of open government data systems. Computer Science and Information Systems 13(1), 237–258 (2016)
4. Alexopoulos, C., Spiliotopoulou, L., Charalabidis, Y.: Open data movement in Greece: a case study on open government data sources. In: Proceedings of the 17th Panhellenic Conference on Informatics. pp. 279–286 (2013)
5. Ansari, B., Barati, M., Martin, E.G.: Enhancing the usability and usefulness of open government data: A comprehensive review of the state of open government data visualization research. Government Information Quarterly 39(1), 101657 (2022)
6. Arlotta, L., Crescenzi, V., Mecca, G., Merialdo, P.: Automatic annotation of data extracted from large web sites. In: International Workshop on the Web and Databases. pp. 7–12 (2003)
7. Attard, J., Orlandi, F., Scerri, S., Auer, S.: A systematic review of open government data initiatives. Government information quarterly 32(4), 399–418 (2015)
8. Bogdanović, M., Gligorijević, M.F., Veljković, N., Puflović, D., Stoimenov, L.: Cross-portal metadata alignment–connecting open data portals through means of formal concept analysis. Information Sciences 637, 118958 (2023)
9. Carducci, G., Leontino, M., Radicioni, D.P., Bonino, G., Pasini, E., Tripodi, P.: Semantically aware text categorisation for metadata annotation. In: Italian Research Conference on Digital Libraries. pp. 315–330. Springer (2019)
10. Davies, T., Walker, S.B., Rubinstein, M., Perini, F.: The state of open data: Histories and horizons. African Minds (2019)
11. Dekkers, M., Kotoglou, S., Nelson, C., Pellegrino, M., Hohn, N., Peristeras, V.: StatDCAT-AP, a common layer for the exchange of statistical metadata in open data portals. In: 6th International Workshop on Semantic Statistics co-located with the 17th International Semantic Web Conference (2016)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of North American Association for Computational Linguistics. vol. 1, p. 2 (2019)
13. Díaz-Corona, D., Lacasta, J., Latre, M.Á., Zarazaga-Soria, F.J., Nogueras-Iso, J.: Profiling of knowledge organisation systems for the annotation of linked data cultural resources. Information Systems 84, 17–28 (2019)
14. Ellen, J.S., Graff, C.A., Ohman, M.D.: Improving plankton image classification using context metadata. Limnology and Oceanography: Methods 17(8), 439–461 (2019)
15. Enríquez-Reyes, R., Cadena-Vela, S., Fuster-Guilló, A., Mazón, J.N., Ibáñez, L.D., Simperl, E.: Systematic mapping of open data studies: Classification and trends from a technological perspective. IEEE Access 9, 12968–12988 (2021)
16. European Commission: DCAT Application profile for data portals in Europe, DCAT-AP Version 2.1.0 (2021), https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/210
17. Evans, A.M., Campos, A.: Open government initiatives: Challenges of citizen participation. Journal of policy analysis and management pp. 172–185 (2013)
18. Freire, J., Fan, G., Feuer, B., Koutras, C., Liu, Y., Peña, E., Santos, A.S., Silva, C.T., Wu, E.: Large language models for data discovery and integration: Challenges and opportunities. IEEE Data Eng. Bull. 49(1), 3–31 (2025)
19. Haunss, S., Kuhn, J., Padó, S., Blessing, A., Blokker, N., Dayanik, E., Lapesa, G.: Integrating manual and automatic annotation for the creation of discourse network data sets. Politics and Governance 8(2), 326–339 (2020)

20. Hudon, A., Beaudoin, M., Phraxayavong, K., Dellazizzo, L., Potvin, S., Dumais, A.: Use of automated thematic annotations for small data sets in a psychotherapeutic context: systematic review of machine learning algorithms. JMIR mental health 8(10), e22651 (2021)
21. Hudon, A., Beaudoin, M., Phraxayavong, K., Dellazizzo, L., Potvin, S., Dumais, A.: Implementation of a machine learning algorithm for automated thematic annotations in avatar: A linear support vector classifier approach. Health Informatics Journal 28(4), 14604582221142442 (2022)
22. Huseynov, R., Nikiforova, A., Symeonidis, D., Duenas-Cid, D.: May the Data Be with You: Towards an AI-Powered Semantic Recommender for Unlocking Dark Data. In: International Conference on Electronic Government. Springer (2025)
23. Ibáñez, L.D., Millard, I., Glaser, H., Simperl, E.: An assessment of adoption and quality of linked data in European open government data. In: International Semantic Web Conference. pp. 436–453. Springer (2019)
24. International Organization for Standardization (ISO): Sampling Procedures for Inspection by Attributes–Part 2: Sampling Plans Indexed by Limiting Quality (LQ) for Isolated Lot Inspection, Standard ISO 2859-2:1985 (1985)
25. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, adoption barriers and myths of open data and open government. Information systems management 29(4), 258–268 (2012)
26. de Juana-Espinosa, S., Luján-Mora, S.: Open government data portals in the European union: A dataset from 2015 to 2017. Data in brief 29, 105156 (2020)
27. Kaldeli, E., Menis-Mastromichalakis, O., Bekiaris, S., Ralli, M., Tzouvaras, V., Stamou, G.: CrowdHeritage: crowdsourcing for improving the quality of cultural heritage metadata. Information 12(02), 64 (2021)
28. Kirstein, F., Dittwald, B., Dutkowski, S., Glikman, Y., Schimmler, S., Hauswirth, M.: Linked data in the european data portal: A comprehensive platform for applying DCAT-AP. In: International Conference on Electronic Government. pp. 192–204. Springer (2019)
29. Kliimask, K., Nikiforova, A.: TAGIFY: LLM-powered tagging interface for improved data findability on OGD portals. In: 2024 Fifth International Conference on Intelligent Data Science Technologies and Applications (IDSTA). pp. 18–27. IEEE (2024)
30. Kubler, S., Robert, J., Neumaier, S., Umbrich, J., Le Traon, Y.: Comparison of metadata quality in open data portals using the analytic hierarchy process. Government Information Quarterly 35(1), 13–29 (2018)
31. Li, B., Zhou, H., He, J., Wang, M., Yang, Y., Li, L.: On the sentence embeddings from pre-trained language models. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 9119–9130. Association for Computational Linguistics, Online (Nov 2020)
32. Lněnička, M., Machova, R., Volejníková, J., Linhartová, V., Knezackova, R., Hub, M.: Enhancing transparency through open government data: The case of data portals and their features and capabilities. Online Information Review 45(6), 1021–1038 (2021)
33. Lnenicka, M., Nikiforova, A., Luterek, M., Milic, P., Rudmark, D., Neumaier, S., Kević, K., Zuiderwijk, A., Bolívar, M.P.R.: Understanding the development of public data ecosystems: From a conceptual model to a six-generation model of the evolution of public data ecosystems. Telematics and informatics p. 102190 (2024)
34. Luna-Reyes, L.F., Bertot, J.C., Mellouli, S.: Open government, open data and digital government. Government Information Quarterly 31(1), 4–5 (2014)
35. Martín-Chozas, P., Montiel-Ponsoda, E., Rodríguez-Doncel, V.: Language resources as linked data for the legal domain. In: Knowledge of the Law in the Big Data Age, pp. 170–180. IOS Press (2019)
36. Miles, A., Brickley, D.: SKOS Core Guide. W3C Working Draft 2 November 2005 (2005), https://www.w3.org/TR/swbp-skos-core-guide

37. Mohamed, M., Pillutla, S., Tomasi, S.: Extraction of knowledge from open government data: The knowledge iterative value network framework. VINE Journal of Information and Knowledge Management Systems 50(3), 495–511 (2020)
38. Morville, P., Rosenfeld, L.: Information Architecture for the World Wide Web. O'Reilly Media, Inc., Canada (2015)
39. Mylonas, P., Voutos, Y., Sofou, A.: A collaborative pilot platform for data annotation and enrichment in viticulture. Information 10(4), 149 (2019)
40. Ni, J., Hernandez Abrego, G., Constant, N., Ma, J., Hall, K., Cer, D., Yang, Y.: Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 1864–1874. Association for Computational Linguistics, Dublin, Ireland (May 2022)
41. Nikiforova, A., McBride, K.: Open government data portal usability: A user-centred usability analysis of 41 open government data portals. Telematics and Informatics 58, 101539 (2021)
42. Nogueras-Iso, J., Lacasta, J., Ureña-Cámara, M.A., Ariza-López, F.J.: Quality of metadata in open data portals. IEEE Access 9, 60364–60382 (2021)
43. Nogueras-Iso, J., Latre, M.Á., Bejar, R., Muro-Medrano, P.R., Zarazaga-Soria, F.J.: A model driven approach for the development of metadata editors, applicability to the annotation of geographic information resources. Data & Knowledge Engineering 81, 118–139 (2012)
44. Paterne Chokki, A., Alexopoulos, C., Matheus, R., Saxena, S., Frénay, B., Vanderose, B.: Do open government data (OGD) portals show signs of knowledge management (KM) practices?: an empirical investigation. Technology Analysis & Strategic Management 36(12), 4829–4844 (2024)
45. Pavia, S., Piraino, N., Islam, K., Pyayt, A., Gubanov, M.N.: Hybrid metadata classification in large-scale structured datasets. Journal of Data Intelligence 3(4), 460–473 (2022)
46. Peng, Y., Wu, Z., Jiang, J.: A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics 43(1), 15–23 (2010)
47. Petrillo, M., Baycroft, J.: Introduction to manual annotation. Fairview research pp. 1–7 (2010)
48. Publications Office of the European Union: Data theme authority table (2022), https://op.europa.eu/s/zBx4
49. Publications Office of the European Union: Metadata quality assessment methodology. Online (2025), https://data.europa.eu/mqa/methodology?locale=en, accessed: 2025-06-06
50. Quin, F., Weyns, D., Galster, M., Silva, C.C.: A/B testing: A systematic literature review. Journal of Systems and Software 211, 112011 (2024)
51. Salih, A.Q.M.: Towards from manual to automatic semantic annotation: based on ontology elements and relationships. International Journal of Web & Semantic Technology 4(2), 21 (2013)
52. Sasse, J., Darms, J., Fluck, J.: Semantic metadata annotation services in the biomedical domain—a literature review. Applied Sciences 12(2), 796 (2022)
53. Shah, S.I.H., Peristeras, V., Magnisalis, I.: A conceptual framework for the government big data ecosystem ('datagov. eco'). Data & Knowledge Engineering p. 102348 (2024)
54. Simonofski, A., Nikiforova, A., Lnenicka, M., Bono Rossello, N.: Artificial intelligence as a catalyzer for open government data ecosystems: A typological theory approach (2025)
55. Tao, W., Yongjia, J., Xiangsheng, R.: A novel two-level one-vs-rest classifier. In: 2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE). pp. 645–648. IEEE (2019)
56. Tuarob, S., Pouchard, L.C., Noy, N.F., Horsburgh, J.S., Palanisamy, G.: ONEMercury: Towards automatic annotation of environmental science metadata. In: Proceedings of the Second International Workshop on Linked Science 2012 - Tackling Big Data, Boston, MA, USA,. CEUR Workshop Proceedings, vol. 951 (2012)

57. Ureña-Cámara, M.A., Nogueras-Iso, J., Lacasta, J., Ariza-López, F.J.: A method for checking the quality of geographic metadata based on ISO 19157. International Journal of Geographical Information Science 33(1), 1–27 (2019)
58. Vandenbussche, P.Y., Vatant, B.: Metadata recommendations for linked open data vocabularies. Version 1, 2011–12 (2011)
59. Verberne, S., D'hondt, E., Van den Bosch, A., Marx, M.: Automatic thematic classification of election manifestos. Information Processing & Management 50(4), 554–567 (2014)
60. W3C: Data Catalog Vocabulary (DCAT). W3C Recommendation 16 January 2014 (2014), https://www.w3.org/TR/vocab-dcat/
61. Wentzel, B., Kirstein, F., Jastrow, T., Sturm, R., Peters, M., Schimmler, S.: An extensive methodology and framework for quality assessment of DCAT-AP datasets. In: International Conference on Electronic Government. pp. 262–278. Springer (2023)
62. Wilkinson, M., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., Bonino da Silva Santos, L.O., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R., Mons, B.: The FAIR guiding principles for scientific data management and stewardship. Scientific data 3(1), 1–9 (2016)
63. Wu, M., Brandhorst, H., Marinescu, M.C., Lopez, J.M., Hlava, M., Busch, J.: Automated metadata annotation: What is and is not possible with machine learning. Data Intelligence 5(1), 122–138 (2023)
64. Yimam, S.M., Biemann, C., de Castilho, R.E., Gurevych, I.: Automatic annotation suggestions and custom annotation layers in webanno. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 91–96 (2014)
65. Zhang, H., Liu, Y., Santos, A., Freire, J., et al.: Autoddg: Automated dataset description generation using large language models. arXiv preprint arXiv:2502.01050 (2025)

Abdul Aziz received the bachelor's degree in computer science from the COMSATS Institute of Information Technology, Lahore, Pakistan, in 2013, and the master's degree in computer science from the National University of Computer and Emerging Sciences, Karachi, Pakistan, in 2018. In 2025 he defended his Ph.D. degree in Computer Science at the University of Zaragoza (Advanced Information Systems Laboratory of the Aragon Institute of Engineering Research), Spain, about the use of feedback mechanisms to promote the Inclusiveness of open government data portals. During his Ph.D. he was an Early Stage Researcher in the ODECO project, a Horizon 2020 Marie Skłodowska-Curie Innovative Training Network (H2020-MSCA-ITN-2020), where he contributed to advancing research on open data ecosystems, user engagement, and data-driven innovation. He is currently working as an AI & Data Consultant at PQNO?, where he applies advanced artificial intelligence and data-driven methodologies to support innovation, strategic decision-making, and digital transformation across diverse domains.

Mohsan Ali is a researcher at the University of the Aegean. He was awarded a Marie Skłodowska-Curie Innovative Training Network (H2020-MSCA-ITN-2020) scholarship in 2021 to pursue his PhD in Greece, focusing on open data ecosystems. His current research centers on the technical interoperability of open data within the Information Systems Laboratory, as part of the ODECO-funded project. His expertise includes open data, data interoperability, data science, natural language processing, and artificial intelligence. In addition, he has specialized in deep learning, a skill developed through his academic and professional training. He holds a Master's degree in Computer Science (MSCS) with

distinction and was awarded a Gold Medal from Air University, Islamabad, Pakistan. He also earned his Bachelor's degree from Pir Mehr Ali Shah Arid Agriculture University, Rawalpindi, Pakistan.

Dagoberto Jose Herrera-Murillo received the bachelor's degree in Business Informatics from Tecnológico de Monterrey and the joint master's degree in Big Data Management from Université Libre de Bruxelles, Universitat Politècnica de Catalunya (BarcelonaTech), and Eindhoven University of Technology. In 2025 he defended his Ph.D. degree in Computer Science at the University of Zaragoza (Advanced Information Systems Laboratory of the Aragon Institute of Engineering Research), Spain, about the evaluation of user interfaces of open data portals . During his PhD he was an Early Stage Researcher for the ODECO project, a Horizon 2020 Marie Sklodowska-Curie Innovative Training Network (H2020-MSCA-ITN-2020).

Maria Ioanna Maratsi is a researcher at the University of the Aegean (Department of Information and Communication Systems Engineering) and PhD candidate in the area of semantic interoperability and open data. She is a graduate (BSc) of Computer Science and Telecommunications from University of Piraeus, Greece, and alumna of the MSc in Information Security of Stockholm University (Department of Computer and Systems Sciences - DSV), Sweden. Ioanna was also a member of the Systems Analysis and Security Unit of Stockholm University. During her PhD she was an Early Stage Researcher for the ODECO project, a Horizon 2020 Marie Sklodowska-Curie Innovative Training Network (H2020-MSCA-ITN-2020). Her academic interests include open and linked data, interoperability, knowledge graph engineering, data privacy, digital forensics, AI ethics, and multidisciplinary approaches in information technology.

Francisco Javier Lopez-Pellicer received the M.S. and Ph.D. degrees in computer engineering from the University of Zaragoza. In 2004, he started his research with the Advanced Information Systems Laboratory, University of Zaragoza (Spain). Currently, he is an Associate Professor of computer science at the University of Zaragoza. Over the past ten years, his professional career has been linked to open data initiatives and spatial data infrastructures. Within this context, he has coauthored numerous publications in books, journals or conference proceedings; and has collaborated in several R+D projects. His research interests include open data infrastructures, service-based geographic information systems, and various information systems.

Javier Nogueras-Iso received the M.S. and Ph.D. degrees in computer science from the University of Zaragoza, Spain. In 1998, he started his research with the Advanced Information Systems Laboratory, University of Zaragoza (Spain), where he is currently a Full Professor of computer science. From 2011 to 2017, he was the Director of the Catedra Logisman on Technological Document Management. From 2015 to 2019, he was the Associate Director of the Aragon Institute of Engineering Research (I3A). His research interests include information retrieval and semantic web technologies applied to different domains, although with a special emphasis on geographic information infrastructures.

Received: October 29, 2025; Accepted: April 6, 2026.