

# An Augmentation-Based System for Diagnosing COVID-19 Using Deep Learning

Mohamad Shady Alrahal<sup>1</sup>, Mohammad A. Mezher<sup>2</sup>,  
Osamah A.M. Ghaleb<sup>3</sup>, Mohammad Al-Hjouj<sup>4</sup>, Raghad Sehly<sup>5</sup>, Samir Bataineh<sup>6</sup>  
AIMD Laboratory, Computing College, Fahad Bin Sultan University, Tabuk City, KSA<sup>1, 2, 3, 4, 5, 6</sup>  
Institute of International Education, SRF program, New York City, United States of America<sup>1</sup>

**Abstract**—Recently, due to the dangerous spread of COVID-19, there has been strong competition among computer science researchers within the scientific research community to employ deep learning for the development of intelligent medical systems that diagnose this illness. Enhancing accuracy is considered the most important objective, and augmentation techniques are used in this context. This study addresses two main issues related to applying augmentation on X-ray and CT-scan images: losing the positional information of augmented medical images and the integration of extracted features while scanning them. The use of the Vision Transformer Structure, supported by a Position-Aware Embedding (PAE) method, is proposed to deal with these issues. Moreover, in this study, a student–teacher-based approach was adopted to enable considerable resistance against training on a small batch of training images. Due to the sensitivity of medical data, preserving the privacy of patients was taken into account by using a pseudonym-based anonymity approach. After evaluations based on accuracy, precision, recall, and specificity metrics, the results showed that the proposed system has a high-level capability to predict class images (X-ray or CT-scan) as well as considerable resistance against training on small medical images.

**Keywords**—COVID-19; medical images; augmentation; vision transformer; training data ratio

## I. INTRODUCTION

The medical sector is considered the most important sector in people's lives because staying healthy leads to higher work productivity and increased happiness. Therefore, governments always rank fighting epidemics at the top of their priorities. In 2019, strange medical signals were observed in some patients in Wuhan, China. These signals showed that the patients had been infected with a serious new strain of coronavirus disease (known as COVID-19) [1, 2]. Subsequently, various parts of the world experienced COVID-19 outbreaks, prompting the World Health Organization (WHO) to inform the public that the COVID-19 crisis had become a global pandemic after the disease was found to have exceptional spreadability. Governments all over the world colluded to fight against the dangerous spread of COVID-19 and its manifold consequences that came in different forms (losses of human life, economic recessions, faltering medical health systems, and necessary lifestyle changes). Most scientists, physicians, and politicians have agreed that the best way to stop the spread of COVID-19 is early diagnosis and detection and adherence to preventive measures such as social distancing [3]. Fig. 1 summarises and arranges the historical events outlined above.

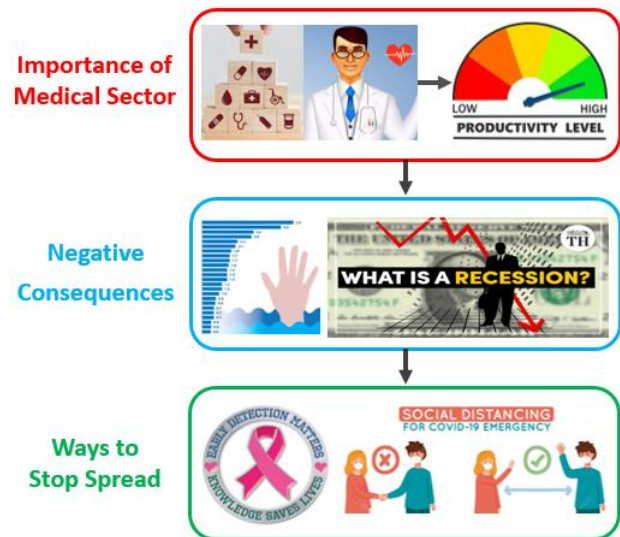


Fig. 1. Historical events resulting from the outbreak of COVID-19.

## A. Motivation

The traditional ways of detecting COVID-19 are antibody testing and Quantitative Reverse Transcription Polymerase Chain Reaction (QRT-PCR), both of which have some issues. Antibody testing generates high false-negative rates when it comes to dealing with early active infections [4, 5]. With QRT-PCR, generating results is time-consuming; thus, it is unsuitable when working under severe time constraints. Relying on medical image analysis and making diagnoses via the use of CT images has proven effective in the early detection of COVID-19. However, techniques based on CT images are time-consuming, as they require some pre-activities to be carried out (e.g., transferring patients to a CT facility and the sterilization of medical devices) and other post-activities (e.g., discussing with consultants to analyse the results of the test). Using X-ray images is preferable to the use of CT scan images in terms of cost. However, some research has shown that when COVID-19 progresses in its infection of patients' bodies, some visual features of X-ray images become blurred (i.e., less informative) [6]. As a result, there is a strong motivation among researchers to develop advanced diagnostic systems to detect COVID-19, and here, deep learning (DL) techniques come to the fore. Fig. 2 summarizes this motivation.

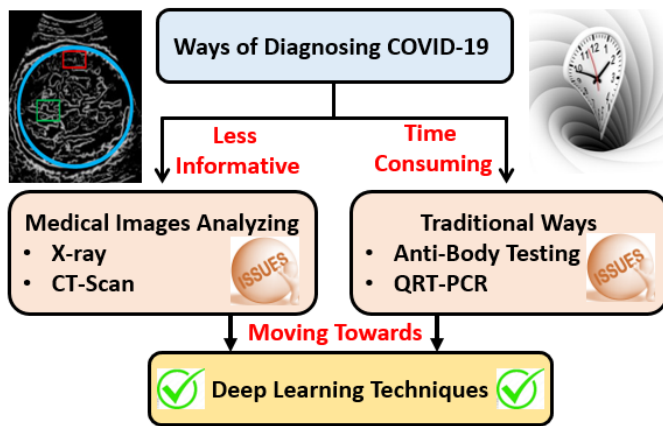


Fig. 2. Motivation.

### B. Statement of the Problem

In order to fight against COVID-19, researchers in the field of computer science have made invaluable contributions to the development of intelligent diagnosing systems by enhancing the accuracy of the intelligent systems developed for detecting COVID-19. They have employed the principles of Medical Image Analysis (MIA) and deep learning (DL) to train intelligent systems. X-rays and CT scans are considered the most common sources of medical images. Among the intelligent systems used for diagnosing COVID-19, their low accuracy is the root cause of why many of these systems fail, as insufficient accuracy leads to an increase in the rate of deaths among infected people because of the high number of cases that are falsely classified as negative [7]. Fig. 3 illustrates the low accuracy problem.

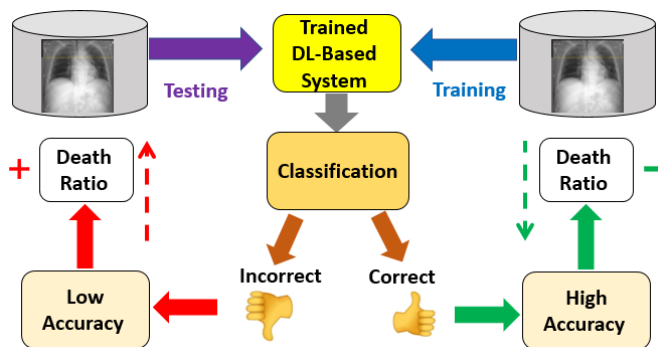


Fig. 3. The low accuracy problem leads to a high death ratio.

One possible way to enhance accuracy is to carry out effective data pre-processing before the system's training stage. In this context, augmentation-based techniques, which center around rescaling medical images to enable the extraction of learning features in an efficient way, are widely used [8]. However, the use of augmentation-based techniques can lead to some issues, including the following: 1) the loss of positional information because the original input medical image is divided into segments, and incorrectly re-ordering segments leads to confusion in the learning process [9]; 2) the integration of extracted features across the entire medical image is crucial to ensure learning on all features [10]; and 3) resistance against different sizes of training datasets becomes a barrier for ensuring

that the diagnosing system can be applied in real-time [11]. Fig. 4 illustrates the issues that can arise when using augmentation-based techniques.

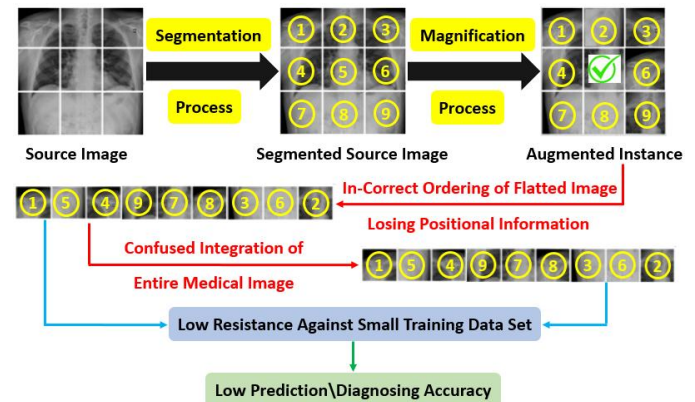


Fig. 4. Issues resulting from poor augmentation.

### C. Research Questions

This study aims to address the following research questions:

- 1) How can one avoid losing the positional information of augmented medical images that are used as inputs for the training process?
- 2) How can one ensure the integration of extracted features while scanning whole medical images?
- 3) How can one enable resistant real-time diagnosing against small training datasets?
- 4) How can the preservation of the privacy of patients, which is a mandatory requirement in the medical sector, be ensured?

### D. Contributions

In addressing the research questions listed above, this work aims to make the following contributions to the literature:

- Regarding the first research question, a Position-Aware Embedding (PAE) method that takes into account the corresponding positional information during the flattening process is presented.
- Regarding the second research question, an Integration-based Vision Transformer Structure (IbVTS) is proposed. The self-attention mechanism provided by this VTS, which ensures the integration of the information extracted from each segment of the flattened medical images, is explained.
- Regarding the third research question, the IbVTS is supported by a distillation tactic to activate less training data for the purpose of supporting the classification layer.
- Regarding the fourth research question, a pseudonym-based anonymity approach (PAA) is proposed to hide the personal information of patients.

### E. Structure of Study

The rest of this study is structured as follows: Section II outlines related works. In Section III, the architecture of the

proposed intelligent diagnosing system is presented in detail, along with the role of each component. The experimental results are documented in Section IV, and finally, Section V concludes the study and lists directions for future research.

## II. RELATED WORK

Many survey-based papers have addressed intelligent systems used to diagnose COVID-19 and provided various taxonomies, such as [12, 13, 14]. The authors of [12] relied on the application of different deep learning approaches used to diagnose COVID-19 (convolutional neural network, Generative Adversarial Network, and Long Short-Term Memory) to derive their taxonomy. The author of [13] presented a taxonomy based on methods used for the data pre-processing stage and techniques used for feature extraction. Convolutional neural networks, recurrent neural networks, deep belief networks, and reinforcement learning are the deep learning architectures used to shape the base of the taxonomy presented in [14]. This work groups intelligent systems into five categories based on the type of learning, as shown in Fig. 5.

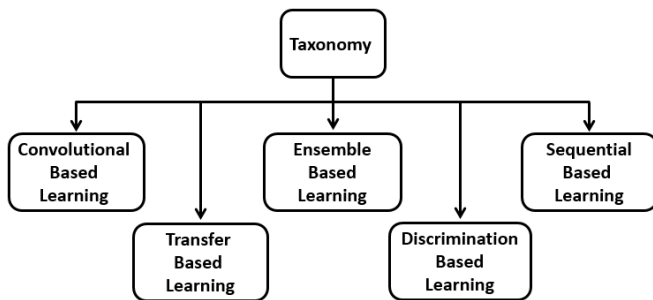


Fig. 5. Taxonomy of intelligent systems used to diagnose COVID-19.

### A. Convolutional-Based Learning

In this category, convolutional neural networks dominate the architectures of the proposed systems. CNNs consist of two main types of layers: a convolutional layer and a pooling layer. CNNs employ filtering principles used in the domain of image processing to extract features.

In [15], the researchers proposed an automatic system for diagnosing COVID-19 using a pre-trained U-net architecture. The purpose of this U-net architecture is to segment medical images of the lungs. Next, the segments are used as inputs for a deep neural network to output the class of the given image. They used a dataset of 540 patients, with 313 (138 males; 175 females) of these patients being COVID-19 positive and 229 (88 males; 141 females) being COVID-19-negative. The experiments were conducted on the CT-scan dataset and produced 90.1 % accuracy on binary classes. The CNN-based architecture presented in [16] consists of 25 layers (19 layers for convolution and 6 max-pooling layers). Convolutional layers are connected to batch normalization for input standardization and to regulate the intelligent model. To ensure the continuity of neurons, a leaky rectified linear unit is utilized. The maxpool function is used for the down-sampling of the inputs. The authors used an X-ray dataset comprising 600 medical images for binary classification. An accuracy of 87 % was obtained.

### B. Transfer-Based Learning

In this category, two main domains are involved: the source domain and the target domain. The source domain is characterized by models that are data-rich and have high-quality feature extraction capabilities. In contrast, the target domain suffers from low-quality feature extraction capabilities and requires help from pre-trained models. The purpose of transfer-based learning is to shift the knowledge from the source domain to the target domain. The degree of maturity of the transferred knowledge depends on the degree of similarity between the source and target.

Transfer-based learning was applied to CT scan medical images in [17] to avoid the high costs and computational complexity associated with constructing a CNN model. Two main DL-based algorithms were used (ResNet18 and ResNet50) as pre-trained models. Then, discriminant correlation was employed for feature fusion to generate a better image representation. The data of the training dataset were gathered from a local medical center, with there being 420 samples for each class (normal and infected). The results showed that an accuracy of 96.35 % was achieved. The authors of [18] chose to use VGG16, VGG19, ResNet, DenseNet, and InceptionV3 as pre-trained DL-based models with rich knowledge. These architectures were used to transfer information derived from chest X-ray images. To enhance the classification accuracy, a rotation-based augmentation technique was employed and referred to as a heading model. The results obtained after conducting experiments showed that among the five models used, VGG19 achieved the best accuracy (80%).

### C. Ensemble-Based Learning

The key idea of ensemble learning is to combine multiple models (referred to as learners) to generate an enhanced model. There is a wide spectrum of tactics used to perform a combination, such as majority voting, plurality voting, and weighted voting.

The authors of [19] proposed an intelligent model that combines EfficientNet and SE-ResNext to predict one of the three classes that label input images (COVID-19, pneumonia, and normal). They chose to use X-ray medical images and a weighted voting-based combination tactic. The results showed that the proposed model achieved a higher accuracy (about 95%) than that achieved by each combined model separately. In the context of ensemble learning, the authors of [20] utilized a stack of pre-trained deep learning models. They used VGG-16 as base learners, trained with a diverse set of inputs, followed by a logistic regression model, the meta learner, to combine the base learner predictions. This combination tactic relies on a fusion technique to enable the system to predict the class of the medical images. They achieved a high level of accuracy (about 89%).

### D. Discrimination-Based Learning

The key idea behind discrimination-based learning is to, as a first step, generate tuples that are similar in features and class to real training data. Then, the resultant tuples and real ones are mixed and act as inputs for the discriminator to be handled under a distinguishing process (i.e., to distinguish the real tuples from generated ones). In the context of discrimination-based learning, Generative Adversarial Networks (GANs) rules supreme.

A GAN-based system was proposed by the authors of [21] to construct synthetic CT scan images of both COVID-19 patients and healthy controls. The construction process was performed during the data augmentation stage. The Whale Optimization Algorithm (WOA) was used to optimize the hyperparameters of the GAN. The model was trained using the SARS-CoV-2 CT-Scan dataset, which consists of 2482 images (1252 COVID-19 cases and 1230 healthy cases). In terms of accuracy, the model obtained an accuracy value of 99.22%. A U-Net-based GAN designed for lung segmentation using chest X-ray images was proposed in [22]. This strategy involved employing a fully convolutional neural network in conjunction with the GAN model to segment images of the lungs in order to make COVID-19 diagnoses.

#### E. Sequential-Based Learning

In this category, the Long Short-Term Memory (LSTM) dominates because it was originally developed to enhance neural networks by enabling learning from sequential data.

The authors of [23] presented a hybrid intelligent model for diagnosing COVID-19 using an LSTM with a CNN. The development of the model consisted of three steps. First, medical images were enhanced in the pre-processing stage based on the contrast technique. Then, the enhanced images were used as model inputs for the purpose of learning. Finally, the Softmax function was used in the classification layer to distinguish between three classes: COVID-19, normal, and pneumonia. This model was trained on the COVID-19 Radiography dataset, which consists of 1143 images of COVID-19 cases, 1341 images of normal cases, and 1345 images of Pneumonia cases. The authors achieved an accuracy of 98.97 %. The authors of [24] utilized a pre-defined histogram threshold to segment CT scan-derived medical images. For feature extraction, both the Q-Deformed entropy (QDE) and a CNN were used. Finally, the extracted features were fused and used as inputs for the LSTM classifier to predict the classes of the images. The highest classification accuracy value obtained was 99.68%.

### III. PROPOSED SYSTEM

This section presents the proposed intelligent system. Firstly, the framework of the system is presented, followed by the system's architecture and the role of each component involved in constructing the architecture.

#### A. Framework of Proposed System

The environment within which the proposed system operates is medical health care centers/hospitals. For a given number of patients (Pat), medical machines capture chest images derived from two types of imaging techniques: X-ray, denoted as ( $MI_{xr}$ ), and CT scan, denoted as ( $MI_{ct}$ ). Both types form a real dataset [see Eq. (1)]:

$$RDS = \bigcup_{i=1}^{Pat} \{MI_{xr}^i, MI_{ct}^i\} \quad (1)$$

Each medical image is modeled as follows [see Eq. (2)]:

$$MI_{type \in \{xr, ct\}}^i = \langle IDp, NAMEp, BODp, SEXp, Dis \rangle | i = 1, 2, \dots, Pat \quad (2)$$

where, IDp, NAMEp, AGEp, SEXp, Dis represent the personal information of the patient, specifically their identifier, name, date of birth, sex, and medical description or health status.

In reality, in the usage stage of the proposed system (i.e., after training and testing),  $MI_{type \in \{xr, ct\}}^i$  is used as an input to predict one of three classes (healthy, pneumonia, or COVID-19), as shown in Fig. 6.

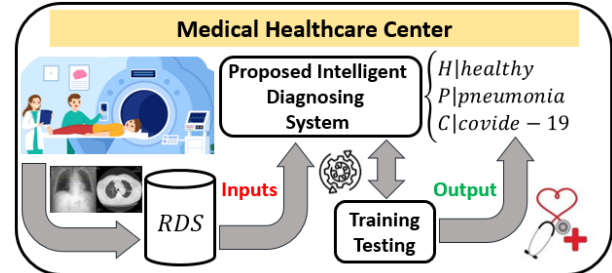


Fig. 6. Framework of the proposed system.

#### B. System Architecture

The objective of the proposed system is to, via learn mapping from inputs, predict/classify medical images to the correct class label. The proposed system is managed by six components. From a DL-based perspective, the components are grouped according to three phases, as shown in Fig. 7.

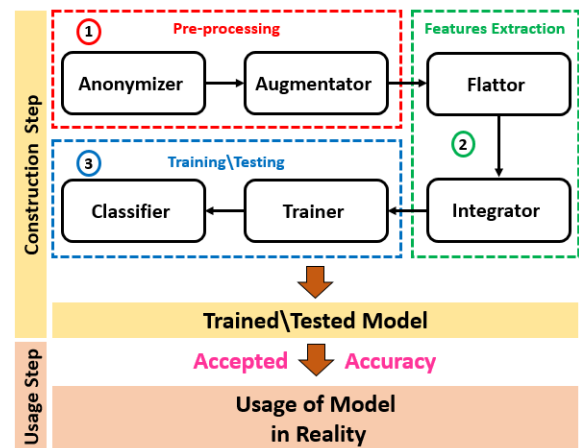


Fig. 7. Architecture of the proposed system.

Table I summarizes the main task of each component, the technique used for each task, and the quality attribute each task maintains.

TABLE I. COMPONENTS

Name of component	Main task	Used technique	Quality attribute
Anonymizer	Privacy preservation	Pseudonym	Privacy
Augmentator	Augmentation	Cutmix/Mixup	Contrast
Patch Embedder	Solving positional information losses	PAE	Resistance
Integrator	Effective feature extraction	IbVTS	Performance
Trainer	Training on integrated extracted features	Distillation	Accuracy
Classifier	Prediction of output	Softmax	



### C. Role of the Anonymizer Component

This component is responsible for protecting the privacy of patients by executing the pseudonym-based anonymity approach (PAA). For this, a pseudonym technique is employed. This technique is widely used to protect location privacy in location-based services. The key idea of the pseudonym technique is to handle personal data so that no data can be attributed to a specific person or data subject without additional information [25]. In this study, PAA was performed through three main steps: removing, replacing, and generalising. Fig. 8 illustrates these steps with an example.

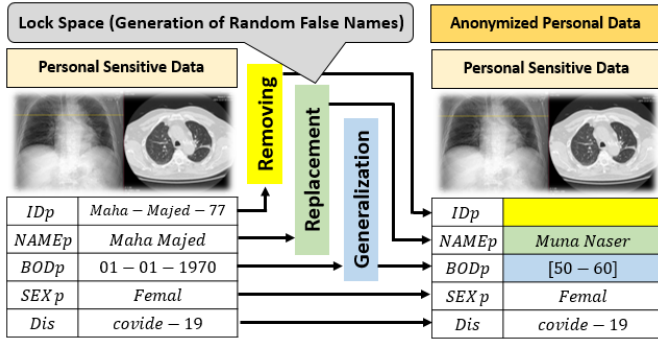


Fig. 8. Application of the three steps of the PAA.

As shown in Fig. 8, Maha Majed is the name of a real patient with a clear ID and clear date of birth, which, in turn, compromises this patient's privacy. Through using the PAA, their ID is removed. Their real name is replaced by "Muna Naser", which is a false name selected from the lock space that generates many pseudonyms based on some rules that help to create a false name adapted from the patient's real name. In our example, the real name consists of two parts (first name and last name), and so does the false one. In addition, the length of each part is 4 and 5 letters, respectively, which also matches the length of the first and last names of the pseudonym (false name). It is worth mentioning that privacy protection is a separate research field, and we will not discuss privacy protection in detail in this work, as we are only concerned about DL. The date of birth of the real patient is generalized in the anonymized profile so that no specific value is given. As for the sex and description fields, they are kept untouched. This way, the privacy of the patient is ensured, since the attacker cannot construct a malicious profile of a patient, and they also cannot make any inferences or link any information to a specific patient.

After performing the PAA task using the Anonymizer component, the resultant medical image ( $MI_{type \in (xr, ct)}^I$ ) is privacy protected [see Eq. (3)].

$$(\overline{MI_{type \in (xr, ct)}^I}) = \langle \overline{NAMEp}, \overline{BODp}, \overline{SEXp}, \overline{Dis} \rangle | i = 1, 2, \dots, Pat \quad (3)$$

### D. Role of the Augmentator Component

This component is responsible for the augmentation of medical images. The objective of augmentation is to enhance the quality of training by enabling the extraction of more meaningful information. There is a wide spectrum of techniques

used for augmentation [26]. The simplest techniques rely on some geometric transformation, such as rotating and cropping. However, such geometric transformation-based techniques are costly and complex, and they also require some manual operations to determine the angles and measures that suit the situation. Due to these limitations, researchers have moved towards more advanced techniques. Random erasing-based techniques such as the Cutout technique are used in this context [27]. The key idea behind the Cutout technique is to randomly remove/mask a square region of pixels in order to force the model to focus on a specific region and thus extract more robust features. For example, if the removed pixels form the background of a given image, then this noise will disappear. However, a critical weakness of the Cutout technique is related to its randomizing process. This means that sometimes it removes some informative parts of the given image, and consequently, some important features will be completely missing from the extracted space. The Mixup [28] technique saves all parts of a given image. The key idea behind this technique is to mix a pair of input images with their classes during training. Thus, more generalizable features that cover all input images are achieved. However, it may generate blurry images. An effective solution to the issue of blurring is discrimination, which is achieved by the use of the Cutmix technique [29]. Instead of masking the region, as is the case in the Cutout technique, it is replaced by the corresponding region from a different medical image. In this study, since we dealt with both X-ray- and CT scan-derived medical images, the Cutmix technique was used for the augmentation of the X-ray medical images, while the Mixup technique was used for the augmentation of the CT scan images. The reason behind this is that the ribs may blur X-ray scans, whereas there is no source of blurring in CT scans.

From a mathematical modeling perspective, let  $\langle \overline{\alpha_A}, \overline{\beta_A} \rangle$  and  $\langle \overline{\alpha_B}, \overline{\beta_B} \rangle$  represent two privacy-protected training samples, where  $\overline{\alpha} = (MI_{type \in (xr, ct)}^I)$  is a training image and  $\overline{\beta} = Dis$  is the corresponding label. For Mixup, the augmented sample is denoted as  $(\overline{\alpha}, \overline{\beta})$ , and the mixing process is given by the following Eq. (4) and Eq. (5):

$$\overline{\alpha} = \delta \overline{\alpha_A} + (1 - \delta) \overline{\alpha_B} \quad (4)$$

$$\overline{\beta} = \delta \overline{\beta_A} + (1 - \delta) \overline{\beta_B} \quad (5)$$

where,  $\delta \in [0, 1]$  is sampled from a beta distribution.

As for Cutmix, the generated augmented privacy-protected image is modeled as follows [see Eq. (6) and Eq. (7)]:

$$\overline{\alpha} = \vartheta [\Psi] \overline{\alpha_A} + (1 - \vartheta) [\Psi] \overline{\alpha_B} \quad (6)$$

$$\overline{\beta} = \delta \overline{\beta_A} + (1 - \delta) \overline{\beta_B} \quad (7)$$

where,  $\vartheta \in \{0, 1\}$  is a binary mask for determining which pixel is removed and replaced by the second image, and  $[\Psi]$  is the multiplication operation.

From a visual perspective, the application of both Mixup and Cutmix to medical images is illustrated in Fig. 9 and Fig. 10, respectively.

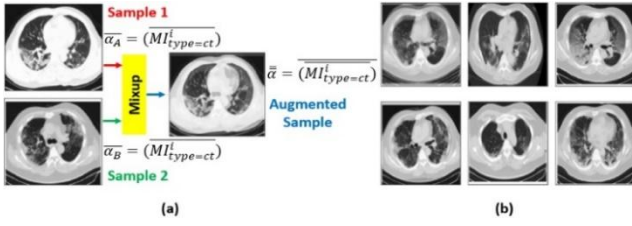


Fig. 9. Mixup-based augmentation. (a) Detailed operation; (b) different augmented samples.

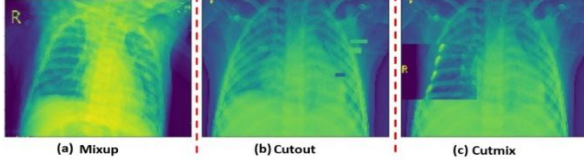


Fig. 10. Effectiveness of the Cutmix technique compared to Mixup and Cutout.

In Fig. 10(a), compared to Fig. 10(c), there are clear blurring regions that minimize the contribution of extracted features in the enhancement of training. Fig. 10(b) compared to Fig. 10(c), there are some informative parts out of the extracted space (the lower-right part of the left lung). In Fig. 10(c), the Cutmix technique avoids the pitfalls of both Mixup and Cutout, as a clear view of each specified part is used for the effective extraction of features.

#### E. Role of the Patch Embedder Component

This component is responsible for solving the problem of positional information losses by carrying out the Position-Aware Embedding (PAE) method on the augmented medical images. In detail, for a given augmented medical image ( $\bar{\alpha} = \text{MI}_{\text{type} \in \{xr, ct\}}^1$ ), which is generated from the source image ( $\text{MI}_{\text{type} \in \{xr, ct\}}^1$ ), these two images are divided into parts of the same area, and then they are flattened into a sequence of non-overlapping patches. Then, corresponding positional information and learnable class tokens are manipulated, as shown in Fig. 11.

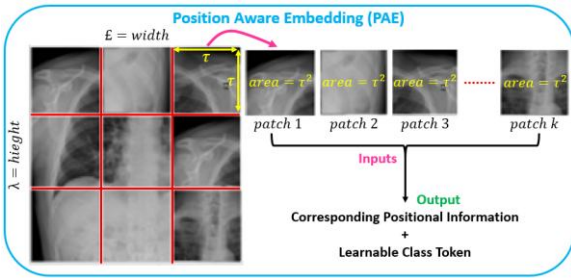


Fig. 11. Task of the patch embedder component when dealing with an X-ray image.

As shown in Fig. 11,  $\lambda$  and  $\mathcal{E}$  are the two dimensions of an augmented and privacy-protected medical image  $\bar{\alpha}$ . The area of  $\bar{\alpha}$  is given by the following Eq. (8):

$$\bar{\alpha}_r = \lambda \times \mathcal{E} \quad (8)$$

$\bar{\alpha}$  is divided into a number of patches (each one has an area equal to  $\tau \times \tau = \tau^2$ ). Relying on the total area of the medical

image and the area of a patch, the number of patches  $k$  is given by the following Eq. (9):

$$k = \frac{\bar{\alpha}_r}{\tau^2} \quad (9)$$

In terms of  $k$  patches,  $\bar{\alpha}$  is modeled as follows [see Eq. (10)]:

$$\bar{\alpha} = \bigcup_{i=1}^k (\bar{\alpha}_{pt}^i) \quad (10)$$

The generated flattened patches are passed to a linear embedding layer ( $\text{LYR}_e^1$ ) to adjust their dimensions in a way that suits the dimensions of the model ( $\text{DIM}_m$ ). The PAE method is applied to both the original image and the augmented instance. Fig. 12 provides a comprehensive view of a scene after the PAE method has been carried out.

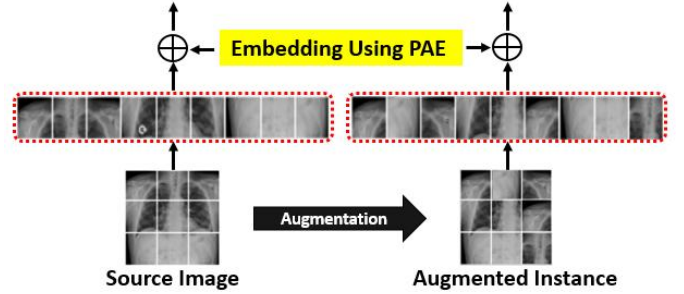


Fig. 12. Application of PAE on both a source image and an augmented image.

The application of PAE results in a series of patches that have the same spatial arrangement as in the source image. Let ( $\mathcal{T}_{lct}$ ) denote learnable class tokens and ( $\text{PS}_{info}$ ) denote the positional information. Then, the output of the PAE is given by the following Eq. (11):

$\text{PAE}_0^{\text{out}} = [\mathcal{T}_{lct}; \bar{\alpha} \text{LYR}_e^1] + \text{PS}_{info}$ , which can be detailed as follows:

$$\text{PAE}_0^{\text{out}} = [\mathcal{T}_{lct}; \bar{\alpha}_{pt}^1 \text{LYR}_e^1; \bar{\alpha}_{pt}^2 \text{LYR}_e^1; \dots; \bar{\alpha}_{pt}^k \text{LYR}_e^1] + \text{PS}_{info} \quad (11)$$

where,  $\text{LYR}_e^1 \in \mathcal{R}^{(\tau^2 ch) \times \text{DIM}_m}$ , and  $\text{PS}_{info} \in \mathcal{R}^{(k+2) \times \text{DIM}_m}$ .

#### F. Role of the Integrator Component

This component is responsible for ensuring effective feature extraction. It employs the concepts of the Vision Transformer (VisTrs) to carry out its mission [30]. The key idea behind the VisTrs is based on two fundamentals of performing image classification: 1) dealing with medical images as a series of patches that can be mapped into a semantic label, and 2) using a self-attention technique that enables effective feature extraction by handling the patches within an integration-based scan (i.e., integrate information at the level of the whole area of the image). This study adapts these two fundamentals and proposes an Integration-based Vision Transformer Structure (IbVTS). The structure of the VisTrs follows the encoder-decoder approach. In the context of the adaptation proposed in this work, the first fundamental of the VisTrs is satisfied by using a distill classifier (along with a corresponding distillation token), which will be discussed further on in this section when providing details on the role of the Trainer component. The second fundamental of the

VisTrs focuses on feature extraction and is satisfied as described below.

From a detail-based hierarchical perspective, the encoder of the VisTrs consists of a number of layers (NoL), and each layer contains two major units. The first unit is called multi-head self-attention (MhSa), which is responsible for carrying out the self-attention technique to ascertain the dependencies between the different patches that represent a given input image. The second unit is called the feed-forward unit (FeFoU), which is responsible for generating a fully/densely connected network of features. Relying on a normalization layer ( $LY_{nrm}$ ), the MhSa and FeFoU are connected. From a mathematical perspective, the two units can be modeled as follows [see Eq. (12) and Eq. (13)]:

$$\widehat{PAE}_{\varphi}^{out} = \text{MhSa} \langle LY_{nrm} (PAE_{\varphi-1}^{out}) \rangle + PAE_{\varphi-1}^{out} | \varphi = 1, 2, \dots, \text{NoL} \quad (12)$$

$$PAE_{\varphi}^{out} = \text{FeFoU} \langle LY_{nrm} (\widehat{PAE}_{\varphi}^{out}) \rangle + \widehat{PAE}_{\varphi}^{out} | \varphi = 1, 2, \dots, \text{NoL} \quad (13)$$

The operations that are executed within the self-attention technique are as follows:

- 1) For a given input sequence of image features,  $\begin{cases} k: \text{key} \\ q: \text{query} \\ v: \text{value} \end{cases}$  are calculated for each element included in the sequence.
- 2) Relying on the inner/dot product, the relevance between the current element and other elements is obtained. In other words, this step highlights and integrates the relative importance of the patches in the sequence.
- 3) The results of step 2 are scaled based on the dimension of key ( $K_{dim}$ ) as a scaling factor and then used as inputs for the Softmax function.
- 4) To focus on more important values, the output of the Softmax function is multiplied by  $v$ .

Eq. (14) and Eq. (15) summarize the operations of the self-attention technique.

$$\begin{cases} k: \text{key} \\ q: \text{query} = PAE^{input} \cup_{k,q,v} | \cup_{k,q,v} \in \mathcal{R}^{DIM_m \times 3K_{dim}} \\ v: \text{value} \end{cases} \quad (14)$$

$$\text{SeAt}(PAE^{input}) = \text{softmax} (qk^T / \sqrt{K_{dim}}) \cdot v \quad (15)$$

The environment (patches) within which the self-attention (SeAt) is executed allows for the parallel execution of these operations. Here, the MhSa, considered an extension of SeAt, comes to the fore. To activate self-attention (SeAt) in a parallel manner, instead of using a single value for the triplet ( $k, q, v$ ), multiple values are used. Consequently, we deal with ( $h$ ) heads (i.e., a head for each attention). The results of all the attention heads are concatenated to express the first unit of the encoder of the VisTrs [see Eq. (16)].

$$\text{MhSa} (PAE^{input}) = \text{concat} \langle \text{SeAt}_1 (PAE^{input}); \text{SeAt}_2 (PAE^{input}); \dots; \text{SeAt}_h (PAE^{input}); \rangle | lw \quad (16)$$

where,  $lw$  represents the learnable weights,  $lw \in \mathcal{R}^{h.K_{dim} \times DIM_m}$ .

As for FeFoU, it is constructed by a fully connected layer, followed by GeLU, followed by another fully connected layer.

Fig. 13 presents the Integrator component, which is based on the Patch Embedder component.

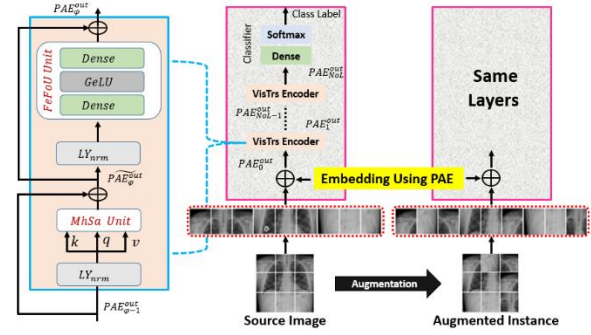


Fig. 13. A VisTrs architecture-based Integrator built based on the mission of the patch embedder component.

### G. Role of the Trainer and Classifier Components

These components are responsible for training the model in a way that enhances both model performance and prediction accuracy. Originally, the VisTrs is driven from the data-efficient image transformer [31], where minimizing training is a major goal based on the student-teacher approach. The student-teacher approach relies on a distillation token generated from the augmented image instance. So, the left part of Fig. 13 represents the student, and the right part represents the teacher. Based on this, Eq. (11) is updated to become the following Eq. (17):

$$PAE_0^{out} = [\mathcal{T}_{lct}; \mathcal{T}_{dst}; \bar{\alpha}_{pt}^1 LY_e^1; \bar{\alpha}_{pt}^2 LY_e^2; \dots; \bar{\alpha}_{pt}^k LY_e^k] + PS_{info} \quad (17)$$

where,  $LY_e^1 \in \mathcal{R}^{(\tau^2 ch) \times DIM_m}$ ,  $PS_{info} \in \mathcal{R}^{(k+2) \times DIM_m}$ , and  $\mathcal{T}_{dst}$  denote the distillation token. In other words, the Trainer uses extracted features to train two classifiers, where each one has a dense layer followed by the Softmax function. The student classifier provides the class token, while the teacher classifier provides the distill token. Both of them are used as inputs for the fusion layer, which provides the final class of the medical image according to the following Eq. (18):

$$C_{final} = \frac{1}{2} \times (C_{token}^{class} + C_{token}^{distill}) \quad (18)$$

Fig. 14 illustrates the Trainer's task, which is based on that of the Integrator.

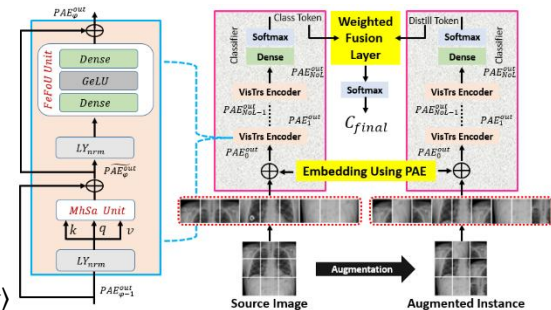


Fig. 14. Mission of the trainer component.

In reality, in the context of classification, a given medical image can be classified as healthy/normal, COVID-19, or pneumonia. To enable binary (healthy or COVID-19) or multi-class modes (binary plus pneumonia), a loss function is used in



the Classifier component, and this function is expressed as follows [see Eq. (19)]:

$$Q(\mathfrak{B}_{ij}, \mathfrak{W}_{ij}) = \frac{-1}{\text{pat}} \sum_{i=1}^{\text{trg}} \sum_{j=1}^{\text{ds}} \mathfrak{W}_{ij} \log \left( \frac{1}{1+e^{-\mathfrak{B}_{ij}}} \right) + (1 - \mathfrak{W}_{ij}) \log \left( 1 - \frac{1}{1+e^{-\mathfrak{B}_{ij}}} \right) \quad (19)$$

where, trg, ds are the number of training images and defined classes.  $\mathfrak{W}_{ij}$  represents the ground-truth labels.  $\mathfrak{B}_{ij}$  is the predicted probability.

#### H. Details of the Proposed System's Architecture

A sequence diagram, Fig. 15, is used to show the tasks of the components involved in the architecture of the proposed system.

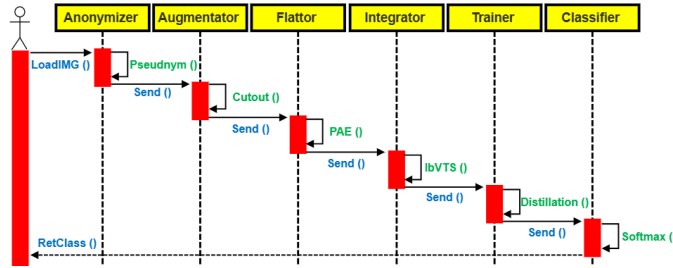


Fig. 15. Sequence diagram based on architecture details.

#### IV. EXPERIMENTAL RESULTS AND EVALUATION

In this work, experiments were conducted on two different datasets to construct (train) and test the proposed diagnosing system. One of these datasets is the COVIDx dataset, created to store X-ray medical images [32]. This dataset consists of 13,962 X-ray images distributed over three classes, as summarized in Table II. We also used the SARS-CoV-2 dataset, which contains CT scan-derived medical images [33]. This dataset comprises 2,482 CT scan images distributed over two classes, as summarized in Table III.

TABLE II. DESCRIPTION OF THE COVIDX DATASET

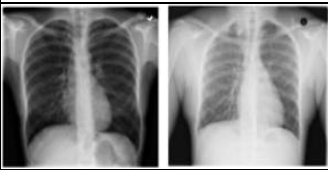
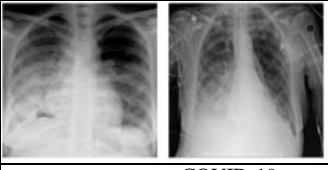

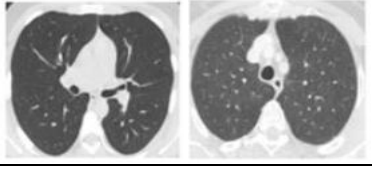
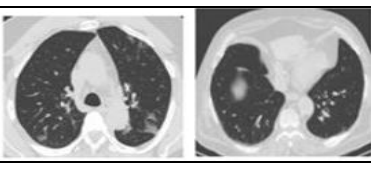
Class 1	Healthy	
Number of images	8066	
Samples		
Class 2	Pneumonia	
Number of images	5538	
Samples		
Class 3	COVID-19	
Number of images	358	
Samples		

TABLE III. DESCRIPTION OF THE COVIDX DATASET OVER TWO CLASSES

Class 1	Healthy	
Number of images	1230	
Samples		
Class 2	COVID-19	
Number of images	1252	
Samples		

#### A. Types of Evaluation

As shown in Fig. 16, we used two types of evaluation metrics: visual evaluation and numeric evaluation.

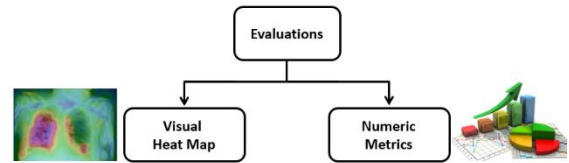


Fig. 16. Types of evaluation metrics used in the present study.

#### B. Used Metrics

Four metrics were used in this study for evaluation purposes. All of them were derived from the confusion matrix summarized in Table IV.

TABLE IV. CONFUSION MATRIX

Predicted class Actual class	$C_{\text{final}}$	$\neg C_{\text{final}}$	Sum
$C_{\text{final}}$	TP	FN	S
$\neg C_{\text{final}}$	FP	TN	Su
Sum	$\bar{S}$	$\bar{S}_u$	All

The elements of the above confusion matrix are as follows:

- True positives (TPs): The positive images that were correctly labeled by the classifier.
- True negatives (TNs): The negative images that were correctly labeled by the classifier.
- False positives (FPs): The negative images that were incorrectly labeled as positive (e.g., images of class healthy = no for which the classifier predicted healthy = yes).
- False negatives (FNs): The positive images that were mislabeled as negative (e.g., images of class healthy = yes for which the classifier predicted healthy = no).

$$\text{Accuracy} = \frac{TP+TN}{\text{All}} \quad (20)$$

Accuracy refers to the recognition rate, which quantifies the percentage of test set images that have been correctly classified.



In other words, it reflects how well the classifier recognizes images from the various classes [see Eq. (20)].

$$\text{Precision} = \frac{TP}{S} \quad (21)$$

Precision refers to the exactness, which refers to the percentage (%) of images that the classifier correctly labeled as positive. Therefore, it determines how many of the positive predictions are correct [see Eq. (21)].

$$\text{Recall} = \frac{TP}{S} \quad (22)$$

Recall (also known as sensitivity) refers to the completeness, which refers to the percentage (%) of positive tuples that the classifier labeled as positive. Recall is used as a measure of a classifier's tendency to identify infected cases [see Eq. (22)].

$$\text{Specificity} = \frac{TN}{Su} \quad (23)$$

Specificity refers to the true negative recognition rate. In other words, it measures the ability of a classifier to detect non-infected cases [see Eq. (23)].

It is worth mentioning that in terms of semantic meaning, a high value for any of the metrics mentioned above is preferable and reflects better performance, and vice versa.

Regarding visual evaluation, heat maps were employed to reflect the attention maps used in the learning process. Heat maps are used to illustrate the gradually increasing focus on regions of interest over layers during the learning process. This, in turn, reflects how regions are tightly coupled with the predicted class when it comes to the visual monitoring of progress.

It is worth mentioning that experiments are conducted in the Artificial Intelligence and Multi Media (AIMD) Lab within the campus of Fahad Bis Sultan University (FBSU) at Computing College.

### C. Evaluation Based on X-ray Dataset

The COVIDx dataset was divided into a training part (25 %) and a testing part (75 %), taking into account the four metrics used for each class. Table V summarizes the results obtained with and without using augmentation.

TABLE V. EXPERIMENTAL RESULTS BASED ON X-RAY IMAGES FROM THE COVID DATASET

Approach	Term Metric	AVG	Class		
			Healthy	COVID-19	Pneumonia
Without Augmentation	Accuracy	≈ 87.16	88.23	87.77	85.47
	Precision	≈ 87.37	87.11	86.81	88.2
	Recall	≈ 87.84	89.21	88	86.31
	Specificity	≈ 87.11	86.26	87.62	87.45
With Augmentation	Accuracy	≈ 92.62	95.88	90	91.98
	Precision	≈ 93.47	94.63	92.12	93.66
	Recall	≈ 93.22	97.52	90	92.13
	Specificity	≈ 95.98	93.83	98.57	95.53

The values documented in Table V represent good performance with regard to diagnosing positive and negative cases. However, to highlight the effectiveness of the proposed augmentation technique, Fig. 17 compares the averages of the two approaches (with and without augmentation) for each class.

Fig. 17 reflects considerable enhancements for the four metrics when using augmentation, where it contributes 5.46, 6.1, 5.38, and 8.87 enhancements for accuracy, precision, recall, and specificity, respectively. The reason behind this is related to the effective extraction of features used for training, since the augmentation technique used enabled an increased focus on the regions of interest. This justification is visually supported by the heat map shown in Fig. 18.

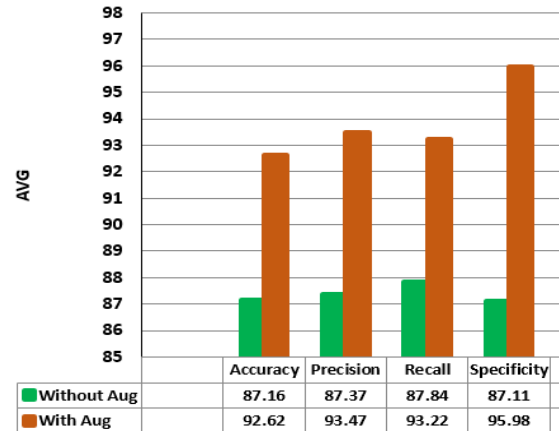


Fig. 17. Comparison between the augmentation-based approach and the non-augmentation-based approach on the X-ray dataset.

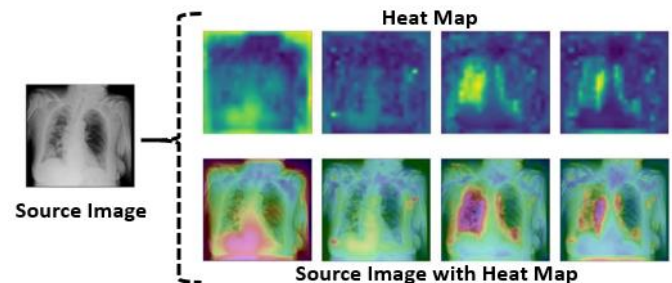


Fig. 18. Heat map of a source COVID-19 X-ray image.

Fig. 18 shows the gradually increasing focus on the zones that are informative for diagnosing/predicting illness. The heat map above reflects the confused focus on random zones of the lung and the point at which the focus becomes more specific on the zones that are tightly coupled with the class of the image, reaching the zones that represent COVID-19 in terms of medical diagnosis.

To measure the resistance of the proposed system against a small-sized training dataset (and to, in turn, evaluate its potential for usage in reality), we repeated our experiments using a different training/testing data ratio (stepwise increase of 10 %), meaning that 20 % of the COVIDx dataset was used for training and 80 % was used for testing, as shown in Fig. 19.

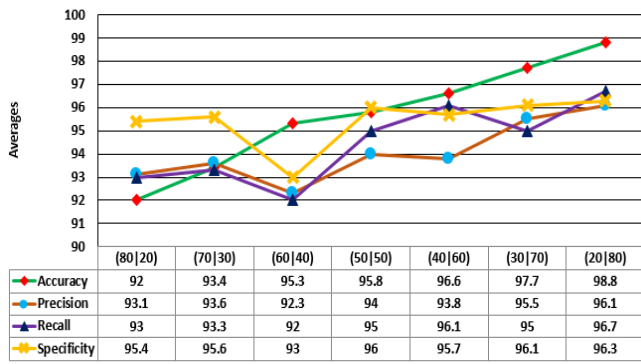


Fig. 19. Average of four metrics over different training/testing data splits.

Fig. 19 presents the results of seven experiments, each of which had its own unique training/testing division ratio. For example, in our second experiment, 30% of the dataset was allocated for training, and 70% was allocated for testing, meaning that this split can be expressed as 30/70. Regarding accuracy, it is obvious that there was a gradual increase in enhancement from about 92% in the first experiment to about 99% in the seventh one. This means that allocating more data for training leads to greater accuracy. However, there is a gap (about 8.8) in accuracy values between the 20/80 training ratio and its counterpart, the 80/20 training ratio. However, in both cases, the accuracy is still within a considerable range, which is a good indicator of the proposed system's potential for use in reality. The other metrics fluctuated randomly in response to the different training ratios. In this context, it is worth mentioning that accuracy is the most important measure to consider when evaluating a system.

#### D. Evaluation Based on CT Scan Dataset

To facilitate a fair comparison between our evaluation of the proposed system based on X-ray images and CT scan images, the same evaluation procedure followed for our experiments on the X-ray images was repeated. Therefore, the SARS-CoV-2 dataset was divided into a training part (25 %) and a testing part (75 %), taking into account the four metrics used for each class. Table VI summarizes the results obtained with and without using augmentation.

TABLE VI. EXPERIMENTAL RESULTS BASED ON CT SCAN IMAGES FROM THE SARS-CoV-2 DATASET

Approach	Term Metric	AVG	Class	
			Healthy	COVID-19
Without Augmentation	Accuracy	≈ 94.72	94.44	95
	Precision	≈ 94.36	93.88	94.83
	Recall	≈ 94.86	94.69	94.98
	Specificity	≈ 94.32	94.54	94.09
With Augmentation	Accuracy	≈ 97.83	97.55	98.11
	Precision	≈ 97.49	97.21	97.76
	Recall	≈ 97.68	97.33	98.03
	Specificity	≈ 96.85	96.57	97.13

The values documented in Table VI indicate that the model's performance was better than that achieved via the use of the X-ray images in terms of diagnosing positive and negative cases.

However, to highlight the effectiveness of the proposed augmentation technique, Fig. 20 compares the averages of the two approaches (with and without augmentation) for each class.

Fig. 20 shows that considerable improvements in the four metrics were achieved when using augmentation, as enhancements of 3.11, 3.13, 2.82, and 2.53 were recorded for accuracy, precision, recall, and specificity, respectively. The results derived from the use of CT scan images are better than those derived from the use of X-ray images. The CT scan-derived results' superiority can be attributed to the ability of the Mixup augmentation technique to overcome the blurring issue that arises due to the ribs obscuring features in X-ray images. This justification is visually supported by the heat map shown in Fig. 21.

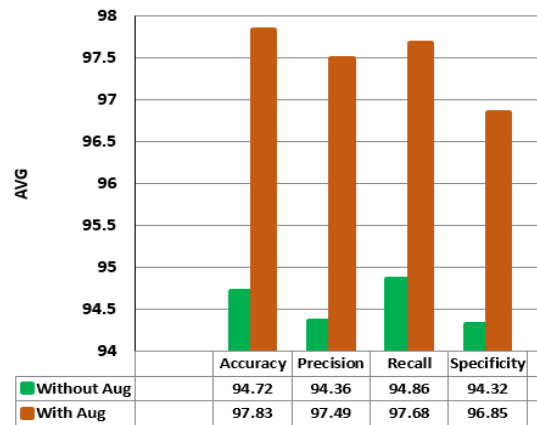


Fig. 20. Comparison between the augmentation-based approach and the non-augmentation-based approach when using the CT-scan dataset.

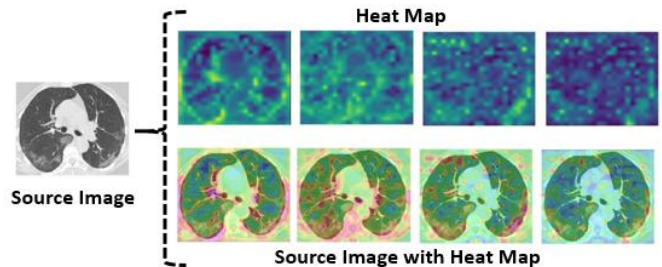


Fig. 21. Heat map of a source COVID-19 CT scan image.

As was the case with the X-ray images presented in Fig. 18, Fig. 21 shows the gradually increasing focus on the zones that are informative for diagnosing/predicting illness. In other words, the system learned to highlight sets of pixels that are strongly linked with illness. It is worth mentioning that the CT scanning technique is more accurate than X-ray imaging, which means that learning on CT scan images leads to higher levels of accuracy, especially if they are supported by augmentation, since augmentation leads to the extraction of more meaningful features.

To measure the resistance of the proposed system against a small-sized CT-based training dataset, the same procedure adopted to assess the resistance of the X-ray dataset was repeated, as shown in Fig. 22.

The results of the seven experiments (shown in Fig. 22) provide strong proof of the effectiveness of applying augmentation on CT scan images to increase the level of accuracy and other metrics as the training ratio increases. Steady enhancements for all metrics were recorded. Although the accuracy increases, there is a small gap (3.15) between the accuracy value when using an 80|20 testing and training ratio and that when using a 20|80 testing and training ratio. This proves that the proposed system has the ability to be trained on a small amount of data, which further indicates its potential for use in reality.

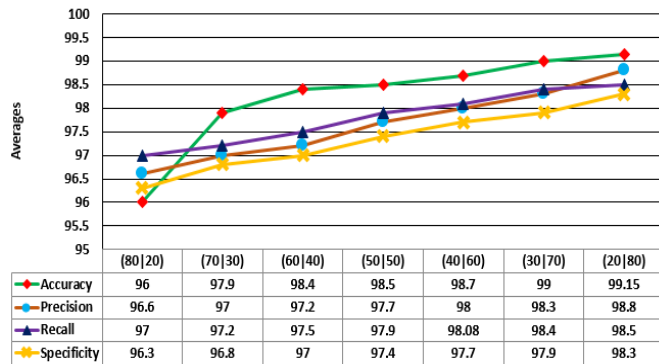


Fig. 22. Average of the four metrics over different training/testing data splits in the CT-based training dataset.

### E. Security Analyzing

To analyze the proposed model's resistance against compromising the privacy of patients, let  $P_{\text{Guess}}^e$  denote the probability that a malicious party can successfully guess whether the event  $e$  is true. The proposed privacy protection method is resistant if [see Eq. (24)]:

$$P_{\text{Guess}}^{e_1=\text{personal info}_i \in \text{RDS}} = P_{\text{Guess}}^{e_1=\text{personal info}_j \in \text{RDS}} \quad \forall (0 < i \neq j \leq \text{pat}) \quad (24)$$

Proof: The attacker cannot obtain any benefit from employing their side's information to recognize the patient. This is because the lock space prevents the attacker from knowing the ID of the patient. Moreover, the first name and last name are masked by dummies of the same size (i.e., a first and last name with the same number of letters as the patient's first and last name), which, in turn, leads to more confusion on the attacker's side. Furthermore, the generalization tactic applied to the date of birth of the patient does not reveal any specific information about the age of the patient. Only the attacker can know the gender of the patient and whether they have COVID-19. As a result, the attacker is forced to randomly guess the personal information of the patient, which leads to uncertainty in inferring sensitive information when attempting to compromise a patient's privacy and potentially harm them.

### V. CONCLUSION

In this study, a deep learning-based system dedicated to detecting COVID-19 using X-ray- and CT scan-derived medical images is presented. The system is managed by six components. The Anonymizer is responsible for preserving the privacy of patients through the use of a pseudonym-based anonymity approach. The Augmentator is responsible for increasing the

quality of the medical images through the use of augmentation techniques (Cutmix and Mixup for X-ray and CT scan images, respectively). The Patch Embedder is responsible for solving the problem of positional information loss by applying the Position-Aware Embedding method to the augmented medical images. The Integrator is responsible for effective feature extraction, which is carried out through employing the concepts of the Vision Transformer Structure. The Trainer and Classifier components are responsible for training the proposed system based on the student-teacher approach and generating the final classes via the Softmax function. The COVIDx (X-ray images) and SARS-CoV-2 (CT-scan images) datasets were utilized in this study to train and test the proposed system. Our evaluations, which were based on accuracy, precision, recall, and specificity metrics, showed that there was an enhancement in values when using augmentation. However, the enhancement achieved after training on CT scan images was better than that achieved after training on X-ray images. Moreover, the system was evaluated using a small amount of data by reversing the training/testing ratios. The results corresponding to the use of X-ray images were promising, as the system achieved 92% accuracy when 20% of the data were allocated for training and 80% of the data were allocated for testing and 98.8% accuracy when 80% of the data were allocated for training and 20% of the data were allocated for testing. However, compared to the results derived from the use of X-ray images, the results corresponding to the use of CT scan images were better, with an accuracy value of 96% being achieved after 20% of the data were allocated for training and 80% of the data were allocated for testing, as well as an accuracy value of 99.15% being achieved after 80% of the data were allocated for training and 20% of the data were allocated for testing. The key limitation of this work is related to the fact that we ignored evaluating performance in terms of the time taken to generate predictions and complexity. In future work, we will address this limitation by adopting a parallel training approach. In addition, testing the proposed system using additional datasets as well as including statistical significance testing (e.g., confidence intervals, p-values) will be in future work.

### REFERENCES

- [1] Guo, Yan-Rong, et al. "The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak—an update on the status." *Military Medical Research* 7.1 (2020): 1-10.
- [2] Lauer, Stephen A., et al. "The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application." *Annals of internal medicine* 172.9 (2020): 577-582.
- [3] Nawaz, Ahsan, et al. "Epidemic spread and its management through governance and leadership response influencing the arising challenges around COVID-19 in Pakistan—a lesson learnt for low income countries with limited resource." *Frontiers in public health* 8 (2020): 573431.
- [4] Chen, Wei, et al. "Reducing false negatives in COVID-19 testing by using microneedle-based oropharyngeal swabs." *Matter* 3.5 (2020): 1589-1600.
- [5] Iliescu, F. S., et al. "Point-of-Care Testing-the Key in the Battle against SARS-CoV-2 Pandemic." *Micromachines* 2021, 12, 1464." (2021).
- [6] Koyyada, Shiva Prasad, and Thipendra P. Singh. "A Systematic Survey of Automatic Detection of Lung Diseases from Chest X-Ray Images: COVID-19, Pneumonia, and Tuberculosis." *SN Computer Science* 5.2 (2024): 229.
- [7] Alafif, Tarik, et al. "Machine and deep learning towards COVID-19 diagnosis and treatment: survey, challenges, and future directions." *International journal of environmental research and public health* 18.3 (2021): 1117.



- [8] Cossio, Manuel. "Augmenting Medical Imaging: A Comprehensive Catalogue of 65 Techniques for Enhanced Data Analysis." arXiv preprint arXiv:2303.01178 (2023).
- [9] Abdollahi, Behnaz, Naofumi Tomita, and Saeed Hassanpour. "Data augmentation in training deep learning models for medical image analysis." *Deep learners and deep learner descriptors for medical applications* (2020): 167-180.
- [10] Altaf, Fouzia, et al. "Going deep in medical image analysis: concepts, methods, challenges, and future directions." *IEEE Access* 7 (2019): 99540-99572.
- [11] Alafeef, Maha, and Dipanjan Pan. "Diagnostic approaches for COVID-19: lessons learned and the path forward." *ACS nano* 16.8 (2022): 11545-11576.
- [12] Wu, Song, et al. "Deep Learning and Medical Imaging for COVID-19 Diagnosis: A Comprehensive Survey." arXiv preprint arXiv:2302.06611 (2023).
- [13] Panday, Aishwarza, Muhammad Ashad Kabir, and Nihad Karim Chowdhury. "A survey of machine learning techniques for detecting and diagnosing COVID-19 from imaging." arXiv preprint arXiv:2108.04344 (2021).
- [14] Awassa, Lamia, et al. "Study of Different Deep Learning Methods for Coronavirus (COVID-19) Pandemic: Taxonomy, Survey and Insights." *Sensors* 22.5 (2022): 1890.
- [15] Wang, Xinggang, et al. "A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT." *IEEE transactions on medical imaging* 39.8 (2020): 2615-2625.
- [16] Singh, Shrinjal, et al. "CNN based Covid-aid: Covid 19 Detection using Chest X-ray." 2021 5th International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2021.
- [17] Lu, Siyuan, et al. "An explainable framework for diagnosis of COVID-19 pneumonia via transfer learning and discriminant correlation analysis." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17.3s (2021): 1-16.
- [18] Sahinbas, Kevser, and Ferhat Ozgur Catak. "Transfer learning-based convolutional neural network for COVID-19 detection with X-ray images." *Data science for COVID-19*. Academic Press, 2021. 451-466.
- [19] Chaudhary, Suman, and Yan Qiang. "Ensemble deep learning method for Covid-19 detection via chest X-rays." 2021 *Ethics and Explainability for Responsible Data Science (EE-RDS)*. IEEE, 2021.
- [20] Upadhyay, Kamini, Monika Agrawal, and Desh Deepak. "Ensemble learning - based COVID - 19 detection by feature boosting in chest X - ray images." *IET Image Processing* 14.16 (2020): 4059-4066.
- [21] Goel, Tripti, et al. "Automatic screening of COVID-19 using an optimized generative adversarial network." *Cognitive computation* (2021): 1-16.
- [22] Ardakani, Ali Abbasian, et al. "Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks." *Computers in biology and medicine* 121 (2020): 103795.
- [23] Er, Mehmet Bilal. "COVID - 19 detection based on pre - trained deep networks and LSTM model using X - ray images enhanced contrast with artificial bee colony algorithm." *Expert Systems* 40.3 (2023): e13185.
- [24] Hasan, Ali M., et al. "Classification of Covid-19 coronavirus, pneumonia and healthy lungs in CT scans using Q-deformed entropy and deep learning features." *Entropy* 22.5 (2020): 517.
- [25] Tyagi, Amit Kumar, and N. Sreenath. "A comparative study on privacy preserving techniques for location based services." *British Journal of Mathematics & Computer Science* 10.4 (2015): 1-25.
- [26] Chlap, Phillip, et al. "A review of medical image data augmentation techniques for deep learning applications." *Journal of Medical Imaging and Radiation Oncology* 65.5 (2021): 545-563.
- [27] DeVries, Terrance, and Graham W. Taylor. "Improved regularization of convolutional neural networks with cutout." arXiv preprint arXiv:1708.04552 (2017).
- [28] Zhang, Hongyi, et al. "mixup: Beyond empirical risk minimization." arXiv preprint arXiv:1710.09412 (2017).
- [29] Yun, Sangdoo, et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.
- [30] Yang, Jianwei, et al. "Focal self-attention for local-global interactions in vision transformers." arXiv preprint arXiv:2107.00641 (2021).
- [31] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." *International conference on machine learning*. PMLR, 2021.
- [32] Wang, Linda, Zhong Qiu Lin, and Alexander Wong. "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images." *Scientific reports* 10.1 (2020): 19549.
- [33] Soares, Eduardo, et al. "SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification." *MedRxiv* (2020): 2020-04.