Computer Science and Information Systems 22(1):1–32 https://doi.org/10.2298/CSIS240229066P

Segmentation of COVID-19 Lesions in CT Scans through Transfer Learning

Symeon Psaraftis-Souranis[1], Christos Troussas[1], Athanasios Voulodimos[2], and Cleo Sgouropoulou[1]

[1] Department of Informatics and Computer Engineering,
University of West Attica, Egaleo 12243, Greece
{cs141049, ctrouss, csgouro}@uniwa.gr
[2] School of Electrical and Computer Engineering,
National Technical University of Athens, Athens 15780, Greece
thanosv@mail.ntua.gr

Abstract. Since its emergence at the end of 2019, SARS-CoV-2 has infected millions worldwide, challenging healthcare systems globally. This has prompted many researchers to explore how machine learning can assist clinicians in diagnosing infections caused by SARS-CoV-2. Building on previous studies, we propose a novel deep learning framework designed for segmenting lesions evident in Computed Tomography (CT) scans. For this work, we utilized a dataset consisting of 20 CT scans annotated by experts and performed training, validation, and external evaluation of the deep learning models we implemented, using a 5-fold cross-validation scheme. When splitting data by slice, our optimal model achieved noteworthy performance, attaining a Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) score of 0.8644 and 0.7612 respectively, during the validation phase. In the external evaluation phase, the model maintained strong performance with a DSC and an IoU score of 0.7211 and 0.5641, respectively. When splitting data by patient, our optimal model achieved a DSC score of 0.7989 and an IoU score of 0.6686 during the validation phase. During the external evaluation phase, the model maintained strong performance with a DSC and IoU score of 0.7369 and 0.5837, respectively. The results of this research suggest that incorporating transfer learning along with appropriate preprocessing techniques, can contribute to achieving state-of-the-art performance in the segmentation of lesions associated with SARS-CoV-2 infections.

Keywords: computer vision, machine learning, deep learning, transfer learning, convolutional neural networks, COVID-19, semantic segmentation, medical imaging, computed tomography.

1. Introduction

According to the World Health Organization, since the first detection of SARS-CoV-2, there have been more than 771 million confirmed cases of COVID-19 infection, of which nearly 7 million cases have resulted in death [1]. Early detection and diagnosis of COVID-19 are key factors in limiting the spread of the virus [2]. Diagnostic tests remain the most common method to detect SARS-CoV-2, with Reverse Transcription

Polymerase Chain Reaction (RT-PCR) tests being the most reliable. However, the fact that they are time-consuming is a significant disadvantage, especially regarding this particular virus, where early diagnosis plays a vital role in successful treatment [2].

The study of medical images, such as CT scans, constitutes a valuable approach for identifying COVID-19 by detecting pathological findings associated with lower respiratory tract infections (pneumonia) [3]. Medical imaging is widely used by specialists as a diagnostic tool for SARS-CoV-2 associated pneumonia, with CT scans providing much higher diagnostic accuracy, as they can identify incipient lesions in the lung parenchyma that cannot be discerned on plain chest X-rays [4].

Through the review of medical images, clinicians can assess the health status of each patient and, if concerning findings are discovered, make the correct diagnosis and follow the appropriate treatment. However, while the process of reviewing medical images by healthcare professionals often leads to correct and timely detection, diagnosis, and treatment, 40-54% of malpractice cases attributable to medical radiologists are associated with errors in interpreting medical images [5].

Given these challenges, leveraging machine learning techniques for the analysis of medical images can be a valuable tool for the early and accurate detection and diagnosis of COVID-19 infections.

The COVID-19 outbreak has resulted in increased interest in how machine learning techniques can contribute to the process of analyzing medical images to detect COVID-19 infections. Consequently, since the start of the COVID-19 pandemic, a multitude of scientific papers featuring noteworthy results have been consistently published by experts in the field of machine learning. The large volume of relevant papers indicates that this research area remains active, which is understandable given the ongoing presence of COVID-19. The opportunities for research in this area, along with the continued interest of the scientific community, motivated our engagement with this specific research topic.

In this work, we investigate the applicability of deep learning techniques for detecting and isolating lesions associated with COVID-19 pneumonia. Our contributions can be enumerated as follows:

- We utilized a segmentation architecture called U-Net [6], which we trained and evaluated using a publicly available dataset. Subsequently, we applied transfer learning principles to train and evaluate variations of the U-Net architecture, replacing the encoder with a pre-trained CNN model. Finally, we compared their performance with the basic U-Net architecture.
- We evaluated the impact of data preprocessing on the performance of the models used in this work. Specifically, we trained the selected architectures on CT slices that were normalized in terms of contrast and brightness, and then segmented to retain only the information within the lung parenchyma for each slice. While contrast and brightness normalization are standard practices, our contribution lies in the systematic integration of these techniques within a deep learning framework tailored for detecting lesions associated with COVID-19 pneumonia. This preprocessing approach proved crucial in achieving state-of-the-art results, demonstrating its effectiveness in enhancing model accuracy and reliability.

The rest of the article is organized as follows: Section 2 provides an overview of related work on COVID-19 lesion segmentation. Section 3 explains the methodology used in this research. In Section 4, we present and discuss the results of the conducted experiments and compare them with state-of-the-art approaches. Finally, in Section 5, we draw conclusions from this research.

2. Related Work

This section provides an overview of the related literature on COVID-19 lesion segmentation in CT scans. To better present the findings, Table 1 summarizes the work conducted by other researchers that is relevant to this research paper. It can be seen that the U-Net architecture is the most popular approach used to address this problem. Furthermore, it is noteworthy that among the approaches detailed in Table 1, only three make use of the transfer learning technique. This observation underscores that the utilization of transfer learning in the context of COVID-19 lesion segmentation on CT scans has not been extensively explored within the existing literature.

Table 1. Summary of the existing literature related to our proposed work

Model Architecture	Training and Validation Dataset	External Evaluation Dataset	Preprocessing & Transfer Learning	Validation Results	External Evaluation Results
Ma et al. [7] 2D U-Net	COVID-19-CT-Seg [8] (5-fold cross-validation Train: 20% Val: 80%)	-	HU Clipping [-1250, 250], No Transfer Learning	DSC: 60.80%	-
Ma et al. [7] nnU-Net [9]	COVID-19-CT-Seg (5-fold cross-validation Train: 20% Val: 80%)	MosMed [10]	HU Clipping [-1250, 250], No Transfer Learning	DSC: 67.30%	DSC: 58.80%
Müller et al. [11] 3D U-Net [12]	COVID-19-CT-Seg (5-fold cross-validation Train: 80% Val: 20%)	-	HU Clipping [-1250, 250], No Transfer Learning	DSC: 76.10%	-
Müller et al. [13] 3D U-Net	COVID-19-CT-Seg (5-fold cross-validation Train: 80% Val: 20%)	An et al. [14]	HU Clipping [-1250, 250], Data Augmentation, No Transfer Learning	DSC: 80.40%	DSC: 66.10%
Owais et al. [15] DAL-Net	Experiment 1: COVID-19-CT-Seg (5-fold cross-validation Train: 80% Val: 20%) Experiment 2: MosMed (5-fold cross-validation Train: 80% Val: 20%) Experiment 3: COVID-19-CT-Seg	Experiment 3: MosMed	Experiment 3: Reinhard Transformation [16], No Transfer Learning	Experiment 1: DSC: 83.23% IoU: 74.86% Experiment 2: DSC: 68.63% IoU: 61.35%	Experiment 3: DSC: 74.93% IoU: 66.50%

Zheng et al. [17] 3D CU-Net	COVID-19-CT-Seg (5-fold cross-validation Train: 80% Val: 20%)	MosMed	HU Clipping [-1250, 250], Data Augmentation, No Transfer Learning	DSC: 77.80%	DSC: 66.80%
Yixin Wang et al. [18] 3D U-Net	COVID-19-CT-Seg (5-fold cross-validation Train: 20% Val: 80%)	-	No Transfer Learning	DSC: 70.04%	-
Singh et al. [19] LungINFSeg	COVID-19-CT-Seg (Train: 70% Val: 10% Test: 20%)	-	Data Augmentation, No Transfer Learning	DSC: 80.34% IoU: 68.77%	-
Amara et al. [20] O-Net	COVID-19-CT-Seg (Only 10 CT) (Train: 70% Val: 30%)	MosMed	Image Cropping, Data Augmentation, No Transfer Learning	DSC: 86.60% IoU: 76.40%	DSC: 58.40% IoU: 42.80%
Aswathy et al. [21] 3D U-Net	COVID-19-CT-Seg (Train: 60% Val: 20% Test: 20%)	-	Lung Parenchyma Segmentation, Patchwise Data Augmentation, No Transfer Learning	DSC: 82.00%	
Xiaoyan Wang et al. [22] SSA-Net	Experiment 1: COVID-19-CT-Seg (5-fold cross validation) Experiment 2: MedSeg Dataset [23] (Only 98 CT Slices) (5-fold cross validation)	-	HU Clipping [-1250, 250], No Transfer Learning	Experiment 1: DSC: 65.22% Experiment 2: DSC: 75.40%	-
Krinski et al. [24] Various CNN Models	COVID-19-CT-Seg (5-fold cross validation Train: 80% Val: 20%)	-	Transfer Learning (ImageNet [25])	Best Model DSC: 73.67% Best Model IoU: 70.91%	-
Mahmoudi et al. [26] 2D U-Net	COVID-19-CT-Seg (4-fold cross validation Train: 70% Val: 30%)	-	CLAHE, Image Cropping, Data Augmentation, No Transfer Learning	DSC: 91% IoU: 85%	-
Qiblawey et al. [27] Various CNN Models	COVID-19-CT-Seg (10-fold cross validation Train: 60% Val: 20% Test: 20%)	-	HU Normalization, Lung Parenchyma Segmentation, Data Augmentation, Transfer Learning (ImageNet)	Best Model DSC: 94.13% Best Model IoU: 91.85%	-
Enshaei et al. [28] COVID-Rate	Private Dataset + COVID-19-CT-Seg (Only 10 CT) (10-fold cross-validation Train: 60% Val: 10% Test: 30%)	MedSeg Dataset (Only 9 CT) + COVID-CT-MD [29]	Lung Parenchyma Segmentation, Data Augmentation, No Transfer Learning	DSC: 80.69%	DSC: 79.98%
Uçar et al. [30] U-Net + Various CNN Models as Encoders	MedSeg Dataset (Train: 80% Val: 20% Test: 10% of Training Data)	-	Transfer Learning (ImageNet)	Best Model DSC: 84.04% Majority Voting DSC: 85.03%	-

Building upon the research summarized in Table 1, it becomes apparent that deep learning methods, particularly those employing the U-Net architecture, are widely recognized for their effectiveness in segmenting lesions within CT scans. During our experimental phase, we followed this approach and utilized a 2D U-Net architecture, along with variations where the encoder of the U-Net model was substituted with pre-trained convolutional neural networks.

Each of the previously mentioned research papers employs various methods and pre-processing techniques to provide a more reliable assessment and enhance the performance of the models on the given task. Notably, in [7], [11], [13], [17] and [22], the authors normalize the CT slices by clipping pixel intensities. In [13], [17], [19], [20], [21], [26], [27] and [28], data augmentation is applied. Moreover, in [21], [27] and [28], segmentation of the lung parenchyma is performed and in [24] and [30], the authors leverage transfer learning.

In this work, we combine methods and pre-processing techniques from the aforementioned research papers, to achieve state-of-the-art performance. Specifically, we adopt a technique similar to the one used in [7], [11], [13], [17] and [22] to normalize CT slices in terms of contrast and brightness. Moreover, we apply data augmentation strategies analogous to those performed in [13], [17], [19], [20], [21], [26], [27] and [28]. Additionally, following the approaches in [21], [27] and [28], we perform lung parenchyma segmentation on CT slices. Furthermore, inspired by [24] and [30], we utilize a transfer learning method to train variations of the U-Net model, where the encoder of the network is replaced with a pre-trained convolutional neural network.

3. Materials and Methods

Analyzing medical images to aid in the diagnosis of COVID-19 pneumonia can be framed as a semantic image segmentation problem. In this type of problem, deep learning methods, such as convolutional neural networks, are provided with CT slices alongside corresponding “masks”, which are images where the regions of the CT slices containing lung lesions have been annotated by experts. The networks are then trained using the input data to map the pixels of each CT slice into distinct categories based on the presence or absence of lesions associated with COVID-19 pneumonia.

The framework proposed in this study is illustrated in Figure 1. In the first step, the selected dataset is preprocessed before being fed into each deep learning model. Subsequently, the deep learning models to be used in this work are selected. For our experiment, we chose to employ pre-trained neural networks as the backbone of a classic U-Net model. Following that, the implemented models undergo training and validation. Finally, after the training process is completed, the models’ ability to perform semantic image segmentation is evaluated. Appropriate performance evaluation metrics are used to assess their performance.

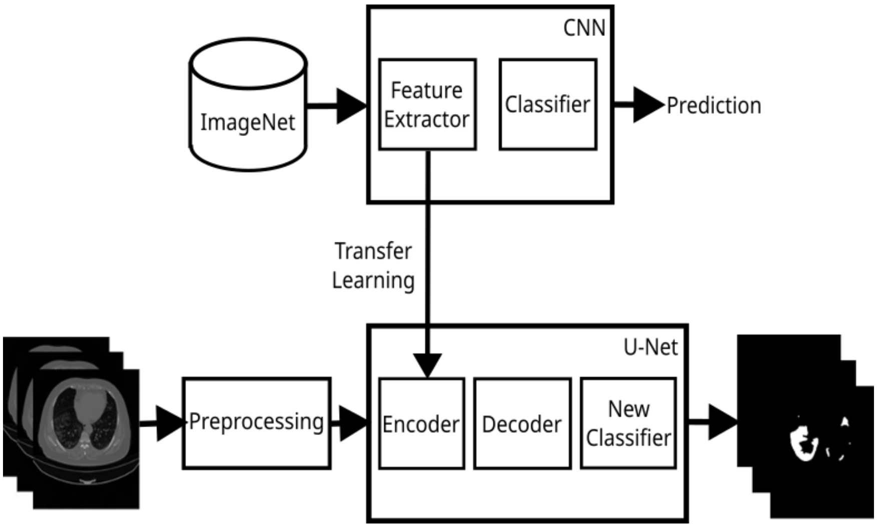

Fig. 1. Block diagram of the proposed method

3.1. Dataset Selection

For this work, we opted to use a publicly available dataset called COVID-19-CT-Seg [7], [8]. This dataset consists of 3,520 CT slices collected from the Coronacases.org [31] and Radiopaedia.org [32] repositories, comprising data from 20 distinct patients [7], [8]. Specifically, 2,581 CT slices originate from 10 patients within the Coronacases.org repository, with the remaining 939 CT slices attributed to the other 10 patients and taken from the Radiopaedia.org repository. Figure 2 illustrates the distribution of CT slices in the two subsets of the COVID-19-CT-Seg dataset.

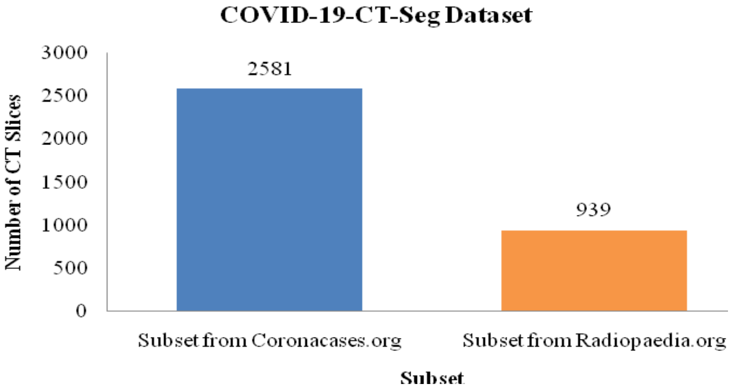

Fig. 2. Bar graph depicting the number of CT slices in the COVID-19-CT-Seg dataset

The creators of this dataset have provided a set of 3,520 masks, in which the lung parenchyma is outlined bilaterally. In addition, they have included another set of 3,520 masks outlining lesions attributed to SARS-CoV-2.

The initial outlining procedure was carried out by two radiologists with 1-5 years of experience [7], [8]. It was then optimized by radiologists with 5-10 years of experience and finally validated and further optimized by a radiologist with over 10 years of experience in respiratory radiology [7], [8].

Each patient's slices are stored in the Neuroimaging Informatics Technology Initiative (NifTI) format. A visual representation of a subset of the COVID-19-CT-Seg dataset, alongside corresponding masks outlining lesions attributed to SARS-CoV-2, is depicted in Figure 3.

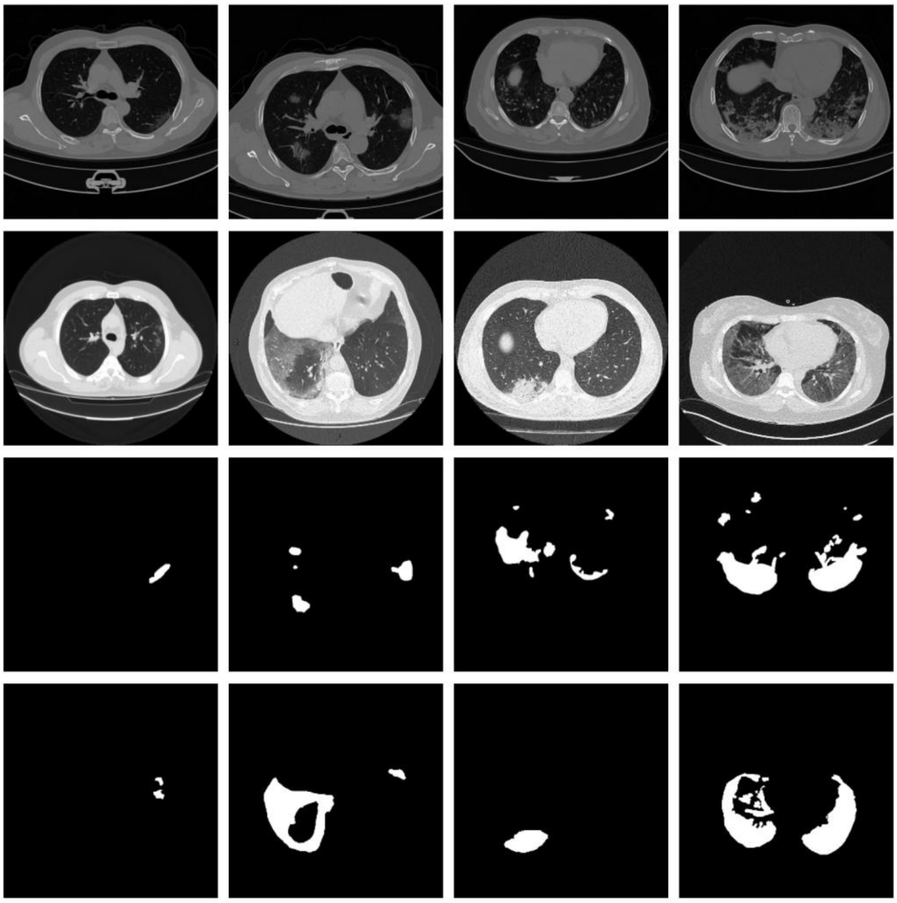

Fig. 3. A subset of CT slices from the COVID-19-CT-Seg dataset, along with their corresponding masks in which areas where lesions attributed to SARS-CoV-2 have been outlined

3.2. Preprocessing

In this section, we list the preprocessing techniques applied to the data before feeding it as input to each of the deep learning models for training and validation. The primary purpose of pre-processing the data is to improve data quality, reduce training time, and produce better results.

In the CT slices, the pixels containing lesions (positive samples) are far fewer than those that do not contain lesions (negative samples). This, combined with the relatively small size of the dataset used in this study, results in class imbalance. To alleviate this problem, we removed CT slices that do not contain lesions.

Therefore, 1,844 slices are retained in the dataset, with 1,351 slices from the Coronacases.org repository and the remaining 492 from the Radiopaedia.org repository. The distribution of CT slices in the two subsets of the COVID-19-CT-Seg dataset is shown in Figure 4. Additionally, since not all CT slices are of the same dimensions, we resize them to 256x256.

The CT slices in the dataset used in this work present significant difference in terms of contrast and brightness. This difference makes the detection of important features for accurate semantic image segmentation difficult. To address this problem, we normalize the CT slices in terms of contrast and brightness. This is achieved by applying a method called "windowing," where we adjust the parameters Window Width and Window Level of the CT slices to 1400 HU and -500 HU respectively. The choice of these values is not random. According to [33], adjusting Window Width and Window Level to these specific values enhances the visibility of features inside the lung parenchyma, which in turn improves lesion detectability. Moreover, by standardizing the appearance of CT images through the method of "windowing," variability can be reduced, which improves the generalization and performance of deep learning models. Figure 5 shows a CT slice before and after applying contrast and brightness normalization.

In addition, we normalize the pixel values of the CT slices to the range [0, 1], which is a common practice in deep learning applications. This preprocessing step facilitates convergence during training by ensuring that all pixel values are within a standardized range. This is achieved by dividing the pixel values of the CT slices by 255, which is the maximum pixel value.

The region of interest in a CT slice, examined for lesions attributed to the SARS-CoV-2 virus, is the lung parenchyma. A beneficial practice that enhances the efficiency of the deep learning models used in this work involves the segmentation of CT to preserve only the lung parenchyma in each slice. This can be achieved by utilizing the masks in which the lung parenchyma has been outlined bilaterally. Figure 6 shows a CT slice before and after the segmentation of the lung parenchyma.

To mitigate the risk of overfitting, data augmentation is applied to the data before it is fed into the neural network. Data augmentation is achieved by making slight modifications to the existing data. These modifications consist of combinations of rotation within the range of 0 to 15 degrees, horizontal flipping, and horizontal and vertical translation within the range of 0 to 15%.

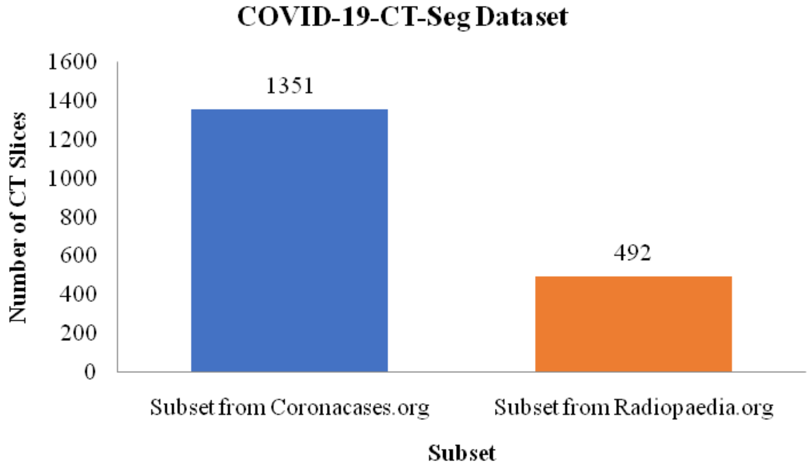

Fig. 4. Number of CT slices in the COVID-19-CT-Seg dataset after removing the slices without lesions

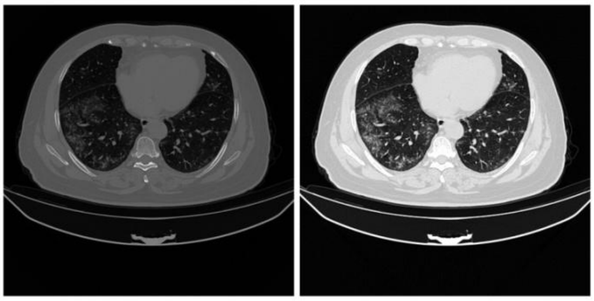

Fig. 5. A CT slice before and after applying contrast and brightness normalization

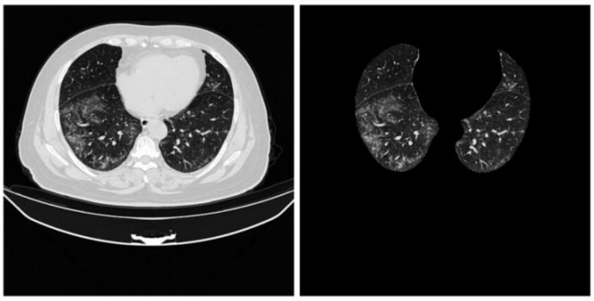

Fig. 6. A CT slice before and after segmenting the lung parenchyma

3.3. Data Splitting

For the purposes of this research paper, we leveraged the 20 CT scans provided in the COVID-19-CT-Seg dataset. Initially, our intention was to train our models using two separate datasets, as exposure to diverse datasets during training can improve a model's robustness. However, due to limited access to public datasets, we chose an alternative approach and opted to use a single dataset sourced from two distinct repositories. To explore this, we employed two different scenarios:

- In the first scenario, a set of 10 CT scans from the Coronacases.org repository, comprising of 2,581 slices, was utilized to construct the training and validation sets for training and evaluation of the deep learning models, respectively. Concurrently, another set of 10 CT scans from the Radiopaedia.org repository encompassing 939 slices, was used as an external evaluation set.
- In the second scenario, the allocation of CT scans was reversed. Here, 10 CT scans from the Radiopaedia.org repository were used to form the training and validation sets, while 10 CT scans from the Coronacases.org repository were designated for the external evaluation set.

The partitioning of data into training and validation subsets was accomplished through a 5-fold cross validation scheme. It is worth mentioning that the splitting into training and validation sets was performed using two distinct approaches:

- At the slice level, where CT slices were randomly split into training and validation sets without any predefined criteria.
- At the patient level, where CT slices were allocated into training and validation sets based on the patient they belonged to.

3.4. Model Selection

This section introduces the deep learning models employed in this research. The models used include the classic U-Net model and variations of it with different "backbones".

Specifically, these U-Net architectures replace the contractive path, or encoder, with convolutional neural networks pre-trained on the ImageNet [25] dataset, thereby leveraging transfer learning. The pre-trained convolutional neural networks selected as the encoder for the U-Net model are the VGG16 [34] and DenseNet121 [35] architectures.

U-Net Architecture. The U-Net architecture, introduced in 2015 by Ronneberger et al., is a convolutional neural network designed for biomedical image segmentation tasks [6]. This network consists of three main components: the contracting path, the bottleneck level, and the expansive path [6]. The contracting path, which acts as the encoder of the network, follows the typical structure of a convolutional neural network [6]. In each block of this path, the spatial dimensions of the images are halved, while the number of the feature channels is doubled [6]. The bottleneck level connects the contracting path to the expansive path [6]. The expansive path, or decoder, consists of blocks where, in each, the spatial dimensions of the images are doubled, and the number of feature channels is halved [6]. The output of each block in the contracting path is concatenated with the input of the corresponding block in the expansive path [6]. This enables the network to preserve high-resolution features from the contracting path [6]. The implementation of the U-Net model allows it to leverage both high-level and low-level features, contributing to improved segmentation accuracy [6]. Another advantage of the U-Net architecture is that it has a relatively small number of parameters, which results in reduced execution time compared to alternative segmentation methods [6].

U-Net Architecture with the VGG-16 Model as Encoder. In this architectural variation, the traditional encoder is replaced with a pre-trained VGG16 model. The VGG16 model consists of 21 layers organized into five blocks [34]. The bottleneck level of the network acts as the intermediary link connecting the VGG16 model with the decoder. The decoder itself consists of five blocks of layers. Notably, the outputs from the last four blocks of the encoder are concatenated with the inputs of the corresponding first four blocks of the decoder.

U-Net Architecture with the DenseNet121 Model as Encoder. In this architectural variation, the traditional encoder is replaced with a pre-trained DenseNet121 model. The DenseNet121 model begins with a convolutional layer followed by a max pooling layer [35]. It then includes four dense blocks, separated by transition blocks [35]. The decoder consists of five blocks of layers. Notably, the outputs from the convolutional layer and the first three dense blocks of the encoder are concatenated with the inputs of the corresponding first four blocks of the decoder.

3.5. Model Training

The training of the implemented architectures was conducted using the Kaggle [36] platform, leveraging the computational capabilities of an Nvidia Tesla P100 GPU with 16 GB of memory. The models were trained on grayscale images of size 256x256 for 200 epochs, with a learning rate of 0.001. Due to memory constraints, the data were divided into batches and incrementally loaded into memory during each epoch. A batch size of 32 was chosen for this purpose. Additionally, Adam [37] was selected as the optimizer. Regarding fine-tuning, we chose not to freeze any layers of the encoder part of the models during training. This decision was motivated by two reasons: First, the problem addressed in this research significantly differs from the original task for which the CNN models were pre-trained. Training all layers from scratch allows the models to better adapt to the specific characteristics of the new problem. Second, since the dataset used is relatively small, freezing layers might limit the models' ability to learn important features specific to the dataset.

3.6. Model Evaluation

In this section, we detail the methodologies used to evaluate the performance of the models described in the previous section. To assess the models' ability to segment lesions in CT slices and detect overfitting, we implemented a 5-fold cross-validation scheme. Additionally, we used graphical representations to monitor the training and validation processes, aiming to identify the presence of underfitting or overfitting and investigate the models' generalization capabilities. The evaluation also includes key metrics, namely Precision, Recall, Dice Similarity Coefficient, and Intersection over Union.

Precision, in the context of semantic image segmentation, is defined as the ratio of true positive pixels to the total number of pixels included in the segmentation by the model, as shown in Equation (**1**).

$$\mathrm{Pr}\mathit{ecision} = \frac{TP}{TP + FP} \tag{1}$$

Recall, in the context of semantic image segmentation, is defined as the ratio of true positive pixels to the total of pixels that should have been included in the segmentation by the model, as shown in Equation (**2**).

$$\mathrm{Re}\mathit{call} = \frac{TP}{TP + FN} \tag{2}$$

The Dice Similarity Coefficient, a fundamental metric for evaluating semantic segmentation tasks, is calculated as the harmonic mean of Precision and Recall. This metric measures the spatial overlap between two segmentation regions, A and B, as shown in Equation (**3**).

$$DSC(A,B) = \frac{2|A \cap B|}{|A| + |B|} = \frac{2TP}{2TP + FP + FN} \tag{3}$$

Intersection over Union, another fundamental metric for evaluating semantic segmentation tasks, measures the spatial overlap between two segmentation regions, A and B, as shown in Equation (**4**).

$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN} \tag{4}$$

4. Results and Discussion

To assess the impact of transfer learning on the performance of the U-Net model, we conducted an ablation experiment. This involved training three variants of the U-Net architecture: the standard U-Net model and two modified versions, where the original encoder was replaced with pre-trained models, specifically VGG16 and DenseNet121.

4.1. Results when Splitting Data by Slice

For our initial experiment, we chose to use 10 CT scans obtained from Coronacases.org for both training and validation, while reserving the remaining CT scans from Radiopaedia.org for external evaluation. This split into training and validation sets was performed at the slice level. Table 2 presents the metrics used to assess the performance of the trained models in semantically segmenting lesions associated with COVID-19 pneumonia.

Among the models evaluated, the U-Net + DenseNet121 model demonstrated superior performance on both the validation and external validation datasets, as indicated by higher mean DSC and IoU values. Despite some decline in performance on the external validation set, all models maintained competitive results.

When comparing the models' performance on the validation data with their performance on the external evaluation data, it is evident that the evaluation metric values are higher in the former than in the latter. This outcome is expected, considering that the external validation dataset contains previously unseen data. Therefore, it is reasonable that the models do not perform as well on this "unknown" dataset. Moreover, the validation set is used during the training process to tune each network's parameters, which may lead the models to learn features specific to that set and, thus perform better on it, despite not being explicitly trained on that specific dataset.

Notably, all models achieved high mean Precision and Recall on the validation set. Higher Precision indicates a lower false positive rate, meaning the model is less likely to incorrectly identify non-lesion regions as lesion-containing areas. Higher Recall implies that the model can effectively identify actual lesions.

In the case of external evaluation, a notable difference between Precision and Recall values is apparent. Although Precision slightly decreased compared to the validation set, Recall exhibited a more significant reduction. A lower Recall suggests that the model is more likely to miss identifying actual lesion regions.

Table 2. Ablation study results when evaluating variations of the U-Net architecture on the validation and external validation sets, with data split at the slice level. Training and validation utilized 10 CT scans sourced from the Coronacases.org repository, while the external validation set comprised the remaining 10 CT scans from the Radiopaedia.org repository

Model	Validation Mean DSC	Validation Mean IoU	Validation Mean Precision	Validation Mean Recall	External Mean DSC	External Mean IoU	Externa Mean Precision	External Mean Recall
U-Net	0.8423 ± 0.0131	0.7277 ± 0.0194	0.8312 ± 0.0267	0.8545 ± 0.0195	0.6587 ± 0.0396	0.4921 ± 0.0435	0.8321 ± 0.0269	0.5487 ± 0.0631
U-Net + VGG16	0.8573 ± 0.0072	0.7503 ± 0.0111	0.8489 ± 0.0120	0.8661 ± 0.0140	0.7118 ± 0.0165	0.5528 ± 0.0200	0.8190 ± 0.0171	0.6305 ± 0.0325
U-Net + DenseNet121	0.8644 ± 0.0091	0.7612 ± 0.0143	0.8597 ± 0.0112	0.8691 ± 0.0094	0.7211 ± 0.0197	0.5641 ± 0.0238	0.8071 ± 0.0450	0.6566 ± 0.0584

Table 3 presents the training and inference times for the models discussed in Table 2. The results show that the U-Net model exhibits the shortest runtime, while the U-Net + DenseNet121 model required the most time. This observation aligns with expectations, as the U-Net + DenseNet121 model includes DenseNet121 as its encoder, a deep and complex model that requires more time to run compared to the standard U-Net model.

Figure 7 illustrates the training and validation curves for the models, the results of which are shown in Table 2. Examination of these curves reveals that all models converge, showing no signs of overfitting.

Table 3. Running times of different variations of the U-Net architecture evaluated with the data split at the slice level. Training and validation used 10 CT scans sourced from the Coronacases.org repository, while the external validation set comprised the remaining 10 CT scans from the Radiopaedia.org repository

Model	Training Time	Validation Time	External Validation Time
U-Net	4577.40 ± 51.57 seconds	1.47 ± 0.08 seconds	2.62 ± 0.15 seconds
U-Net + VGG16	4633.48 ± 17.83 seconds	1.63 ± 0.02 seconds	2.90 ± 0.01 seconds
U-Net + DenseNet121	7807.70 ± 218.71 seconds	2.58 ± 0.05 seconds	4.50 ± 0.098 seconds

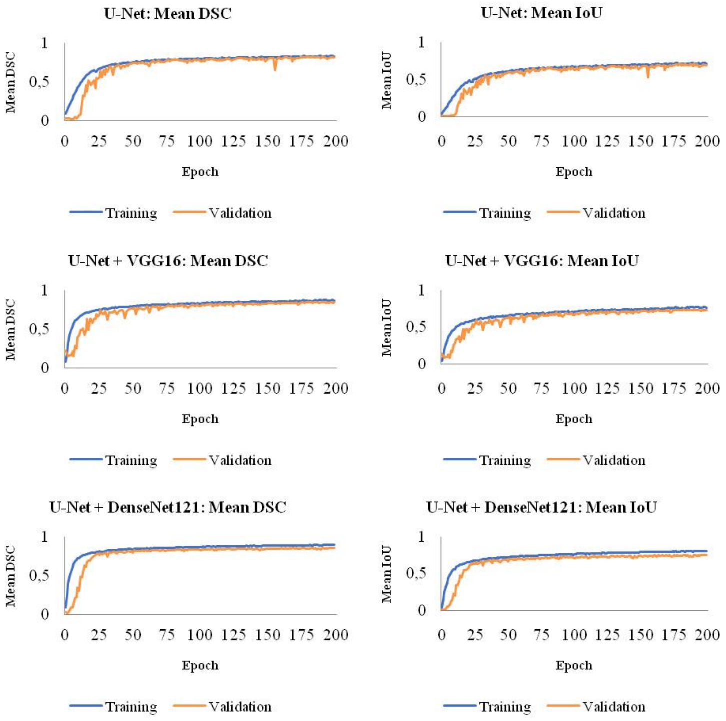

Fig. 7. Training and validation curves for the models trained and evaluated on data split at the slice level. Training and validation used 10 CT scans sourced from the Coronacases.org repository, while the external validation set comprised the remaining 10 CT scans from the Radiopaedia.org repository

In our second experiment, we maintained the data splitting by slice approach, but this time utilized the 10 CT scans taken from Radiopaedia.org for training and validation, and the 10 CT scans from Coronacases.org for external validation.

Upon comparing the model performances, it becomes apparent that the U-Net + DenseNet121 model once again exhibits the highest performance across all metrics in both the validation and external evaluation cases. Notably, a distinct observation arises: the difference in performance between U-Net + DenseNet121 and the other variants is more apparent compared to the corresponding models in Table 2. Given that the dataset sourced from Radiopaedia.org is significantly smaller than the one sourced from Coronacases.org, it seems that U-Net + DenseNet121 performs better than the other two models, particularly in scenarios with limited data availability.

Interestingly, the difference between Precision and Recall values during external evaluation appears to be smaller compared to what was demonstrated in Table 2. This suggests that using the subset sourced from Radiopaedia.org for training and validation, and the subset sourced from Coronacases.org for external evaluation, may lead to models with improved ability to generalize.

Table 4. Ablation study results when evaluating variations of the U-Net architecture on the validation and external validation sets, with data split at the slice level. Training and validation utilized 10 CT scans sourced from the Radiopaedia.org repository, while the external validation set comprised the remaining 10 CT scans from the Coronacases.org repository

Model	Validation Mean DSC	Validation Mean IoU	Validation Mean Precision	Validation Mean Recall	External Mean DSC	External Mean IoU	External Mean Precision	External Mean Recall
U-Net	0.7909 ± 0.0051	0.6541 ± 0.0070	0.7962 ± 0.0253	0.7864 ± 0.0154	0.6546 ± 0.0161	0.4867 ± 0.0180	0.6565 ± 0.0411	0.6559 ± 0.0379
U-Net + VGG16	0.8062 ± 0.0055	0.6753 ± 0.0078	0.7935 ± 0.0241	0.8201 ± 0.0155	0.7152 ± 0.0132	0.5568 ± 0.0159	0.7341 ± 0.0381	0.6991 ± 0.0239
U-Net + DenseNet121	0.8586 ± 0.0068	0.7523 ± 0.0105	0.8568 ± 0.0166	0.8609 ± 0.0158	0.7553 ± 0.0065	0.6069 ± 0.0084	0.7879 ± 0.0235	0.7261 ± 0.0185

Table 5. Running times of different variations of the U-Net architecture evaluated with the data split at the slice level. Training and validation used 10 CT scans sourced from the Radiopaedia.org repository, while the external validation set comprised the remaining 10 CT scans from the Coronacases.org repository

Model	Training Time	Validation Time	External Validation Time
U-Net	1596.20 ± 26.73 seconds	0.53 ± 0.04 seconds	6.47 ± 0.03 seconds
U-Net + VGG16	1805.55 ± 43.33 seconds	0.64 ± 0.05 seconds	7.88 ± 0.02 seconds
U-Net + DenseNet121	3075.20 ± 62.63 seconds	1.15 ± 0.04 seconds	12.17 ± 0.14 seconds

Table 5 showcases the training and inference times for the models discussed in Table 4. The results reveal that, as expected, the U-Net model exhibited the shortest runtime, while the U-Net + DenseNet121 model required the most time. Moreover, the training and validation times in this case are shorter, and the external validation time is longer compared to those in Table 3. This is expected, as the dataset used for training and validation, sourced from Radiopaedia.org, is notably smaller than the one used for external evaluation.

Figure 8 depicts the training and validation curves corresponding to the models whose results are displayed in Table 4. Upon examining these curves, it is evident that all the models converge, indicating no signs of overfitting.

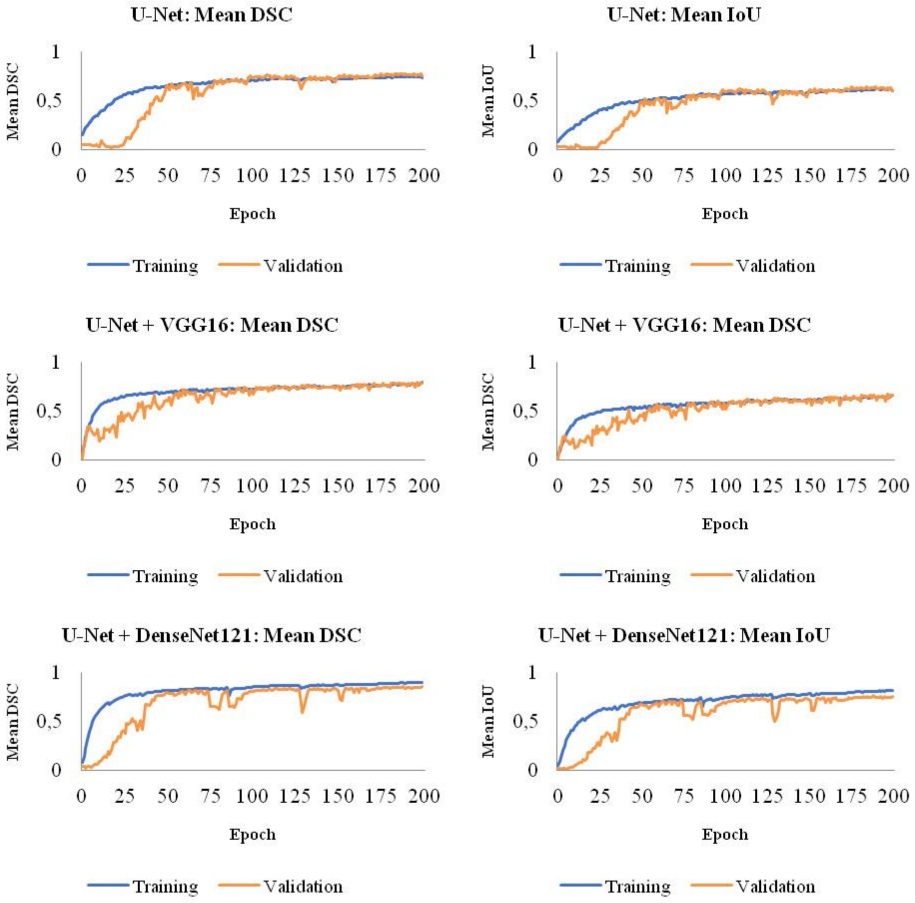

Fig. 8. Training and validation curves for the models trained and evaluated on data split at the slice level. Training and validation utilized 10 CT scans sourced from the Radiopaedia.org repository, while the external validation set comprised the remaining 10 CT scans from the Coronacases.org repository

4.2. Results when Splitting Data by Patient

In this section, we assess the performance of the selected models under a patient-wise data split. For the first experiment, we utilized 10 CT scans from Coronacases.org for training and validation, while 10 CT scans originating from Radiopaedia.org were used for external evaluation. The performance metrics achieved during validation and external evaluation are summarized in Table 6.

Once again, the U-Net + DenseNet121 model exhibits superior performance on both the validation and external evaluation sets. Although the model's performance does not reach the levels observed when trained on data split by slice, it remains satisfactory.

Table 6 shows lower performance metrics compared to Table 2. It is important to note that Table 2 contains performance metrics obtained from models trained using a slice-wise data split, whereas here we have opted for a patient-wise data split. When data are split by slice, the training and validation sets contain more diverse data, which often leads to improved generalization during training. In contrast, when splitting data by patient, the limited number of patients poses challenges due to insufficient training data for the model to learn patterns and relationships. Furthermore, lower performance metrics when splitting by slice could be associated with data leakage [38], a phenomenon where information from the validation or test set is used during the training phase, potentially overestimating model performance [38].

Consistent with the findings in Table 2, all models achieve similar mean Precision and Recall on the validation set, although these values are lower than the corresponding ones reported in Table 2. Notably, the performance on the external evaluation dataset is significantly lower, with Recall being lower than Precision, mirroring the results observed when using a slice-wise data split.

Table 6. Ablation study results when evaluating variations of the U-Net architecture on the validation and external validation sets, with data split at the patient level. Training and validation utilized 10 CT scans sourced from the Coronacases.org repository, while the external validation set comprised the remaining 10 CT scans from the Radiopaedia.org repository

Model	Validation Mean DSC	Validation Mean IoU	Validation Mean Precision	Validation Mean Recall	External Mean DSC	External Mean IoU	External Mean Precision	External Mean Recall
U-Net	0.7637 ± 0.0678	0.6216 ± 0.0883	0.7394 ± 0.0894	0.7912 ± 0.0442	0.6613 ± 0.0372	0.4940 ± 0.0404	0.7302 ± 0.1209	0.6392 ± 0.1431
U-Net + VGG16	0.7669 ± 0.0629	0.6253 ± 0.0830	0.7591 ± 0.0675	0.7759 ± 0.0673	0.7022 ± 0.0217	0.5414 ± 0.0258	0.7592 ± 0.0974	0.6654 ± 0.0639
U-Net + DenseNet121	0.7889 ± 0.0541	0.6540 ± 0.0738	0.8219 ± 0.0521	0.7603 ± 0.0706	0.7079 ± 0.0123	0.5480 ± 0.0146	0.8037 ± 0.0558	0.6369 ± 0.0466

The training and inference times of the models, whose metrics are presented in Table 6, are detailed in Table 7. Similar to the experiments presented in the previous section, the results demonstrate that the U-Net model had the shortest runtime, while the U-Net + DenseNet121 model required the longest time.

The training and validation curves for the models, whose results are depicted in Table 6, are shown in Figure 9. Upon examining these curves, it is notable that all models exhibit convergence. However, unlike the curves presented in the case of data split by slice, it is apparent that the models show signs of overfitting. This could be attributed to the potential lack of diversity in the data when splitting by patient. Furthermore, the reduced incidence of overfitting when splitting by slice could be linked to data leakage.

Within Table 8, we present the metrics used to assess the ability of models trained with a patient-based split approach to accurately segment lesions attributed to SARS-CoV-2. For this evaluation, 10 CT scans sourced from Radiopaedia.org were used for training and validation, while the remaining CT scans from Coronacases.org were used for external evaluation.

In line with all previous experiments, the U-Net + DenseNet121 model demonstrates superior overall performance. Similar to our previous experiment, where we trained and validated using data from the Radiopaedia.org repository, we note that the difference between Precision and Recall values during external evaluation is smaller compared to instances where our models were trained on data from the Coronacases.org repository.

Table 7. Running times of different variations of the U-Net architecture evaluated with the data split at the patient level. Training and validation used 10 CT scans sourced from the Coronacases.org repository, while the external validation set comprised the remaining 10 CT scans from the Radiopaedia.org repository

Model	Training Time	Validation Time	External Validation Time
U-Net	4176.32 ± 143.44 seconds	1.36 ± 0.28 seconds	2.41 ± 0.02 seconds
U-Net + VGG16	4684.79 ± 157.58 seconds	1.64 ± 0.35 seconds	2.92 ± 0.01 seconds
U-Net + DenseNet121	7613.26 ± 223.34 seconds	2.45 ± 0.50 seconds	4.45 ± 0.10 seconds

Table 8. Ablation study results when evaluating variations of the U-Net architecture on the validation and external validation sets, with data split at the patient level. Training and validation utilized 10 CT scans sourced from the Radiopaedia.org repository, while the external validation set comprised the remaining 10 CT scans from the Coronacases.org repository

Model	Validation Mean DSC	Validation Mean IoU	Validation Mean Precision	Validation Mean Recall	External Mean DSC	External Mean IoU	External Mean Precision	External Mean Recall
U-Net	0.7631 ± 0.0813	0.6224 ± 0.1035	0.7845 ± 0.0842	0.7568 ± 0.1285	0.6724 ± 0.0380	0.5075 ± 0.0428	0.6975 ± 0.0746	0.6621 ± 0.0890
U-Net + VGG16	0.7755 ± 0.0079	0.6387 ± 0.1032	0.7680 ± 0.0699	0.7872 ± 0.1056	0.6895 ± 0.0437	0.5275 ± 0.0512	0.7227 ± 0.0433	0.6631 ± 0.0690
U-Net + DenseNet121	0.7989 ± 0.0615	0.6686 ± 0.0835	0.7905 ± 0.0762	0.8146 ± 0.0925	0.7369 ± 0.0189	0.5837 ± 0.0238	0.7642 ± 0.0420	0.7152 ± 0.0484

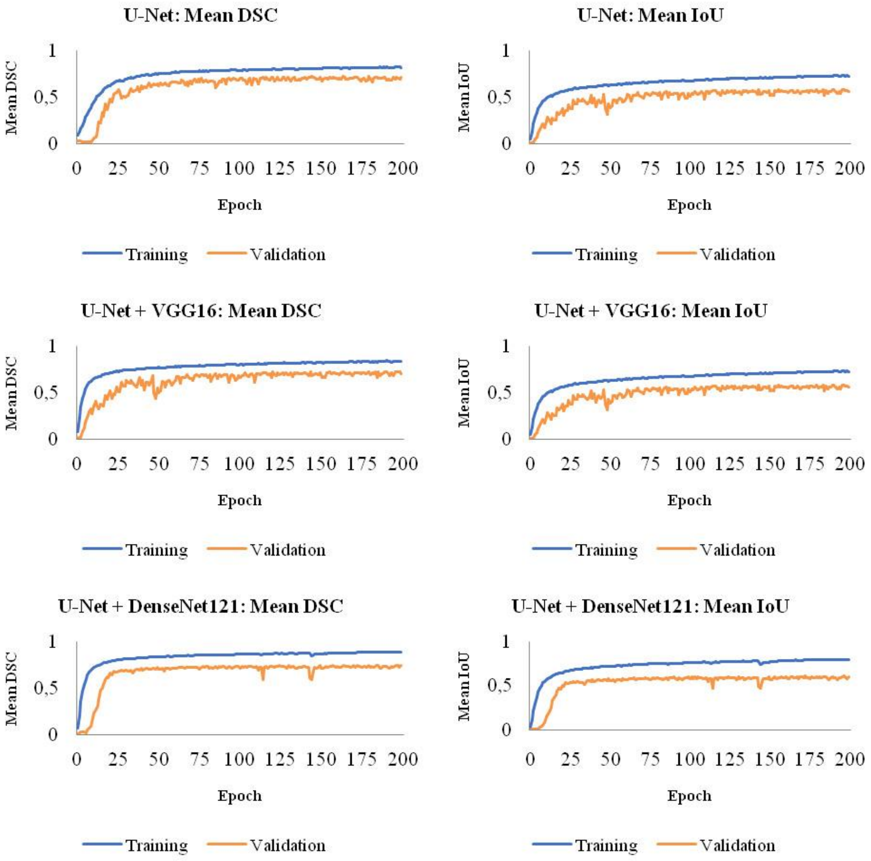

Fig. 9. Training and validation curves for the models trained and evaluated on data split at the patient level. Training and validation utilized 10 CT scans sourced from the Coronacases.org repository, while the external validation set comprised the remaining 10 CT scans from the Radiopaedia.org repository

Table 9. Running times of different variations of the U-Net architecture evaluated with the data split at the patient level. Training and validation used 10 CT scans sourced from the Radiopaedia.org repository, while the external validation set comprised the remaining 10 CT scans from the Coronacases.org repository

Model	Training Time	Validation Time	External Validation Time
U-Net	1581.43 ± 97.51 seconds	0.50 ± 0.18 seconds	6.48 ± 0.02 seconds
U-Net + VGG16	1799.78 ± 83.24 seconds	0.63 ± 0.24 seconds	7.90 ± 0.04 seconds
U-Net + DenseNet121	3101.80 ± 144.39 seconds	1.06 ± 0.36 seconds	12.13 ± 0.21 seconds

Table 9 provides a breakdown of the training and inference times for the models that achieved the performance metrics displayed in Table 8. Once again, the U-Net model had the shortest runtime, in contrast to the U-Net + DenseNet121 model, which required the longest time for training, validation and external evaluation.

In Figure 10, the training and validation curves for the models, whose results are shown in Table 8, are presented. In the experiment conducted for this section, all trained models converge. However, it is evident that each model exhibits some degree of overfitting.

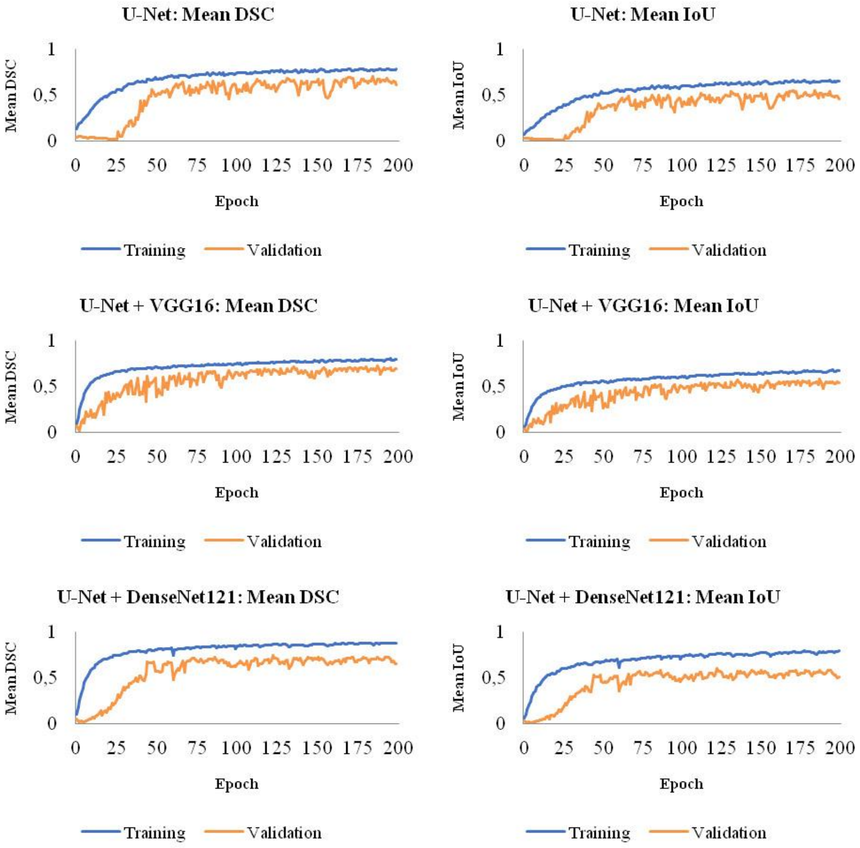

Fig. 10. Training and validation curves for the models trained and evaluated on data split at the patient level. Training and validation utilized 10 CT scans sourced from the Radiopaedia.org repository, while the external validation set comprised the remaining 10 CT scans from the Coronacases.org repository

4.3. Results of the Application of Semantic Segmentation on a Subset of the Validation and External Validation Datasets

Figures 11 and 12 illustrate the results of applying semantic segmentation to subsets of the validation and external evaluation datasets, respectively. Upon examining these results, it is evident that all models effectively isolate lesions present in the CT slices.

In addition, a quantitative analysis was conducted, and the results are presented in Tables 10 and 11. Notably, the U-Net + DenseNet121 demonstrated the most accurate segmentation among the models on both the validation and external evaluation datasets, as evidenced by its superior DSC and IoU values in Tables 10 and 11.

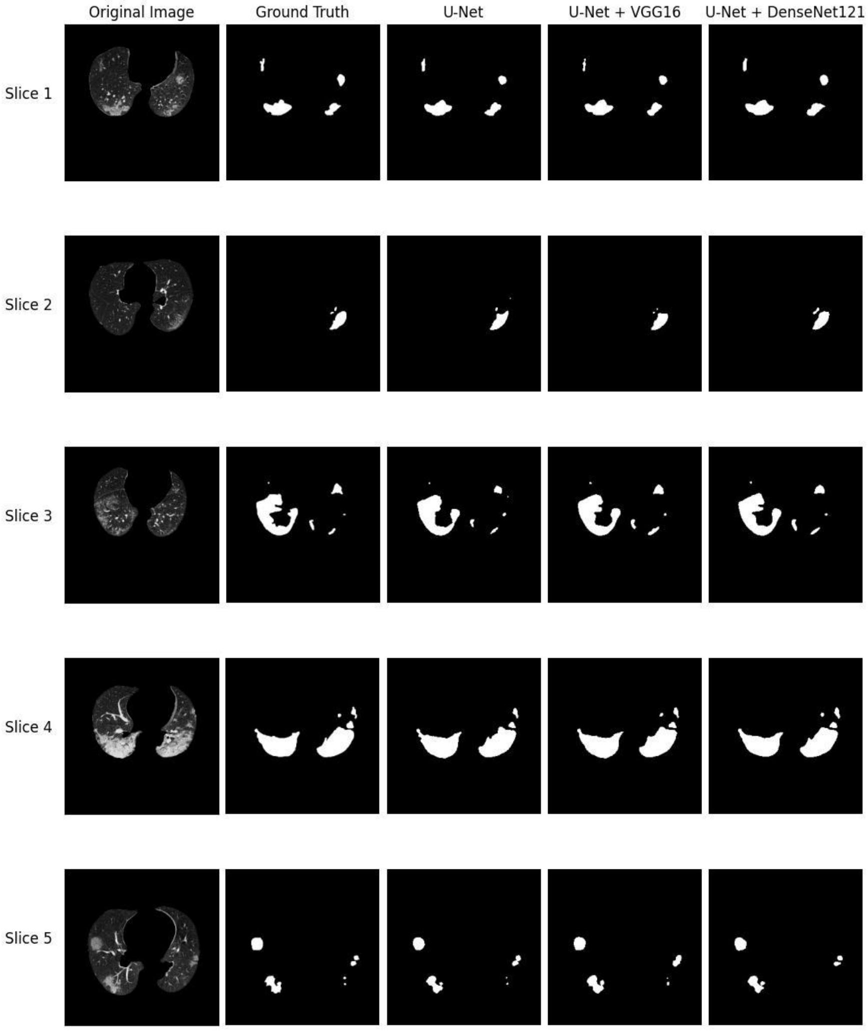

Fig. 11. Semantic segmentation results on a subset of the validation dataset

Table 10. Quantitative analysis of the results displayed in Figure 11

Slice	U-Net		U-Net + VGG16		U-Net + DenseNet121	
	DSC	IoU	DSC	IoU	DSC	IoU
Slice 1	0.8622	0.7577	0.8818	0.7886	0.8830	0.7905
Slice 2	0.8170	0.6906	0.8475	0.7353	0.8927	0.8062
Slice 3	0.8460	0.7331	0.8828	0.7901	0.8988	0.8162
Slice 4	0.9071	0.8300	0.9448	0.8953	0.9461	0.8977
Slice 5	0.8638	0.7602	0.8585	0.7521	0.8653	0.7626

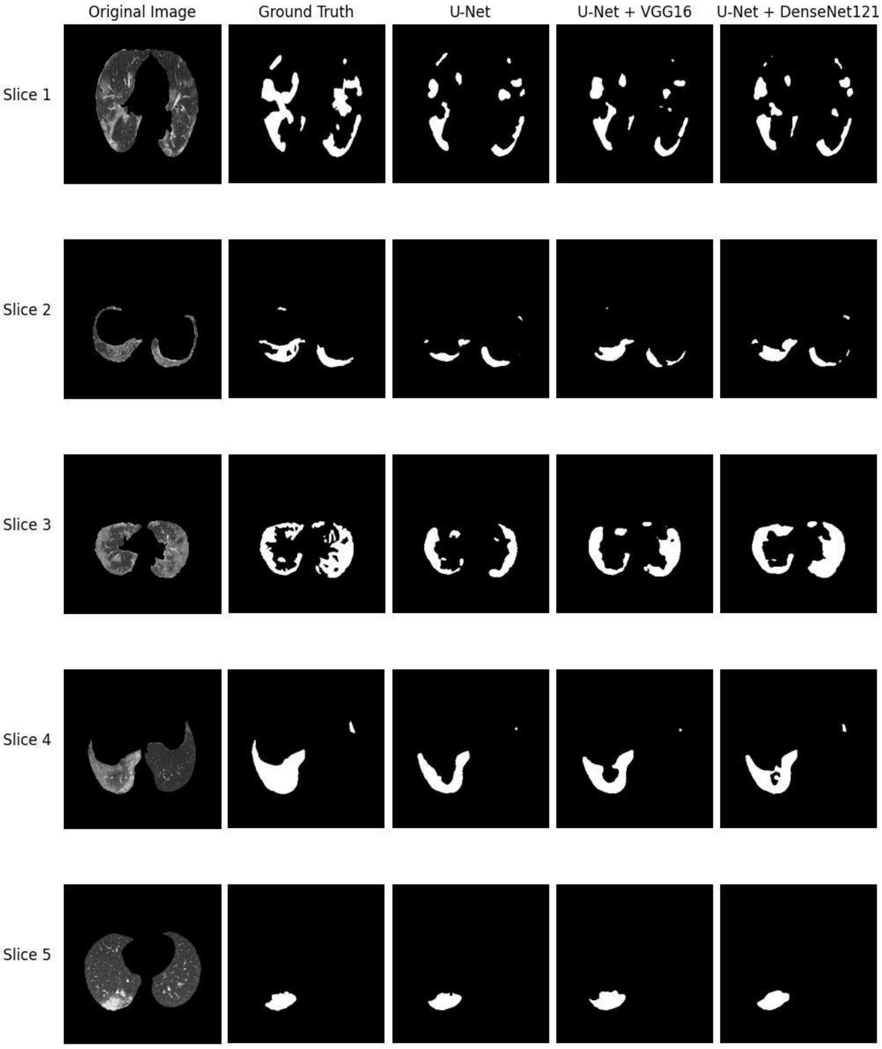

Fig. 12. Semantic segmentation results on a subset of the external validation dataset

Table 11. Quantitative analysis of the results displayed in Figure 12

Slice	U-Net		U-Net + VGG16		U-Net + DenseNet121	
	DSC	IoU	DSC	IoU	DSC	IoU
Slice 1	0.6218	0.4511	0.6417	0.4724	0.6627	0.4955
Slice 2	0.6537	0.4856	0.7074	0.5472	0.7283	0.5726
Slice 3	0.6624	0.4952	0.7595	0.6122	0.8020	0.6695
Slice 4	0.7669	0.6219	0.8131	0.6851	0.8444	0.7307
Slice 5	0.8521	0.7423	0.8704	0.7705	0.8853	0.7942

4.4. Statistical Test Analysis of the Results

In this section, we conduct a comprehensive statistical analysis of the results derived from our experiments. Our goal is to compare the performance of the best and worst models, namely U-Net + DenseNet121 and U-Net, respectively. To assess the performance of our models across different cross-validation data partitions, we employed a statistical test called t-test.

A t-test is used to compare the means of two groups [39]. There are two types of t-tests: the independent t-test, which compares the means of two groups that are unrelated to each other, and the paired t-test which compares the means of two groups that are related to each other [39].

Since we compare two CNN models that have been trained and evaluated on the same data using a 5-fold cross-validation scheme, the appropriate t-test to use is the paired t-test. This is because the same data folds are used for evaluating both models, meaning the results from each fold are paired. The aim of an analysis using a paired t-test is to discern whether there exists a statistically significant difference in the models' mean performance scores.

The paired t-test employs two contradictory research hypotheses: the null hypothesis and the alternative hypothesis [40]. The null hypothesis states that the mean difference between the paired observations is zero [40]. The alternative hypothesis states that the mean difference between the paired observations is not zero [40].

The steps to compute the paired t-test are summarized below:

1. Calculate the difference between each pair of observations:

$$d_i = y_i - x_i \tag{5}$$

2. Calculate the mean difference:

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{6}$$

3. Calculate the standard deviation of the differences:

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})} \tag{7}$$

4. Calculate the t-statistic:

$$t = \frac{\overline{d}}{s_d}\sqrt{n} \tag{8}$$

Under the null hypothesis, this statistic follows a t-distribution with n-1 degrees of freedom

5. Find the p-value corresponding to the calculated t-statistic.

6. Compare the p-value to the chosen significance level:

- If the p-value is less than the significance level (commonly 0.05), reject the null hypothesis.
- If the p-value is greater than the significance level, fail to reject the null hypothesis.

Table 12 presents the p-values obtained from the paired t-test. In cases where the p-values are less than 0.05, there exists a statistically significant difference in the performance of the compared models. By examining Table 12, we can see that all p-values are less than 0.05. This indicates a statistically significant difference in the performance between U-Net and U-Net + DenseNet121.

Table 12. Statistical test analysis of the U-Net and U-Net + DenseNet121 models. Training and validation utilized 10 CT scans sourced from the Coronacases.org repository, while the external validation set comprised the remaining 10 CT scans from the Radiopaedia.org repository.

Split by Slice				Split by Patient			
Validation		External Validation		Validation		External Validation	
DSC	IoU	DSC	IoU	DSC	IoU	DSC	IoU
0.0026	0.0022	0.0231	0.0202	0.0263	0.0175	0.0093	0.0080

4.5. Comparison of our Results with the Results of other Researchers

In this paper, we employed deep learning methods, specifically convolutional neural networks, to assess their efficacy, after appropriate training, in detecting and outlining lesions in CT slices caused by COVID-19 pneumonia. This proved to be a challenging task for two primary reasons. Firstly, collecting a large enough dataset posed difficulties due to the necessity of annotations provided by medical experts. Secondly, the prevalence of the negative class in medical images contributed to evident class imbalance within the data. To address these issues, we proposed leveraging transfer learning by replacing the encoder of a standard U-Net model with a pre-trained model.

Results indicate that the quality and quantity of the utilized dataset, as well as the use of ideal preprocessing techniques, were critical for our models' performance. Segmentation of the lung parenchyma and normalization in terms of contrast and brightness helped achieve better overall performance. On the other hand, data augmentation was not that helpful.

Employing two different splitting techniques provided us with the opportunity to compare our method with the literature, as some authors chose to split by patient instead of by slice and vice versa.

Verifying our approach using both subsets of the dataset for training, validation and external evaluation showed similar performance. This observation indicates the robustness of the models and their strong generalization capabilities.

Upon observing the performance of the selected models during semantic segmentation on both the validation and external validation datasets, it becomes evident that the standard U-Net model achieves significantly improved performance when transfer learning is applied. Moreover, the results on the external evaluation datasets demonstrate high generalization ability.

Table 13. Comparison of our work with other scientific papers where the data were split by slice.

Model Architecture	Training and Validation Dataset	External Evaluation Dataset	Preprocessing & Transfer Learning	Validation Results	External Evaluation Results
Xiaoyan Wang et al. [22] SSA-Net	Experiment 1: COVID-19-CT-Seg (5-fold cross validation) Experiment 2: MedSeg Dataset [23] (Only 98 CT Slices) (5-fold cross validation)	-	HU Clipping [-1250, 250], No Transfer Learning	Experiment 1: DSC: 65.22% Experiment 2: DSC: 75.40%	-
Krinski et al. [24] Various CNN Models	COVID-19-CT-Seg (5-fold cross validation Train: 80% Val: 20%)	-	Transfer Learning (ImageNet [25])	Best Model DSC: 73.67% Best Model IoU: 70.91%	-
Mahmoudi et al. [26] 2D U-Net	COVID-19-CT-Seg (4-fold cross validation Train: 70% Val: 30%)	-	CLAHE, Image Cropping, Data Augmentation, No Transfer Learning	DSC: 91% IoU: 85%	-
Qiblawey et al. [27] Various CNN Models	COVID-19-CT-Seg (10-fold cross validation Train: 60% Val: 20% Test: 20%)	-	HU Normalization, Lung Parenchyma Segmentation, Data Augmentation, Transfer Learning (ImageNet)	Best Model DSC: 94.13% Best Model IoU: 91.85%	-
Our Approach U-Net + DenseNet121	Experiment 1: COVID-19-CT-Seg (10 CT from Coronacases) (5-fold cross-validation) Experiment 2: COVID-19-CT-Seg (10 CT from Radiopaedia) (5-fold cross-validation)	Experiment 1: COVID-19-CT-Seg (10 CT from Radiopaedia) Experiment 2: COVID-19-CT-Seg (10 CT from Coronacases)	HU Normalization WW: 1400 WL: -500, Lung Segmentation, Data Augmentation Transfer Learning (ImageNet)	Experiment 1: DSC: 86.44% IoU: 76.12% Experiment 2: DSC: 85.86% IoU: 75.23%	Experiment 1: DSC: 72.11% IoU: 56.41% Experiment 2: DSC: 75.53% IoU: 60.69%

Through our experiments, we found that increasing the batch size results in better overall model performance, while increasing the depth of the U-Net model by adding extra layers or changing layer parameters increases the model complexity without significant performance gains.

Another distinctive aspect of our approach that sets it apart from the works of other researchers is the utilization of 2D U-Nets instead of 3D U-Net architectures, which are less computationally expensive. Upon reviewing Tables 13 and 14 it is evident that our best model outperforms the majority of the models featured in other works, demonstrating enhanced performance. Analytically, when splitting by slice and training on the data from the Coronacases.org repository, our best model achieves a DSC and IoU score of 0.8644 and 0.7612 during the validation phase and a DSC and IoU score of 0.7211 and 0.5641 during the external evaluation phase. When splitting by slice and training on the data from the Radiopaedia.org repository, our best model achieves a DSC and IoU score of 0.8586 and 0.7523 during the validation phase and a DSC and IoU score of 0.7553 and 0.6069 during the external evaluation phase. When splitting by patient and training on the data from the Coronacases.org repository, our best model achieves a DSC and IoU score of 0.7889 and 0.6540 during the validation phase and a DSC and IoU score of 0.7079 and 0.5480 during the external evaluation phase. When splitting by patient and validating on the data from the Radiopaedia.org repository, our best model achieves a DSC and IoU score of 0.7989 and 0.6686 during the validation phase and a DSC and IoU score of 0.7369 and 0.5837 during the external evaluation phase. This improvement is crucial for clinical applications, as accurate segmentation of COVID-19 lesions can lead to better monitoring of disease progression and response to treatment. Automated and reliable identification of affected lung regions can help radiologists quantify the extent of disease more efficiently, enabling timely adjustments in patient management strategies. This could potentially reduce diagnostic errors and improve patient outcomes by ensuring that critical cases are identified and treated promptly.

By further examining the results in Table 13, it is evident that our best model is surpassed by the models trained in studies [26] and [27] in terms of performance. While our work demonstrates lower DSC and IoU scores compared to [27], it is important to consider the difference in our data preprocessing strategies. In [27], the authors applied data augmentation techniques before splitting the dataset into training and validation sets. This approach may introduce a risk of data leakage, as slightly modified images could be present in both training and validation sets potentially overestimating model performance during evaluation. In contrast, our method follows a better practice by performing data augmentation after the data split and exclusively on the training set. Regarding [26], their strategy of cropping the CT slices to the size 256x256 likely contributed to their models achieving higher DSC and IoU scores, as it may have helped them better focus on anatomical features relevant to the task. As shown in the results of Table 14, our best model is slightly outperformed by the model presented in [13], as well as by the model implemented in [15].

Table 14. Comparison of our work with other scientific papers where the data were split by patient

Model Architecture	Training and Validation Dataset	External Evaluation Dataset	Preprocessing & Transfer Learning	Validation Results	External Evaluation Results
Ma et al. [7] 2D U-Net	COVID-19-CT-Seg [8] (5-fold cross-validation Train: 20% Val: 80%)	-	HU Clipping [-1250, 250], No Transfer Learning	DSC: 60.80%	-
Ma et al. [7] nnU-Net [9]	COVID-19-CT-Seg (5-fold cross-validation Train: 20% Val: 80%)	MosMed [10]	HU Clipping [-1250, 250], No Transfer Learning	DSC: 67.30%	DSC: 58.80%
Müller et al. [11] 3D U-Net [12]	COVID-19-CT-Seg (5-fold cross-validation Train: 80% Val: 20%)	-	HU Clipping [-1250, 250], No Transfer Learning	DSC: 76.10%	-
Müller et al. [13] 3D U-Net	COVID-19-CT-Seg (5-fold cross-validation Train: 80% Val: 20%)	An et al. [14]	HU Clipping [-1250, 250], Data Augmentation, No Transfer Learning	DSC: 80.40%	DSC: 66.10%
Owais et al. [15] DAL-Net	Experiment 1: COVID-19-CT-Seg (5-fold cross-validation Train: 80% Val: 20%) Experiment 2: MosMed (5-fold cross-validation Train: 80% Val: 20%) Experiment 3: COVID-19-CT-Seg	Experiment 3: MosMed	Experiment 3: Reinhard Transformation [16], No Transfer Learning	Experiment 1: DSC: 83.23% IoU: 74.86% Experiment 2: DSC: 68.63% IoU: 61.35%	Experiment 3: DSC: 74.93% IoU: 66.50%
Zheng et al. [17] 3D CU-Net	COVID-19-CT-Seg (5-fold cross-validation Train: 80% Val: 20%)	MosMed	HU Clipping [-1250, 250], Data Augmentation, No Transfer Learning	DSC: 77.80%	DSC: 66.80%
Yixin Wang et al. [18] 3D U-Net	COVID-19-CT-Seg (5-fold cross-validation Train: 20% Val: 80%)	-	No Transfer Learning	DSC: 70.04%	-
Our Approach U-Net + DenseNet121	Experiment 1: COVID-19-CT-Seg (10 CT from Coronacases) (5-fold cross-validation) Experiment 2: COVID-19-CT-Seg (10 CT from Radiopaedia) (5-fold cross-validation)	Experiment 1: COVID-19-CT-Seg (10 CT from Radiopaedia) Experiment 2: COVID-19-CT-Seg (10 CT from Coronacases)	HU Normalization WW: 1400 WL: -500, Lung Segmentation, Data Augmentation Transfer Learning (ImageNet)	Experiment 1: DSC: 78.89% IoU: 65.40% Experiment 2: DSC: 79.89% IoU: 66.86%	Experiment 1: DSC: 70.79% IoU: 54.80% Experiment 2: DSC: 73.69% IoU: 58.37%

5. Conclusions and Future Work

The primary goal of this study was to investigate the ability of deep learning methods to accurately segment lesions in CT slices caused by pneumonia attributed to SARS-CoV-2. Our focus was on evaluating the performance of a U-Net architecture and two variations of it, where the encoder was replaced with pre-trained convolutional neural networks. The outcomes, presented in Section 4 and compared with the literature in Section 2, indicate that the objectives of this research have been achieved.

This study lays the groundwork for potential future extensions aimed at enhancing the robustness and applicability of the proposed models. Future work could involve collecting more diverse data, featuring variations in patient age, gender and ethnicity from various healthcare facilities. This would allow for further training of the models using more varied datasets. Additionally, training the models on higher-resolution images could enhance segmentation precision and overall model performance. Another potential direction could involve adapting our segmentation method for other respiratory diseases where accurate lesion segmentation is equally critical. Moving our method from research to clinical practice is also a promising prospect. This would require further steps, including extensive clinical trials and validation studies to ensure the robustness and reliability of the segmentation tool in real-world scenarios. Collaborating with healthcare providers to integrate our tool into hospital information systems and workflows will be essential for practical implementation.

Tuning the hyperparameters of the deep learning models automatically using methods such as Grid Search and Random Search could potentially improve model performance. These methods were omitted in this work due to their computational expense. Implementing regularization techniques, such as Lasso and Ridge Regression, could help address the overfitting phenomenon observed when data are split by patient.

References

1. WHO Coronavirus (COVID-19) Dashboard. [Online]. Available: https://covid19.who.int/ (current July 2023)
2. Chen, Y. J., Jian, W. H., Liang, Z. Y., Guan, W. J., Liang, W. H., Chen, R. C., Tang, C. L., Wang, T., Liang, H. R., Li Y. M., et al.: Earlier diagnosis improves COVID-19 prognosis: a nationwide retrospective cohort analysis. Annals of Translational Medicine, Vol. 9, No. 11, 941. (2021)
3. Machnicki, S., Patel, D., Singh, A., Talwar, A., Mina, B., Oks, M., Makkar, P., Naidich, D., Mehta, A., Hill, N. S., et al.: The Usefulness of Chest CT in Patients with Suspected or Diagnosed COVID-19: A Review of Literature. Chest, Vol. 60, No. 2, 652-670. (2021)
4. Borakati, A., Perera, A., Johnson, J., Sood, T.: Diagnostic accuracy of X-ray versus CT in COVID-19: a propensity matched database study. BMJ Open, Vol. 10, No. 11. (2020)
5. Pinto, A., Brunese, L.: Spectrum of diagnostic errors in radiology. World Journal of Radiology, Vol. 2, No. 10, 377-383. (2010)
6. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer-Verlag, Munich, Germany, 234-241. (2015)

7. Ma, J., Wang, Y., An, X., Ge, C., Yu, Z., Chen, J., Zhu, Q., Dong, G., He, J., He, Z., et al.: Toward data-efficient learning: A benchmark for COVID-19 CT lung and infection segmentation. Medical Physics, Vol. 48, No. 3, 1197-1210. (2021)
8. Ma, J., Ge, C., Wang, Y., An, X., Jiantao, G., Ziqi, Y., Zhang, M., Liu, X., Deng., X., Cao, S., et al.: COVID-19 CT Lung and Infection Segmentation Dataset. (2020). [Online]. Available: https://zenodo.org/records/3757476 (current July 2023).
9. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., Maier-Hein, K. H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, Vol. 18, 203-211. (2021)
10. Morozov, S. P., Andreychenko, A. E., Blokhin, I. A., Gelezhe, P. B., Gonchar, A. P., Nikolaev, A. E., Pavlov, N. A., Chernina, V. Y., Gombolevskiy, V. A.: MosMedData: data set of 1110 chest CT scans performed during the COVID-19 epidemic. Digital Diagnostics, Vol. 1, No. 1, 49-59. (2020)
11. Müller, D., Rey, I. S., Kramer, F.: Automated Chest CT Image Segmentation of COVID-19 Lung Infection based on 3D U-Net. (2020). [Online]. Available: https://arxiv.org/abs/2007.04774 (current July 2023)
12. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Bronx, T., Ronnebeger, O.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Proceedings of the 19th International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer-Verlag, Athens, Greece, 424-432. (2016)
13. Müller, D., Rey, I. S., Kramer, F.: Robust chest CT image segmentation of COVID-19 lung infection based on limited data. Informatics in Medical Unlocked, Vol. 25, 100681. (2021)
14. An, P., Xu, S., Harmon, S. A., Turkbey, E.B., Sanford, T. H., Amalou, A., Kassin, M., Varble, N., Blain, M., Anderson, V., et al.: CT Images in COVID-19. [Online]. Available: https://wiki.cancerimagingarchive.net/display/Public/CT+Images+in+COVID-19 (current December 2023).
15. Owais, M., Baek, N. R., Park, K. R.: Domain-Adaptive Artificial Intelligence-Based Model for Personalized Diagnosis of Trivial Lesions Related to COVID-19 in Chest Computed Tomography Scans. Journal of Personalized Medicine, Vol. 11, No. 10, 1008. (2021)
16. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications, Vol. 21, No. 5, 34-41. (2001)
17. Zheng, R., Zheng, Y., Dong-Ye, C.: Improved 3D U-Net for COVID-19 Chest CT Image Segmentation. Scientific Programming, Vol. 2021, No. 1. (2021)
18. Wang, Y., Zhang, Y., Liu, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., He, Z.: Does non-COVID-19 lung lesion help? Investigating transferability in COVID-19 CT image segmentation. Computer Methods and Programs in Biomedicine, Vol. 202, 106004. (2021)
19. Singh, V. K., Abdel-Nasser, M., Pandey, N., Puig, D.: LungINFseg: Segmenting COVID-19 Infected Regions in Lung CT Images Based on a Receptive-Field-Aware Deep Learning Framework. Diagnostics, Vol. 11, No. 20, 158. (2021)
20. Amara, K., Aouf, A., Kennouche, H., Djekoune A. O., Zenati, N., Kerdjidj, O.; Ferguene, F.: COVIR: A virtual rendering of a novel NN architecture O-Net for COVID-19 Ct-scan automatic lung lesions segmentation. Computers & Graphics, Vol. 104, 11-23. (2022)
21. Aswathy, A.L., Chandra, S. S. V.: Cascaded 3D UNet architecture for segmenting the COVID-19 infection from lung CT volume. Scientific Reports, Vol. 12, 3090. (2022)
22. Wang, X., Yuan, Y., Guo, D., Huang, X., Cui, Y., Xia, M., Wang, Z., Bai, C., Chen, S.: SSA-Net: Spatial self-attention network for COVID-19 pneumonia segmentation with semi-supervised few-shot learning. Medical Image Analysis, Vol. 79, 102459. (2022)
23. COVID-19 – Medical segmentation. [Online]. Available: http://medicalsegmentation.com/covid19 (current July 2023)
24. Krinski, B. A., Ruiz, D. V., Todt, E.: Spark in the Dark: Evaluating Encoder-Decoder Pairs for COVID-19 CT's Semantic Segmentation. In Proceedings of the 2021 Latin American

Robotics Symposium (LARS), 2021 Brazilian Symposium on Robotics (SBR), and 2021 Workshop on Robotics in Education (WRE). IEEE, Natal, Brazil. (2021)

25. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Miami, Florida, USA, 248-255. (2009)
26. Mahmoudi, R., Benameur, N., Mabrouk, R., Mohammed M. A., Gacia-Zapirain, B., Bedoui, M. H.: A Deep Learning-Based Diagnosis System for COVID-19 Detection and Pneumonia Screening Using CT Imaging. Applied Sciences, Vol. 12, No. 10, 4825. (2022)
27. Qiblawey, Y., Tahir, A., Chowdhury, M. E. H., Khandakar, A., Kiranyaz, S., Rahman, T., Ibtehaz, N., Mahmud, S., Maadeed, S. A., Musharavati, F. et al: Detection and Severity Classification of COVID-19 in CT Images Using Deep Learning. Diagnostics, Vol. 11, No. 5, 893. (2021)
28. Enshaei, N., Oikonomou, A., Rafiee, M. J., Afshar, P., Heidarian, S., Mohammadi, A., Plataniotis, K. N., Naderkhani, F.: COVID-rate: an automated framework for segmentation of COVID-19 lesions from chest CT images. Scientific Reports, Vol. 12, No. 1, 3212. (2022)
29. Afshar, P., Shahin, H., Enshaei, N., Naderkhani, F., Rafiee, M. J., Oikonomou, A., Fard, F. B., Samimi, K., Plataniotis, K. N., Mohammadi, A.: COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning. Scientific Reports, Vol. 8, 121. (2021)
30. Uçar, M.: Automatic segmentation of COVID-19 from computed tomography images using modified U-Net model-based majority voting approach. Neural Computing and Applications, Vol. 34, No. 24, 21927-21938. (2022)
31. CORONACASES.ORG – by RAIOSS.com. [Online]. Available: https://coronacases.org/ (current February 2023).
32. Playlist 'COVID-19 pneumonia' by Dr Yaïr Glick. [Online]. Available: https://radiopaedia.org/playlists/25887/ (current February 2023)
33. Chau, S., Hayre, C. M.: Computed Tomography: A Primer for Radiographers. CRC Press, Boca Raton, Florida, USA. (2022)
34. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations. San Diego, California, USA. (2015)
35. Huang, G.. Liu, Z.. van der Maaten, L.. Weinberger, K. Q.: Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Honolulu, Hawaii, USA, 4700-4708. (2017)
36. Kaggle: your machine learning and data science community. [Online]. Available: https://www.kaggle.com/ (current February 2023).
37. Kingma, D. P, Ba, J.: Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations. San Diego, California, USA. (2015)
38. Yagis, E., Atnafu, S. W., García Seco de Herrera, A., Marzi, C., Scheda, R., Giannelli, M., Tessa, C., Citi, L., Diciotti, S.: Effect of data leakage in brain MRI classification using 2D convolutional neural networks. Scientific Reports, Vol. 11, No. 1, 22544. (2021)
39. Kim, T. K.: T test as a parametric statistic. Korean Journal of Anesthesiology, Vol. 68, No. 6, 540-546. (2015)
40. Hasija, Y.: All About Bioinformatics: From Beginner to Expert. Academic Press, London, UK. (2023)

Symeon Psaraftis-Souranis received his M.Eng. degree from the Department of Informatics and Computer Engineering at the University of West Attica in 2023. During his studies, he worked on several academic projects, gaining hands-on experience in applying artificial intelligence across various fields, including medicine. He has developed strong proficiency in programming languages and artificial intelligence applications, with a keen interest in computational biology. His research interests include artificial intelligence, machine learning, image processing, natural language processing, data mining, and bioinformatics.

Christos Troussas is currently an Assistant Professor in the Department of Informatics and Computer Engineering at the University of West Attica. He has been PostDoc Researcher in the field of Software Engineering and Applied Artificial Intelligence at the same Department. He has received a PhD degree in Informatics, a MSc degree with specialization in “Intelligent Technologies of Human-Computer Interaction” and a BSc degree in Informatics from the Department of Informatics, University of Piraeus. He has participated in national and international research projects. His academic influence is indicated by thousands of citations of his work and by having received best paper awards in international conferences in the field of computer science. His current research interests are in the areas of personalized software technology, human-computer interaction and applied artificial intelligence. He is among the world’s top 2% of scientists, according to the Stanford University ranking for the years 2020, 2021 and 2022 (Ioannidis, John P.A. (2023), “October 2023 data-update for "Updated science-wide author databases of standardized citation indicators"”, Elsevier Data Repository, V6, doi: 10.17632/btchxktzyw.6). Additionally, he holds the position of Research Fellow at INTI International University, Malaysia.

Athanasios (Thanos) Voulodimos is currently an Assistant Professor in the School of Electrical and Computer Engineering at the National Technical University of Athens (NTUA). He received his Dipl.-Ing., MSc, and PhD degrees from the same institution with the highest honors. His research interests include machine learning, artificial intelligence, image and signal processing, multimedia, and multimodal data fusion, management, and analysis. He has participated in more than 15 European and national research and development projects as a researcher, senior researcher, and/or technical manager. Dr. Voulodimos has co-authored more than 140 papers in refereed international journals, conference proceedings, and books, with his work receiving over 7,000 citations. He has served as an Organizing and Program Committee member at various international conferences and workshops and has guest-edited collective book volumes and special issues in international journals. He is a member of IEEE and ACM.

Cleo Sgouropoulou holds a Ph.D. in Electrical and Computer Engineering from the National Technical University of Athens (NTUA) and is currently a Professor in the Department of Informatics and Computer Engineering at the University of West Attica (UNIWA). Her research focuses on software engineering, with emphasis on the design, development, and standardization of Learning Technology and Research Information Systems. She has coordinated numerous national and European projects related to e-Learning, Open Educational Resources (OER), and skills modeling. Prof. Sgouropoulou has published over 140 articles and has received more than 3000 citations. She plays a key role in European Learning Technologies Standardization, leading the Greek delegation to CEN Technical Committees and contributing to projects on learning outcomes, digital skills, and learner mobility. Her work has resulted in several European

Norms, including the EuroLMAI, and she has been widely recognized within the field for her contributions.

Received: February 29, 2024; Accepted: September 09, 2024.