Diagnosis of the Stage of COVID-19 Disease from CT Scan Images of the Lung by Using a Swin Transformer

Amir Mohammad Hamedani [1], Mahsa Akhbari [1,*], Parisa Gifani [2]

[1]Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran, akhbari.mahsa@srbiau.ac.ir
[2]MRC London Institute of Medical Sciences, Imperial College London, London, UK, Pgifani@ic.ac.uk

Abstract

COVID-19 became a global pandemic, necessitating a strategy of testing, tracing, and isolating infected individuals. Accurately assessing disease severity from lung CT scans is crucial for treatment, but manual analysis is time-consuming and error-prone. This paper proposes an automated system using the Swin Transformer to identify COVID-19 severity levels from CT images. The authors used the smallest Swin Transformer model (Swin Tiny) and three popular networks: ResNet50, DenseNet121 convolutional networks, and ViT (ViT-base-patch16) for comparison. These pre-trained models were then fine-tuned on the training portion of the authors' dataset, and their performance was evaluated on the testing set. The dataset consisted of five classes, sourced from 689 individuals suspected of COVID-19. Each obtained 3D-CT scan was divided into 1 cm slices. After preprocessing and selecting the most informative images (those with a larger lung volume), a total of 10,902 2D images were extracted. The proposed network (Swin Transformer) achieved an accuracy of 0.95 on the test dataset after 100 epochs and 0.97 after 300 epochs, outperforming other models. Furthermore, this network outperformed all others in terms of Macro F1-Score, Macro recall, and Macro precision. This study demonstrates that novel self-attention-based neural networks can effectively detect and classify CT images of COVID-19 patients, outperforming well-known convolutional neural networks.

Keywords: COVID-19, Lung CT scan, Swin transform, Stages of disease, Diagnosis
Article history: Received 2024/12/24; Revised 2025/03/06; Accepted 2025/1104/02, Article Type: Research paper

1. Introduction

COVID-19 was caused by a virus that was first called the new Coronavirus and later SARS-CoV-2 (1). The most apparent symptoms of patients with COVID-19 are runny nose, cough, phlegm, fatigue, fever, loss of sense of smell, and shortness of breath (2, 3). Although this disease has more symptoms than the common cold and flu, due to having some common symptoms, it was confused with them at first (4). This delayed the identification of patients.

The COVID-19 virus generally enters the human respiratory system through the air and the nose. After that, it finds its way into the alveoli through the trachea, bronchioles, and cilia. Therefore, the patient's lungs are the main organs affected by the COVID-19 virus (5). When the infection enters the alveoli, type II cells secrete an inflammatory signal. In response to this signal, macrophages are recruited to the alveoli. This immune cell releases cytokine, which leads to the opening of blood vessels and the influx of more immune cells. This inflammatory response evokes alveolar fluid accumulation, diluting surfactant and hindering gas exchange. If this process continues, the alveoli will be destroyed, and acute respiratory distress syndrome (ARDS) will occur (6). These events lead to the creation of abnormal areas in the human lung that can be recognized on CT images by the eye. These abnormalities are observed in various forms, such as dilated intra-infiltrate vessels, ground glass opacities (GGO), rounded opacities, and consolidation (7).

Abnormalities created in the lungs of sick people can be seen easily through Computed Tomography (CT) scan images. Although there are other ways to diagnose COVID-19, a CT scan of the lung is the most accurate way, especially for the diagnosis of the stages of COVID-19 lung involvement, which is the aim of our research. The CT images of the lungs of people with COVID-19 can be divided into four stages based on the progress

of the disease in time, which are called (1) early stage, (2) progressive stage, (3) peak stage, and (4) absorption stage.

CT images are three-dimensional, in that each lung image (three-dimensional) contains ordinarily exceeding 250 images (two-dimensional), each of which displays a cross-sectional surface of the lung (8). Therefore, it is a time-consuming and challenging task for physicians to diagnose COVID-19 or its stage from images. Since the number of patients with COVID-19 is high during the epidemic, the possibility of physicians mistakes naturally increases. Also, some physicians may be unfamiliar with the pattern observed in the lungs of COVID-19 patients. Scientists in the field of artificial intelligence suggest using deep networks to solve these problems. During these years, deep learning networks achieved high accuracy in machine vision tasks in such a way that even in some studies, their accuracy has exceeded the accuracy of humans (9).

Detecting COVID-19 and determining its stage from CT images always comes with challenges that hinder accurate diagnosis. Just as a lack of expertise and experience among doctors and radiologists increases the likelihood of misdiagnosis, neural networks also suffer from reduced accuracy due to insufficient training data. Additionally, variations in imaging protocols (such as different devices, doses, and settings), low image quality caused by patient movement, scanner settings, and noise, as well as the similarity of COVID-19-related lung abnormalities to those of other diseases, are all factors that prevent achieving maximum accuracy.

Vision transformers are a type of neural network architecture that has gained popularity in the field of computer vision, because they utilize attention mechanisms to attend to important parts of the image. They are based on the transformer architecture, which was originally developed for natural language processing tasks (10). Vision transformers are designed to process visual data, such as images, and they calculate self-attention for pixels.

The main benefit of vision transformers over traditional convolutional neural networks (CNNs) is their ability to capture long-range dependencies in the input data. CNNs are typically designed to process local features in an image, which can limit their ability to capture global relationships between different parts of the image. Vision transformers, on the other hand, use self-attention mechanisms to capture both local and global dependencies, allowing them to better understand the context of different image regions (11).

Swin Transformer is a recent development in the field of vision transformers. It introduces a hierarchical architecture that divides the input image into smaller patches, allowing for more efficient processing of large images. This hierarchical approach enables the Swin Transformer to capture both local and global dependencies in the input data, making it particularly effective for tasks such as object detection and image classification. The benefits of using Swin Transformer in computer vision include its ability to handle large-scale visual data with high efficiency. Its hierarchical structure allows for parallel processing of image patches, enabling it to scale to larger input sizes without a significant increase in computational cost. Additionally, the Swin Transformer has been shown to achieve state-of-the-art performance on various computer vision tasks, making it a promising architecture for future research in this field (12, 13).

Most of the research conducted on the classification of COVID-19 has been done in the field of diagnosis and binary classification of COVID-19 (COVID and non-COVID). In (14), Li Zhang and Yan Wen proposed a framework based on the Swin transformer for automatic detection of COVID-19 using chest CT scan, which consists of a preprocessing step for lung segmentation by U-net, and classification has been done by using Swin transformer. They then compared this structure with two structures using BigTransfers (BiT-M) and EfficientNetV2 as the main body, on the COV19-CT-DB dataset (which contains about 5000 3D CT series). Finally, their proposed structure, which used Swin-B architecture which is one of the four Swin Transformer models, named for its size of parameters.in its classification section, could show a better performance than the other two structures by obtaining an F1 score of 0.935. In another study (7) with the same dataset, Dimitrios Kollias and his colleagues developed a simple network with CNN-RNN architecture which is Combination of convolutional neural network and recurrent neural network, using ResNet50, and used it as a baseline for comparison with the results of the ICCV 2021 COV19D COVID competition. In the following, they discussed the results of some models participating in this competition, the best of which had reached the F1 score of 0.904. This model was a deep learning network with the ResNet backbone architecture that utilizes the Periphery-aware Spatial Prediction (PSP) technique to extract high-level features and calculate the boundary distance map of pixels (used to identify the location of pixels within the lungs). Additionally, it employs the Contrastive Representation Enhancement (CRE) mechanism to enhance the similarity between data within the same class and distinguish between different classes, leading to better network training. The weakness of the models in the two mentioned studies, and all models trained on this dataset, is that due to the lack of labels for CT slices, they can only examine the whole CT scan (in 3D).

In some research, X-ray images have been used to classify COVID-19. For example, Juntao Jiang and Shuyi Lin (15) identified COVID-19 using a database with 17,955 chest X-ray images and with the help of a Swin transformer. They combined the Swin transformer and the Transformer in Transformer (TNT) with the weighted average method (ratio 2 to 1). Then they used it to classify X-ray images into three categories: normal, pneumonia, and COVID-19. Finally, with their proposed model, they achieved an accuracy of 0.94.

Chih-Chung Hsu and his colleagues in (16) presented two designs for identifying patients among suspected COVID-19 cases, one based on 2D-CT scans (conventional CT slice) and the other on 3D-CT scans (each series of CT scans). The 2D-CT scan-based design, named ADLEaST, employed adaptive distribution learning along with statistical hypothesis testing. In this design, the features of the dataset are extracted by a Swin Transformer and these features were divided into different probability distributions by a fully-connected layer.

Subsequently, outlier slice removal (removing slices that are unlikely to contain useful lung information) and Wilcoxon signed-rank test were performed on them to classify them into two groups: COVID-19 infected and non-infected. The proposed deep neural network is called ADLEaST and obtained an accuracy of 0.92 and an F1 score of 0.92, better than DenseNet201. In their 3D CT-based approach, they did not use statistical analysis and left feature extraction, slice importance selection, and classification to their proposed neural network, Convolutional CT-Aware Transformer (CCAT). This model used a ResNet50 along with two vision transformers and achieved an accuracy of 0.93 and an F1 score of 0.93.

The use of the Swin network is not limited to the field of COVID-19, and it has been used successfully in other areas as well, for example for coastal wetlands classification (17) and breast cancer classification (18).

The novelty of this paper lies in creating an accurate automated method based on the Swin Transformer architecture for COVID-19 diagnosis and severity determination (five classes: normal, early stage, progressive stage, peak stage, absorption stage) using a distinct dataset of CT images and transfer learning. Since this method, in addition to diagnosing COVID-19 patients from healthy individuals, also determines the severity of their disease, it can, in addition to helping radiologists, also guide physicians in determining treatment priorities and treatment methods for COVID-19 patients. In this research, in addition to the Swin Tiny pre-trained model, three well-known pre-trained Vision Transformer (ViT-base-patch16), ResNet50 and DenseNet121 models were also tested on our dataset for comparison.

The paper is organized as follows: methodology including the database, data preprocessing procedure, and proposed networks are illustated in section 2. Experiments are descibed in section 3, results are presented in section 4 and finally section 5 concludes the paper.

2. Methodology

A) Database

In this research, a dataset of 5 classes collected from people suspected of COVID-19 is used. Two hundred seventy-nine women and 410 men (689 cases in total) participated in collecting the data. These 3D-CTs were divided into 1 cm slices each. Between 20 and 30 2D images were provided in total from each person. The size of the 2D images is 512 × 512. Two professional radiologists divided the two-dimensional images obtained into five classes: normal with 314 images, initial with 80 images, progressive with 84 images, peak with 110 images, and absorption with 101 images. See Figure 1 for sample images.

Detection of the first stage of COVID-19 is done by identifying the ground-glass opacity, the second stage by identifying the crazy-paving pattern, the third stage by identifying consolidation, and the fourth stage by the absence of crazy-paving pattern and the gradual clarity of consolidation. Ground-glass opacity refers to an area in the lung tissue that has increased density and is similar to ground glass, foggy and cloudy (19). The crazy-paving pattern is called the pattern of stones arranged irregularly on the mortar. With the progression of the disease, when the septa between the lung lobes thicken and are placed on the ground-glass opacities, an appearance similar to crazy-paving appears (20). Consolidation is the term used to describe the compaction of part of the lung tissue so that a space that should be filled with air appears solid. It occurs when the inside of the lung alveoli is filled with pus, blood liquid, or other substances (21).

The CT images in this dataset were prepared at the Qaboos Teb Golestan imaging center located in the city of Gonbad Kavos (Iran). The device used to prepare these images was an American Hispeed CT dual scanner (GE Healthcare company). This device has imaged the patient's chest in spiral mode and holding breath at a speed of 15 millimeters/rotation and sagittal view. You can access the database through below link1 (22). During the imaging and examination of CT images of all individuals, the

ethical standards of Shahid Beheshti Faculty of Medical Sciences have been considered (22).

A fundamental principle for the success of classification tasks is the separability of classes. In Figure 2, two sample images from each class are presented. From the perspective of a less experienced individual, significant similarities may be observed between inter-class images, making it nearly impossible to distinguish between these classes. However, if features are extracted using a powerful model (e.g., ResNet50) and the inter-class variance and intra-class variance are calculated for the dataset, the inter-class variance is found to be significantly larger (approximately 48 times larger for ResNet50) than the intra-class variance. This indicates that the data is highly separable.

B) Data preprocessing

As previously explained, a substantial number of 512×512 two-dimensional images were acquired from each person as a result of chest CT imaging procedures. Among these obtained images, some (such as the first and last images) exhibit a smaller lung volume than others. Therefore, a threshold is needed to select more valuable images. To achieve this, the following tasks were performed. First, a 333×333 pixel region was extracted from the center of each image. Then, using the Otsu32 method, a digital image thresholding technique, the two-dimensional images were binarized. The number of pixels with a value of 1 (white color) now represents the lung volume, meaning that the greater the number of 1s, the larger the lung volume seen in the image. To find an image with these specifications, the normalized area was plotted against the frame number of CT scans, as shown in Figure 3.

Therefore, the authors identified the image with the greatest lung volume, selected the eight images preceding and following it, and separated these from the remaining images for subsequent use. This process yielded 17 images per individual. Subsequently, the images were qualitatively reviewed, and some were removed due to poor quality. It is important to note that the original 512×512 images (not 333×333) were used for the experiments.

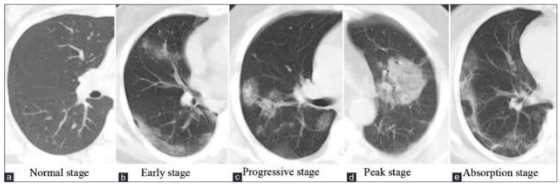

Fig. 1. Examples of images related to the CT scan of a normal person's lung and those affected by COVID-19 with different stages of involvement and tissue patterns. (a): normal, (b): early stage (0-4 days), (c): progressive stage (5-8 days), (d): peak stage (9-13 days), (e): absorption stage (≥14 days) (22).

Table.1.
The number of examined people from each stage was divided into three sets: training, validation, and testing sets. The last row shows the number of images after preprocessing.

Stage	***All persons***	***Training***	***Validation***	***Test***
Normal	314	251	31	32
Early	80	64	8	8
Progressive	84	67	8	9
Peak	110	87	11	12
Absorption	101	80	10	11
Total people	689	549	68	72
Total images	10902	8724	1098	1080

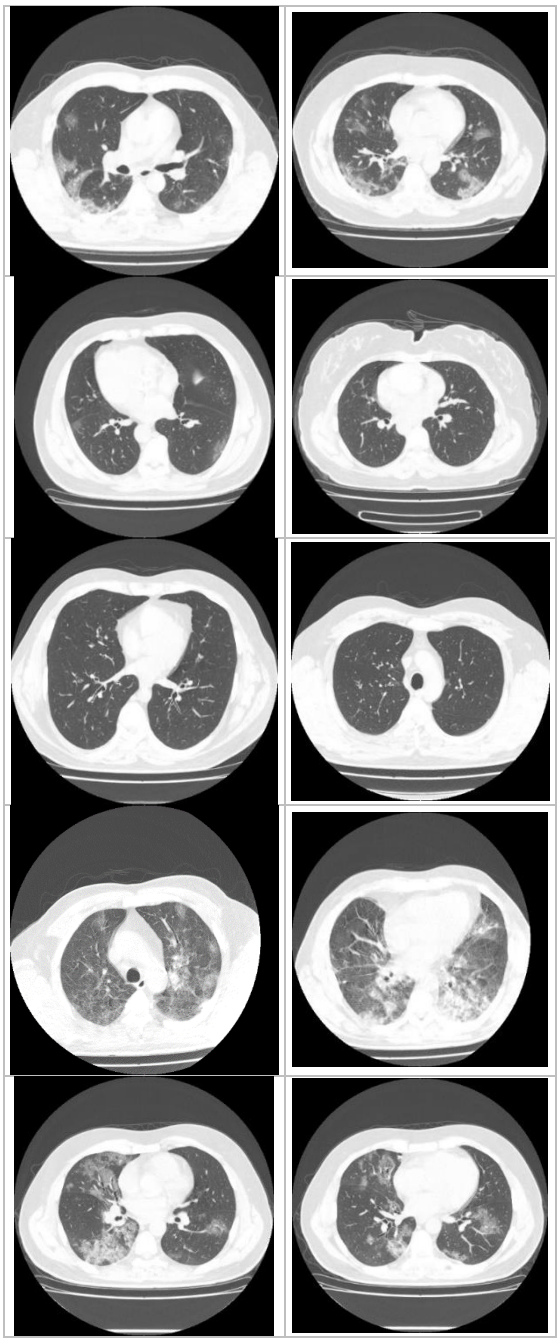

Fig. 2. Two data samples from each class are shown before preprocessing. The images in each row belong to the same class.

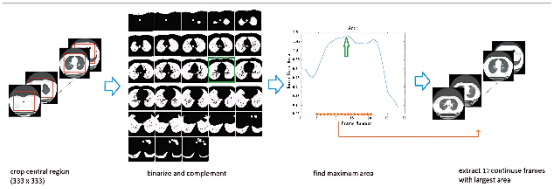

Fig. 3. Preprocessing and extracting the best lung CT images of each patient (reprinted from (22)).

C) *Proposed Network*

This study proposes a Swin transformer for the classification of COVID-19 CT images. This transformer is a vision transformer that limits self-attention computation to local windows, unlike the Vision Transformer (ViT), which computes self-attention across the entire image. The Swin transformer's creators posit that most relevant information is contained in nearby pixels, thus local windows represent a more efficient approach. ViT was the first transformer to enter the field of computer vision. It requires a large amount of training data and utilizes a Multi-head Self-Attention (MSA) module, which has high computational complexity. With the advent of the Swin transformer, these weaknesses were relatively addressed.

Swin is a multifunctional backbone transformer, that can be used in various vision tasks. As depicted in Figure 4, this transformer consists of four nearly identical stages, and at the end of each stage, an image with a different resolution is formed as an output. Moving towards the final stage, the resolution of these outputs, which are called feature maps, decreases but they represent more semantic information. At the start of this structure and after receiving the image, there is a patch partition module that is responsible for dividing the images into small, equal patches. The first stage consists of a linear embedding module and an even number of Swin transformer blocks. The subsequent three stages consist of a patch merging module and an even number of Swin transformer blocks, these stages have similar functionality. Each transformer block comprises two multi-window self-attention blocks named W-MSA (Window multi-head self-attention) and SW-MSA (Shifted window multi-head self-attention) that are responsible for computing self-attention within windows before and after shifting them. Each W-MSA block consists of a normalization layer (LN), a W-MSA module, and a multilayer perceptron (MLP). Also, each SW-MSA block has the same components as the W-MSA block, but it uses the SW-MSA module instead of the W-MSA module [23]. The following describes the most prominent features of this network.

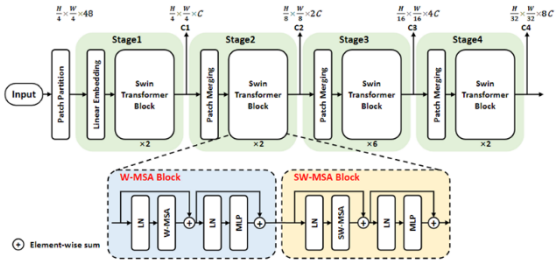

Fig. 4. The general framework of Swin transformer (Swin-Tiny) (23).

Reducing the load of calculations: In Vit, the computational complexity of calculating self-attention increases quadratically with the number of input image patches, which causes more resources to be used to calculate self-attention. Swin transformer uses non-overlapping windows to solve this problem. In the sense that it considers windows of equal size and no overlap for the images, in a way that an equal number of image patches are placed within each window. Then it calculates the self-attention within these non-overlapping windows, which makes the computational complexity have a linear relationship with the size of the image. Equations (1) and (2) calculate the computational complexity for the MSA and W-MSA modules for an image divided into h×w patches and C number of channels. In the second equation, M×M is the number of patches within each non-overlapping window related to W-MSA. As you can see, the first equation is quadratic for hw, and the second is linear for hw when M is constant [23].

$$\Omega(\mathrm{MSA}) = 4hwC^2 + 2(hw)^2C \quad (1)$$

$$\Omega(\mathrm{W-MSA}) = 4hwC^2 + 2M^2hwC \quad (2)$$

The shifted window approach: In this approach, after partitioning the image, equal-sized windows are initially considered as shown in the left figure, and self-attention is calculated in them by the W-MSA module. Then, the windows are shifted half their width to the bottom right, creating the right figure, and this time self-attention is calculated by the SW-MSA module in the new windows (shifted windows). This window transition is done so that patches that were close together but were not previously in the same window can also learn from each other. This approach is illustrated in Figure 5.

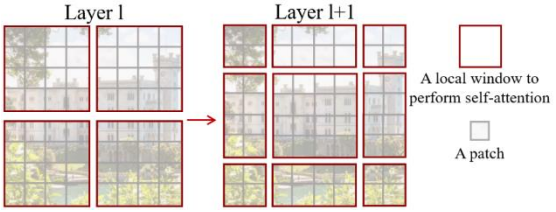

Fig. 5. Shifted window approach (12).

Relative position bias: Relative bias is a proposed method for embedding the position in transformers, which is preferred in the Swin structure over techniques such as absolute position embedding. This approach encodes the distance between patches and gives position information to keys and values in Attention instead of simply combining it by embedding patches (24). Equation (3) shows how to calculate attention with relative bias.

$$Attention(Q.K.V) = SoftMax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (3)$$

where $B \in R^{M^2 \times M^2}$ represents the relative position bias, $Q,K,V \in R^{M^2 \times d}$ are the query, key and value matrices; d is the query/key dimension, and M2 is the number of patches in a window [12]. This research has been implemented in Python using the PyTorch library. Also, the codes were run on Google Colab with a T4 GPU and 12.7 GB of RAM, and on Kaggle with a P100 GPU and 13 GB of RAM.

3. Experiments

Experiment 1: To simplify simulations, the authors chose the smallest Swin transformer model (Swin Tiny) for experiments. To compare the results and evaluate the strength of this model, the style models of three other popular networks were used. These models were ResNet50 and DenseNet121 convolutional networks, and ViT (ViT-base-patch16). ViT was pre-trained on ImageNet 21K images, and the other models were pre-trained on ImageNet 1K. To employ the data, the dataset was divided into three random separate parts with an approximate ratio of 80% (training), 10% (validation), and 10% (testing), as shown in Table 1. Also, before the training process , the MLP heads of the last layer of the models were removed and replaced with new MLP heads which have 5 neurons in the outer layer. Then, the Swin model was trained with a learning rate of 0.0001 and the AdamW optimizer, ResNet with a learning rate of 0.0001 and the Adam optimizer, DenseNet and ViT with a learning rate of 0.001 and the Stochastic gradient descent (SGD) optimizer, for 100 epochs on training data, to achieve stability. It is worth noting that these hyperparameters provide the best performance for the mentioned networks (25). Simultaneously with the training process and after each epoch, the authors performed a validation epoch and finally evaluated the model on the test data and displayed the results.

After each training epoch, the training data is shuffled, but the validation data is not shuffled. For data augmentation and normalization steps, a random resized crop with 224 pixels size and random horizontal flipping has been used on training data, and the center crop technique has been used on validation and testing data. Also, since the models examined in this experiment were already trained on the ImageNet dataset, the entire dataset has been normalized based on the average values and standard deviation of the ImageNet dataset. This procedure stabilizes the gradient. Finally, the images given as input to the models had a length and width of 224 pixels.

Experiment 2: The Swin Tiny model was trained with the same conditions as before for 300 epochs (Experiment 1 was done for 100 epochs). In this situation, the model becomes more stable and more accurate on the testing dataset.

4. Results and Discussion

A) Evaluation metrics

Accuracy, precision, recall, and F1 score are the most common metrics used to evaluate the performance of neural networks. Formulas for these metrics are provided below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP+FP+FN} \quad (7)$$

TP, TN, FP, and FN are acronyms for true positive, true negative, false positive, and false negative, respectively.

A Receiver Operating Characteristic curve, or ROC curve is a probability diagram that shows the effect of different sensitivity values on the specificity. The vertical axis of this diagram shows the true positive rate, and the horizontal axis shows the false positive rate. The Area under the ROC Curve (AUC) metric is used to compare and evaluate the performance of models. The higher the value (area), the better the overall performance of the model.

Macro average and micro average are two components of averaging methods. In the macro average method, first, the desired metric (e.g., precision) of each class is calculated, and then the average of all of them is taken. But in the micro average method, after calculating the all-classes desired metric, their weighted average is calculated.

The authors used macro-averaged precision, recall, and F1-score metrics, and macro and micro-averaged AUC to compare the performance of the networks. The use of macro averages is useful in cases where the number of data points in each class is not equal (26).

B) Quantitative evaluations

Experiment 1: You can see the curves of accuracy, and cross-entropy loss function obtained from training and validation of models for 100

epochs in Figure 6 and Figure 7. In the curves related to training, as we move from epoch 1 to epoch 100, the model's accuracy increases, and the amount of loss function decreases. This matter is seen with more distortion in the curves related to validation, which is normal. You can also see that in the last epoch, the accuracy gradient has decreased, and the models have reached relative stability, which means the models are close to the global minimum.

After fine-tuning the models, they were tested on the test data. The obtained results for classification (accuracy, F1 macro score, macro precision, and macro recall) are shown in Table 2. Since the recognition of each class has equal importance, macro metrics were used. As can be seen, Swin achieved the highest accuracy, which is 0.95. Due to its higher accuracy than ViT (0.9), its success in improving the classification accuracy of vision transformers can be realized. After the Swin Transformer, the ResNet50 and DenseNet121 convolutional networks ranked next with 0.93 accuracy. Despite the equal accuracy of these two networks, if other metrics are considered, the ResNet50 network performed slightly better than DenseNet121. This proximity of percentages indicates the almost equal ability of these two convolutional networks in classification tasks. After these three networks, the ViT is ranked last with an accuracy of 0.9, despite having more complexity and parameters than other models. Also, the recall of the Swin transformer is significantly greater than other models, which is important for medical issues like the present work. Another metric is the F1 score, which results from the combination of two metrics: precision and recall. In this metric, Swin likewise obtained the highest score (0.94), which shows the complete superiority of this transformer over the rest of the compared models.

Another metric considered to compare the networks' ability (on test data) was AUC. Since the dataset has five classes and is not binary, the authors averaged the AUC value of all classes with micro and macro methods. For this purpose, the ROC curve of each network was drawn (Figure 8). The highest value of micro AUC, 0.96, was obtained by Swin, ResNet, and DenseNet. The highest value of macro AUC, 0.95, was acquired by Swin, followed by the two convolutional networks, which achieved a value of 0.94. This small difference indicates the almost equal resolution ability of these three networks.

In general, these results indicate the significant effect of the transfer learning technique in addressing the data scarcity problem. Transfer learning is a technique whereby the network leverages experience gained from training on a large dataset to solve another related task with a smaller dataset. This technique is employed when the tasks of the network(s) are closely related (27). With this, the network will no longer have random weights at the beginning of the work, and instead, it will use weights obtained from training with previous data.

Experiment 2: Figure 9 and Figure 10 show the curves of the loss function and accuracy for 300 training and validation epochs. It is evident that the Swin transformer has reached complete stability without overfitting, and it is fair to say that its changes have become imperceptible after 200 epochs.

Table 3 shows the scores obtained by this network. The network has achieved about 0.97 accuracy in our test data, which is an outstanding result. The macro F1 score, macro precision, and macro recall of the model equal 0.96.

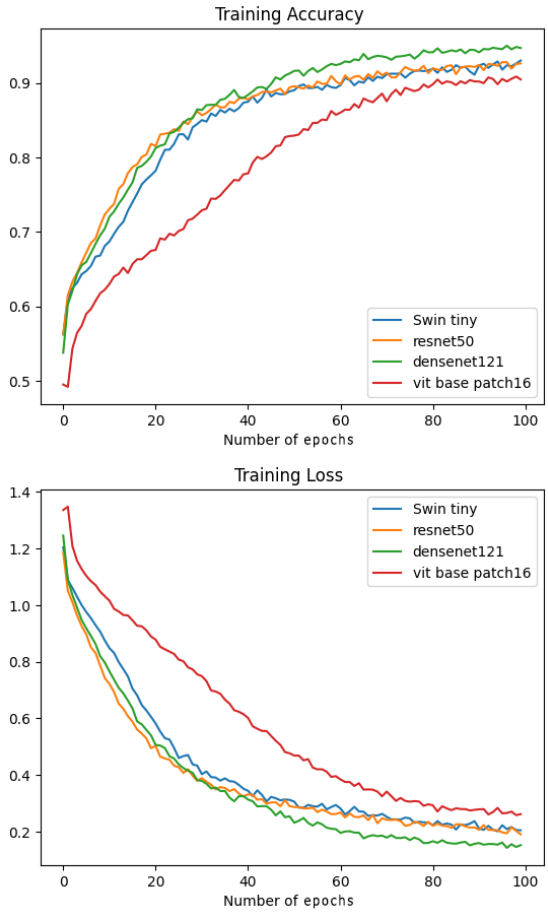

Fig. 6. Training accuracy and Training loss function (Cross-Entropy) in 100 training epochs.

Table.2.
Results of models on test data

Model	*Macro precision*	*Macro recall*	*Macro F1-Score*	*Accuracy*
Swin Transformer	0.95	0.93	0.94	0.95
Resnet50	0.93	0.90	0.92	0.93
Densenet121	0.92	0.90	0.91	0.93
ViT	0.88	0.87	0.87	0.90

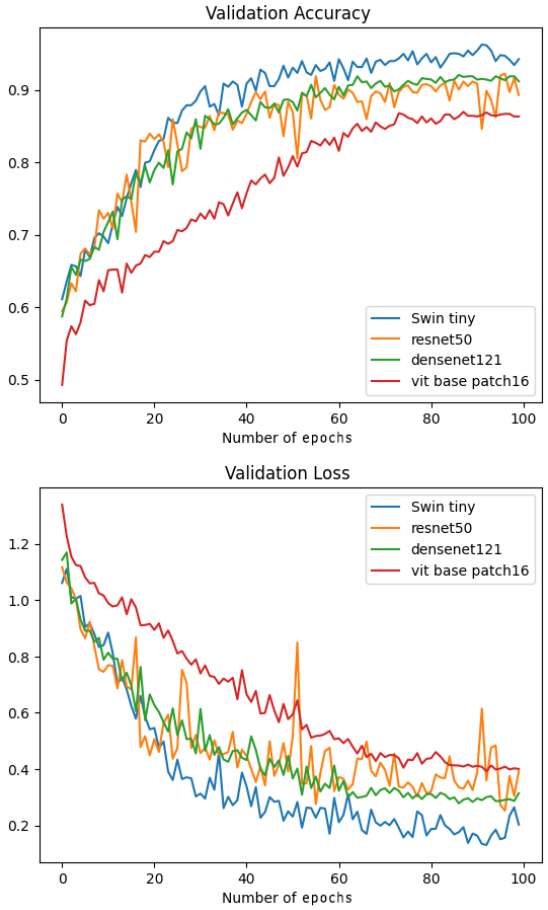

Fig. 7. Validation accuracy and Validation loss function (Cross-Entropy) in 100 validation epochs

Figure 11 shows the network confusion matrix. The abbreviations in this matrix are as follows: AB (absorption stage), E (early stage), NL (normal stage), PE (peak stage), and PR (progressive stage). The diagonal elements of this matrix represent the number of images that are correctly classified, and the other numbers represent incorrect predictions where one class label is mistaken for another class label. Out of 512 lung images of normal people (NL), only five cases were incorrectly predicted as other classes. Additionally, only one out of 163 peak-stage images (PE) was incorrectly predicted as a normal lung, which is an acceptable result. Note that the higher accuracy of the network in distinguishing the class of normal people from the rest of the classes is due to the clear difference between the images of normal lungs and those of lungs affected by COVID-19. The authors also illustrated the ROC curve of this model for each of the five classes of the test dataset, separately using the Scikit-learn Python library and in the thresholds selected by the library (0, 1, and 2) (Figure 12). The ideal state of a network is a state whose curve is close to the upper-left corner of the diagram because, in this state, the recall is higher, the FPR is lower, and the AUC is also larger. After checking the curve of all five classes, it was found that the NL (Normal) class has the highest AUC with 0.99.

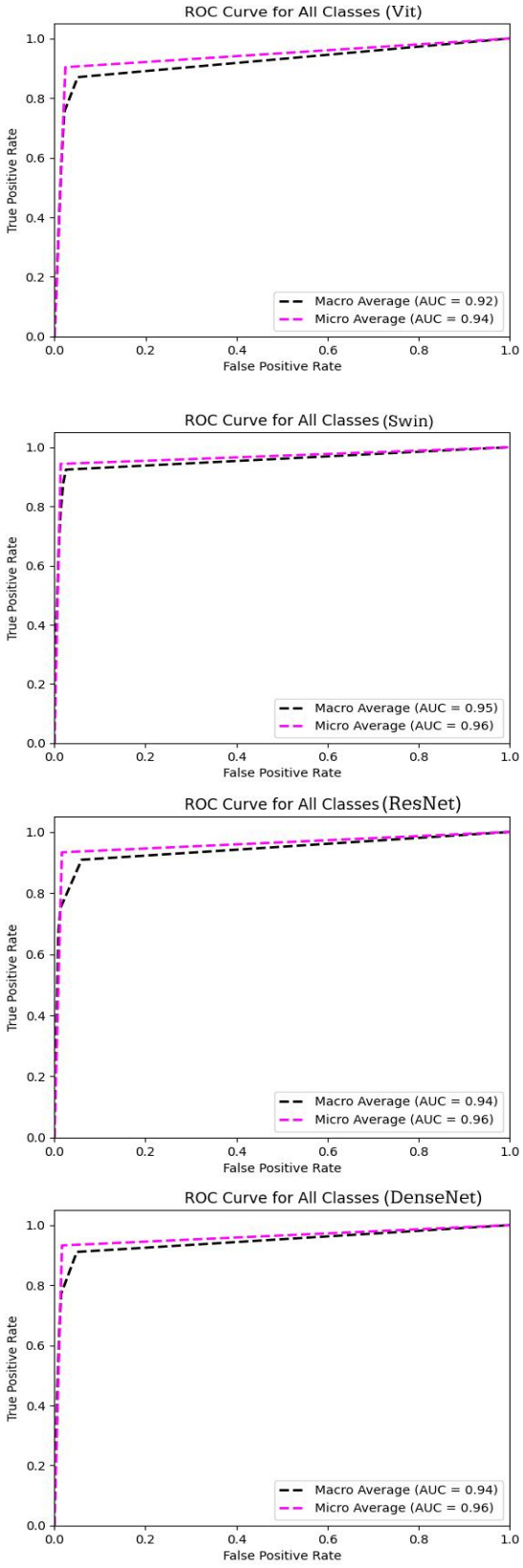

Fig. 8. Macro & Micro average ROC curve for four models: Vit-base-patch16, Swin Tiny, ResNet50 and Densenet121.

After that, class E (Early stage), PE (Peak stage), and PR (Progressive stage) ranked second with a value of 0.98, and class AB (Absorption stage) ranked last with the lowest AUC (0.96). Therefore, all of them have a high AUC, which indicates the excellent performance of the Swin model in distinguishing all classes from each other. As a result, this model can help radiologists and physicians in diagnosing normal lungs and determining the stage of lung involvement, as well as allocating resources for patient hospitalization and such decisions.

The Table 4 presents a comparison of several studies on COVID-19 stage classification based on lung involvement severity with our study. The results obtained from this research are significantly better compared to the paper (22) that worked on the same dataset. The authors believe that this is due to using more 2D images for training (more than five frames per person), using a different preprocessing technique, and the power of the Swin transformer. A key advantage of our approach over studies such as Paper (28), Paper (29), and similar research is that it eliminates the need for lung and lesion segmentation. Furthermore, Paper (29) has an additional limitation, as it requires access to the patient's previous CT scans and clinical metadata for classification. Despite these advantages, a limitation of our study is that the dataset was collected from a single source, which may affect the model's generalizability to external data.

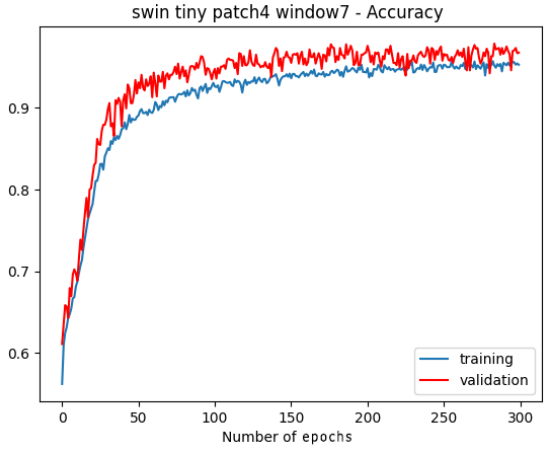

Fig. 9. The training and validation accuracy of the Swin transformer in 300 epochs.

Table.3.

Results of the Swin model on test data after 300 epochs of training

Model	*Accuracy*	*Macro precision*	*Macro recall*	*Macro F1 Score*
Swin transformer	0.97	0.96	0.96	0.96

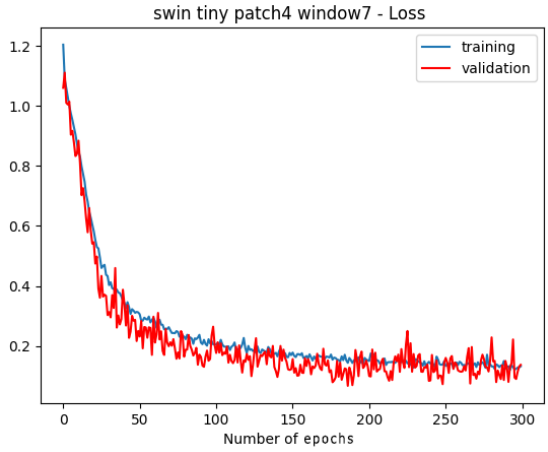

Fig. 10. The training and validation loss function of the Swin transformer in 300 epochs.

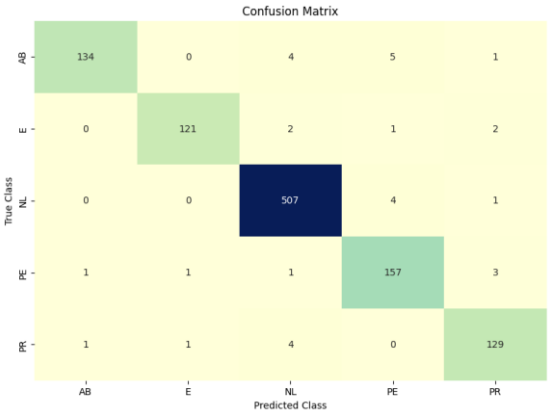

Fig. 11. Confusion matrix of the Swin transformer on the test data after 300 training epochs (AB: absorption stage, E: early stage, NL: normal, PE: peak stage, and PR: progressive stage).

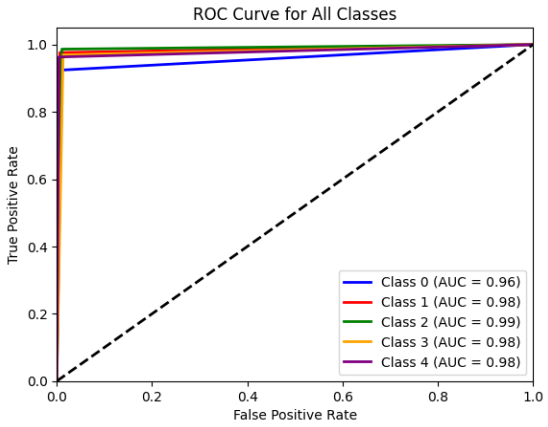

Fig. 12. The ROC curve of the Swin tiny model is for the test data (class 0: absorption stage (AB), class 1: early stage (E), class 2: normal (NL), class 3: peak stage (PE), and class 4: progressive stage (PR)).

Table.4.
Comparison of Our Work with Others in COVID-19 Classification Based on CT Images.

Author	*Dataset*	*Methodology*	*Results*
Zhidan Li et al., 2021 (29)	A combination of multiple public and private datasets.	Lung and lesion segmentation, followed by classification through comparison with the patient's previous images (dual-Siamese channels) and the use of clinical metadata.	Accuracy: 86.7
Qiblawey Y et al., 2021 (28)	Four public datasets	Classification based on Percentage of Infection (PI) using lung and lesion segmentation. (4 classes)	Accuracy: 97
Gifani P et al., 2023 (22)	Private Dataset	Ensemble of transfer learning models with convolutional neural networks	Accuracy: 91.94
Our work	Private Dataset	Swin Transformer with Transfer Learning	Accuracy: 97

5. Conclusion and future work

The authors selected three well-known networks to compare with the Swin transformer, then trained, validated, and tested all four networks on a 5-class dataset of suspected COVID-19 patients. They observed that the Swin transformer outperformed the other models in all scores. The authors then trained the Swin model for an additional 300 epochs on their training data to maximize the network's accuracy in the validation data. Subsequently, they classified the test data using this network, resulting in improved results compared to before, with the model achieving 0.97 accuracy in test data and 0.967 accuracy in validation data. These results demonstrated the strong performance of the model in data classification, which was attributed to the efficiency of the local attention of this network and the transfer learning technique used in the networks.

One limitation of the work was the small amount of data and the absence of a class with lung images of non-COVID pneumonia, which would have allowed the model to detect non-COVID pneumonia in addition to identifying the stage of COVID-19 involvement and distinguishing a COVID-19 patient from a healthy person. Therefore, the authors intend to address this in future work. Additionally, it is advisable to assess the generalizability of this model within the same dataset. Furthermore, as the Swin transformer is a multifunctional network, it may be beneficial to first segment the images using Swin and then classify them, followed by comparing the results with the current findings.

Acknowledgment

The authors would like to thank Prof Christian Jutten, for his helpful comments on the paper.

References

[1] Ciotti M, Ciccozzi M, Terrinoni A, Jiang W-C, Wang C-B, Bernardini S. The COVID-19 pandemic. Critical reviews in clinical laboratory sciences. 2020;57(6):365-88.

[2] Alimohamadi Y, Sepandi M, Taghdir M, Hosamirudsari H. Determine the most common clinical symptoms in COVID-19 patients: a systematic review and meta-analysis. Journal of preventive medicine and hygiene. 2020;61(3):E304.

[3] Elibol E. Otolaryngological symptoms in COVID-19. European Archives of Oto-Rhino-Laryngology. 2021;278(4):1233-6.

[4] Czubak J, Stolarczyk K, Orzeł A, Frączek M, Zatoński T. Comparison of the clinical differences between COVID-19, SARS, influenza, and the common cold: A systematic literature review. Advances in Clinical and Experimental Medicine. 2021;30(1):109-14.

[5] Mortazavi H, Beni HM, Aghaei F, Sajadian SH. SARS-CoV-2 droplet deposition path and its effects on the human upper airway in the oral inhalation. Computer Methods and Programs in Biomedicine. 2021;200:105843.

[6] Birman D. Investigation of the Effects of Covid-19 on Different Organs of the Body. Eurasian Journal of Chemical, Medicinal and Petroleum Research. 2023;2(1):24-36.

[7] Kollias D, Arsenos A, Soukissian L, Kollias S, editors. Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021.

[8] Khehrah N, Farid MS, Bilal S, Khan MH. Lung nodule detection in CT images using statistical and shape-based features. Journal of Imaging. 2020;6(2):6.

[9] LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436-44.

[10] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

[11] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020.

[12] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al., editors. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021.

[13] Dümen S, Kavalcı Yılmaz E, Adem K, Avaroglu E. Performance of vision transformer and swin transformer models for lemon quality classification in fruit juice factories. European Food Research and Technology. 2024:1-12.

[14] Zhang L, Wen Y, editors. A transformer-based framework for automatic COVID19 diagnosis in chest CTs. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021.

[15] Jiang J, Lin S. Covid-19 detection in chest x-ray images using swin-transformer and transformer in transformer. arXiv preprint arXiv:211008427. 2021.

[16] Chen G-L, Hsu C-C, Wu M-H, editors. Adaptive distribution learning with statistical hypothesis testing for COVID-19 CT scan classification. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021.

[17] Jamali A, Mahdianpari M. Swin Transformer and Deep Convolutional Neural Networks for Coastal Wetland Classification Using Sentinel-1, Sentinel-2, and LiDAR Data. Remote Sensing. 2022;14(2):359.

[18] Tummala S, Kim J, Kadry S. BreaST-Net: Multi-Class Classification of Breast Cancer from Histopathological Images Using Ensemble of Swin Transformers. Mathematics. 2022;10(21):4109.

[19] Remy-Jardin M, Remy J, Giraud F, Wattinne L, Gosselin B. Computed tomography assessment of ground-glass opacity: semiology and significance. Journal of thoracic imaging. 1993;8(4):249-64.

[20] Kunal S, Gera K, Pilaniya V, Jain S, Gothi R, Shah A. "Crazy-paving" pattern: a characteristic presentation of pulmonary alveolar proteinosis and a review of the literature from India. Lung India: Official Organ of Indian Chest Society. 2016;33(3):335.

[21] Lee KS, Han J, Chung MP, Jeong YJ, Lee KS, Han J, et al. Consolidation. Radiology Illustrated: Chest Radiology. 2014:33-47.

[22] Gifani P, Vafaeezadeh M, Ghorbani M, Mehri-Kakavand G, Pursamimi M, Shalbaf A, et al. Automatic diagnosis of stage of COVID-19 patients using an ensemble of transfer learning with convolutional neural networks based on computed tomography images. Journal of Medical Signals & Sensors. 2023;13(2):101-9.

[23] Xu Z, Zhang W, Zhang T, Yang Z, Li J. Efficient transformer for remote sensing image segmentation. Remote Sensing. 2021;13(18):3585.

[24] Shaw P, Uszkoreit J, Vaswani A. Self-attention with relative position representations. arXiv preprint arXiv:180302155. 2018.

[25] Peng L, Wang C, Tian G, Liu G, Li G, Lu Y, et al. Analysis of CT scan images for COVID-19 pneumonia based on a deep ensemble framework with DenseNet, Swin transformer, and RegNet. Frontiers in microbiology. 2022;13:995323.

[26] Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv preprint arXiv:200805756. 2020.

[27] Shaha M, Pawar M, editors. Transfer learning for image classification. 2018 second international conference on electronics, communication and aerospace technology (ICECA); 2018: IEEE.

[28] Qiblawey Y, Tahir A, Chowdhury ME, Khandakar A, Kiranyaz S, Rahman T, et al. Detection and severity classification of COVID-19 in CT images using deep learning. Diagnostics. 2021;11(5):893.

[29] Li Z, Zhao S, Chen Y, Luo F, Kang Z, Cai S, et al. A deep-learning-based framework for severity assessment of COVID-19 with CT images. Expert Systems with Applications. 2021;185:115616.