Interpretable CNN-Multilevel Attention Transformer for Rapid Recognition of Pneumonia From Chest X-Ray Images

Shengchao Chen, Sufen Ren, Guanjun Wang, Mengxing Huang, and Chenyang Xue

Abstract—Chest imaging plays an essential role in diagnosing and predicting patients with COVID-19 with evidence of worsening respiratory status. Many deep learning-based approaches for pneumonia recognition have been developed to enable computer-aided diagnosis. However, the long training and inference time makes them inflexible, and the lack of interpretability reduces their credibility in clinical medical practice. This paper aims to develop a pneumonia recognition framework with interpretability, which can understand the complex relationship between lung features and related diseases in chest X-ray (CXR) images to provide high-speed analytics support for medical practice. To reduce the computational complexity to accelerate the recognition process, a novel multilevel self-attention mechanism within Transformer has been proposed to accelerate convergence and emphasize the task-related feature regions. Moreover, a practical CXR image data augmentation has been adopted to address the scarcity of medical image data problems to boost the model’s performance. The effectiveness of the proposed method has been demonstrated on the classic COVID-19 recognition task using the widespread pneumonia CXR image dataset. In addition, abundant ablation experiments validate the effectiveness and necessity of all of the components of the proposed method.

Index Terms—Pneumonia recognition, self-attention mechanism, convolutional neural network, vision transformers, interpretability.

I. INTRODUCTION

COVID-19 was first discovered in December 2019 in Wuhan, Hubei Province, China, and its ability to spread robustly has led to a global pandemic of it and its variant strains (Omicron BA.1, Omicron BA.2, etc.) [1]. As of September 10, 2022, more than 607,362,097 people have been diagnosed with COVID-19 positive worldwide, with a death toll of 6,510,110. Its main symptoms include fever, shortness of breath, and loss of taste [2]. In some cases, the infection can lead to serious respiratory complications, most notably pneumonia, which often requires admission to Intensive Care Unit (ICU). Unfortunately, the dramatic increase in diseases has paralyzed many health systems, and the shortage of wards and medical equipment has forced many countries to designate a series of epidemic prevention policies, such as distance restraints and national lockdowns [3].

Real-time reverse transcription polymerase chain reaction (RT-PCR) is the primary diagnostic tool for COVID-19 [4]. In clinical practice, chest X-ray (CXR) have long been used to evaluate patients’ lungs because of their portability and cost-effectiveness compared to computed tomography (CT) [5]. However, the resulting images require manual analysis by radiologists, making the diagnostic process inefficient and difficult to cater to large-scale diagnostic needs. This diagnostic strategy, which requires the practical intervention of experts and is time-consuming and laborious, needs urgent improvement.

Rapid advances in deep learning (DL) techniques have enabled the development of a practical computer-aided diagnostic tool to detect COVID-19-related diseases from large-scale CXR images. DL allows computational models consisting of multiple layers that contain several processing units to learn distinct feature abstraction, thus avoiding direct interaction with skilled experts. For example, lots of work trains convolutional neural network (CNNs)-based models to recognize COVID-19 from CXR images [6], [7], [8], [9], [10], [11], [12], [13], the key is to establish the relationship between the feature distribution in lung region and disease labels. However, the shape and location of lungs, speckled, hazy, irregular, hairy glassy cloudiness, and reticular cloudiness [14] contribute to the recognition results to varying degrees, which means not all features are efficient and discovering task-related features is crucial to boost the recognition accuracy. Inspired by the human visual system, the attention

Digital Object Identifier 10.1109/JBHI.2023.3247949
mechanism has been widely used in image-based tasks [15], [16], [17], [18], [19], [20], [21]. These encourage models to focus on task-related information regions through reshaping features while suppressing the interference of useless peripheral information to boost performance. Nevertheless, adding the attention mechanism inevitably increases the computational complexity, indirectly leading to its slowing down of learning and inference, contrary to the original intention of achieving high-speed recognition.

What’s worse, the superiority of DL relies heavily on pre-collected datasets, a common bottleneck of data-driven approaches. The major problem with medical images is the disadvantage in size due to the cumbersome collection process, making it challenging to train a robust network adequately. Image-based augmentation approaches, such smoothing, noise addition, flipping, image masking, are considered an effective solution to above problem by extending datasets. However, their generations’ distribution may be detached from the original image distribution. DL-based augmentation methods such as autoencoders (AEs) [22] and generative adversarial networks (GANs) [23] are widely used to generate data with similar distribution to the original dataset [24], [25]. Unfortunately, requiring tedious data processing and network training makes them hard to satisfy the low-resource medical applications. Moreover, DL models that regarded as ‘black boxes’ because of large randomness are hard to gain confidence in practical CXR images-based pneumonia recognition tasks [26].

Therefore, this paper aims to design a framework with interpretability and fast recognition of pneumonia, especially COVID-19-positive, for providing reliable analytic support to computer-aided medical diagnosing systems.

To this end, we propose a model cascades CNN and Transformer, dubbed CMT, to understand complex lung’s feature correlation with specific pneumonia with CXR images. First, pre-trained CNN-based models have been considered feature extractors for abstracting complex features from input CXR images and convert them into feature embedding. Secondly, the Transformer gets these embedding to explore the importance of a feature so that we understand the relationship between task-related features and specific pneumonia diseases.

Specifically, to relieve the negative implication of medical data scarcity on recognition performance, a practical CXR image data augmentation is proposed to extend the dataset as the pre-work of recognition, which generated new images undeviating from original distribution at a low cost. Moreover, we introduce a novel self-attention mechanism within Transformer, named Multilevel Multi-Head Self-Attention (MMSA), to reduce the overall computational complexity while keeping excellent recognition accuracy by computing different levels’ importance. The proposed CMT provides a novel platform for pneumonia recognition and can be extended to other medical image recognition tasks.

To verify the effectiveness of the proposed CMT, We fused patient lung images collected from five different medical institutions to produce a complete pneumonia identification dataset and conducted extensive experiments on this dataset. The main contribution of this work can be summarized in four-fold:

- To provide high-efficient pneumonia (especially COVID-19) recognition solution with credibility in a clinic medical diagnosing practices, we proposed CNN-MMSA-Transformer (CMT) that cascaded CNN and Transformer with low computational complexity attention mechanism, to analyze the CXR images.
- To reduce the computational complexity of Transformer and improve the task-related lung feature perception ability for adapting low-resourced medical applications, we propose Multilevel Multi-Head Self-Attention (MMSA) (Section III-C2), which outperform conventional attention mechanisms with ultra-fast convergence and powerful feature perception while guaranteeing excellent performance.
- To mitigate the negative impact of medical image data scarcity on the model recognition effect, a random beta sampling (RBS)-based feature fusion data augmentation approach, named FFA (Section II), was proposed by distributing non-divergent augmentation of the original CXR image dataset.
- To boost the credibility of the proposed CMT in clinical medical diagnosing applications, the model’s interpretability has been explored and validated by by visualizing the weight matrix and feature activation map during inference process (Section IV-G).

The remainder of this work is in the following: Section II reviews related works. Section III describes the proposed approach in detail. Section IV describes the specific implementation details of the experiment and shows the results of experiments. Section V concludes this work.

II. RELATED WORK

This section summarizes the relevant representative approaches on this work, which are divided into three parts, including COVID-19 recognition with CNN, Self-Attention mechanism, and Vision Transformers (ViTs).

A. COVID-19 Recognition With CNN

Since COVID-19 has become widespread, many works have described the application of CNN and DL in diagnosing the disease based on medical images and have achieved promising results. CNNs such as ResNet [27], VGGNet [28], and AlexNet [29] have shown excellent performance in computer vision tasks such as object detection, image classification, and semantic segmentation. Lots of works have recently applied the CNN-based models in COVID-19 recognition [6], [7], [8], [9], [10], [11], [12], [13]. For example, Wang et al. proposed COVID-Net to perform a three-way recognition in COVID-19/Non-COVID-19 [6]. Heidari et al. used vanilla VGGNet to classify COVID-19-based disease from CXR image [7]. Considering the low-resourced computer constrain in medical institutions, fully training a CNN-based model based on knowledge is expensive. For this reason, Apostolopoulos et al. applied a pre-trained CNN-based model based on transfer learning (TL) to achieve three-way CXR image recognition while satisfying the computational restriction [8]. A series of related works that used pre-trained models to accelerate the convergence has surface, e.g., CNN-LSTM [9], CoroNet [11], etc. However, these methods ignore other pneumonia that remains in the CXR image dataset that not limited to classifying COVID-19/Non-COVID-19, so that could be lead to a high cost to build different models to recognize different pneumonia. To overcome the problem, some researchers view the COVID-19 recognition task as a multi-class classification problem [30], [31], [32], [33], [34], [35], [36], [37].
network models by improving the attention mechanism, i.e., introducing formaldehyde and class-specific attention for each channel. Fan et al. [46] proposed a multiscale ViT structure with multiscale feature pyramids. Zhou et al. [47] increased their diversity at different levels of representation by regenerating attention maps. However, these methods inevitably bring higher computational cost while enhancing performance. This paper proposes a pneumonia recognition framework based on Transformer with a novel low computational complexity attention mechanism, which can effectively capture the relationship between global and local features while maintaining the high-speed training and inference process.

### III. METHODOLOGY

The architecture of the proposed recognition framework is shown in Fig. 1, which consists of two parts. The first part is the CXR image feature fusion data augmentation (FFA) that takes effect only in the model training phase, which receives the original CXR images and generates augmented images with a similar distribution. The second part is the CMT consisting of a CXR image feature extractor (CIFE) and Multilevel-Self-Attention Transformer (MMSA Transformer), where CIFE is used to extract deeper feature information in CXR images. MMSA Transformer is used to receive this feature information to explore and construct complex nonlinear relationships between global feature, local feature, and disease information. The CMT receives data from the original and augmented datasets for training. The details of their design will be specified next.

### A. CXR Image Feature Fusion Augmentation (FFA)

Unlike other types of image dataset, the scarcity of medical image data hinders further performance improvement of data-driven approaches. An efficient CXR image augmentation strategy, named Feature Fusion Augmentation (FFA), is proposed to overcome this problem. FFA is used for online augmentation based on the original CXR images to extend the original dataset to improve recognition accuracy. The specific implementation of the proposed FFA algorithm is shown in Algorithm 1. The basic idea is to randomly fuse two CXR images under the same category based on the weight matrix generated by random beta sampling (RBS), which makes the generated images not deviate from the original distribution while maintaining the essential features.

The schematic diagram of the FFA algorithm is shown in Fig. 2. Specifically, first, two CXR images, $I_1 \in \mathbb{R}^{h \times w}$ and $I_2 \in \mathbb{R}^{h \times w}$, under the same class are randomly selected and divided equally according to patch size $p$, such that the size of the divided region is $I_d \in \mathbb{R}^{\frac{h}{p} \times \frac{w}{p}}$. Then, fixed weights $\zeta$ and $1-\zeta$ are obtained according to RBS, and similarly, RBS is applied at each point of the $\frac{h}{p} \times \frac{w}{p}$ region to obtain different weights $\bar{z}$ and $1-\bar{z}$. The Beta distortion is the probability distribution over the continuous variable $x \in [0, 1]$, which can be formulated as (1).

$$
\text{Beta}(x|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}
$$

where $a$ and $b$ are determined parameters that are constants and more than 0, and $\Gamma(x)$ and $B(x, y)$ is Gamma function and Beta
Fig. 1. Schematic diagram of the proposed CMT, which mainly consists of a CXR image feature fusion augmentation (FFA) module, CXR image feature extractor (CIFE), and Multilevel Multi-Head Self-Attention Transformer (MMSA Transformer). Note that the Shared phase means the training and inference phase.

Algorithm 1: CXR Image Feature Fusion Augmentation (FFA) Algorithm.

**Input:** Two random CXR Images matrix $I_1, I_2$, patch size $p$, random coefficient $\alpha$ less than 1, and positional parameter $m, n$ for the feature fusion augmentation system $FFA = (I_1, I_2, p, \alpha, m, n)$.

**Output:** Generated image $\hat{I}$.

1: $z' = BRS(\alpha = \alpha, \beta = \alpha); z'_d = 1 - z'$; // Step 1: Random Beta Sampling (RBS) to generate fixed weights $z'$ and $z'_d$
2: Define two all-zero matrix $M$ and $M_d$;
3: Define the $h$ and $w$ denotes the height and weight of the image matrix $I_1$ and $I_2$.
4: for ($i = 0; i \leq h$; $i + +$) do
5: for ($j = 0; j \leq w; j + +$) do
6: if $m * \frac{h}{p} < i < (m + 1) * \frac{h}{p}$ and $n * \frac{w}{p} < j < (n + 1) * \frac{w}{p}$ then
7: $\hat{z} = BRS(\alpha = \alpha, \beta = \alpha); \hat{z}' = 1 - \hat{z}'$; // Step 2: Generate random weights using random beta sampling in the loop
8: $M[i, j] = \hat{z}; M_d[i, j] = \hat{z}'$
9: else
10: $M[i, j] = z'; M_d[i, j] = z'_d$
11: $\hat{I}[i, j] = I_1[i, j] * M[i, j] + I_2[i, j] * M_d[i, j]$ \// Step 3: The weight matrix and the image matrix do Hadamard product to generate a new image matrix $\hat{I}$
12: end
13: return $\hat{I}$

B. CXR Image Feature Extractor (CIFE)

The architecture of CIFE is shown in Fig. 3. To extract global and local features from CXR images more comprehensively

function that can be formulated as (2) and (3).

\[
\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt, \quad (2)
\]

\[
B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt. \quad (3)
\]

RBS, on the other hand, randomly selects one of the beta divisions calculated from the pre-determined $a, b$ as the weights. These weights are combined as the weight matrices $M \in \mathbb{R}^{h \times w}$ and $M_d \in \mathbb{R}^{h \times w}$, respectively. It is worth noting that the weights of the points in the $\frac{h}{p} \times \frac{w}{p}$ region of the matrix are inconsistent. In contrast, the other regions’ weights are consistent, allowing the generated images to ensure the presence of the necessary features. Finally, the two images and the weight matrices are Hadamard-product and summed, as shown in Line 11 of Algorithm 1, to obtain the generated image $\hat{I} \in \mathbb{R}^{h \times w}$.
while maintaining high efficiency. The first four convolutional and pooling layers of pre-trained ResNet101 are selected as the feature extractor. The ResNet is pre-trained based on the ImageNet dataset containing several images of common objects and scenes in life, which cannot be adapted to the medical image classification problem due to the significant differences in the dataset. Therefore, we consider using only its shallow layers (i.e., the first four layers mentioned) with powerful feature extraction capabilities for CXR images and discard the deep layers that are usually classifiers. This means that we can maximize the role of pre-trained ResNet in medical image classification tasks. Given the input CXR image is \( I \in \mathbb{R}^{h \times w \times c} \), the feature maps extracted by CIFE is \( I_F \in \mathbb{R}^{h' \times w' \times c'} \). These feature maps are rich in deep lung features that were encoded as feature embedding \( E \in \mathbb{R}^{C' \times (h' \times w')} \) for input to the subsequent Transformer, which can better understand the relationship between lung features and disease labels.

### C. Multilevel Multi-Head Self-Attention Transformer (MMSA-Transformer)

The architecture of the MMSA Transformer is shown in Fig. 4. It mainly consists of a Multilevel Multi-Head Self-Attention (MMSA) used to emphasize the importance of information while ignoring invalid information. Moreover, a feed-forward network (FFN) is applied to understand further the relationship among task-related feature information by receiving the reconstructed information from the MMSA. FFN then outputs the corresponding probabilities of each label simultaneously rather than in isolation, and the process can be formulated as

\[
P = FFN(X) = \text{Sigmoid} \left( W_1 W_2 X + b \right),
\]

where

\[
\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}}.
\]
where the $X$ is the feature information, the $W^1$ and $W^2$ denotes the weight matrix of the first the second layer of FFN, respectively, the $Sigmoid$ is the activation function.

Traditionally, the Transformer-based models rely heavily on the Multi-Head Self-Attention (MHSA) to capture the critical information on data, as shown in Left of Fig. 5. However, the high computational complexity of this mechanism inevitably increases the model’s inference time while obtaining excellent performance. To reduce the computational complexity of the attention mechanism, so that satisfy the high-speed recognition requirement while keeping outstanding accuracy, we design MMSA, an efficiently and practical self-attention mechanism, as shown in Right of Fig. 5. Next, we first briefly introduce the conventional MHSA, while our proposed MMSA is presented and analyzed afterward.

1) Multi-Head Self-Attention (MHSA): Given the feature map $I_F \in \mathbb{R}^{c \times h \times w}$, where $c$, $h$, and $w$ denotes the channels, height, and width, respectively. The $I_F$ was reshaped before fed into the attention module to define and obtain the query $Q$, key $K$, and value $V$, these can be formulated as follow:

$$I_F \in \mathbb{R}^{c \times h \times w} \xrightarrow{\text{reshape}} I_F \in \mathbb{R}^{c \times (h \times w)},$$

$$Q = W^q I_F, K = W^k I_F, V = W^v I_F,$$

where the $W^q$, $W^k$, and $W^v$ denotes the trainable weight matrices for $Q$, $K$, and $V$, respectively. The attention matrix can be obtained from $Q$ and $K$, that can be presented as

$$\text{Att} = Softmax \left( \frac{QK^T}{\sqrt{d}} \right),$$

where the $\sqrt{d}$ is an approximate normalization constant. Then the output obtain from $Q$, $K$, and $V$ can be expressed as:

$$A = \text{Att} V.$$  

The matrix product $QK^T$ first computes the similarity between each pair of tokens. Then, based on the similarity, each new token is derived by the combination of all tokens. Finally, the attention’s output can be computed by a residual connection, which can be formulated as

$$I_F \xrightarrow{\text{reshape}} I_F \in \mathbb{R}^{c \times h \times w},$$

in which the $W^p$ is a trainable matrix that shape is $W^p \in \mathbb{R}^{c \times c}$ for feature projection. Given that this attention module contains $N$ heads internally, i.e., $N$-heads Self-Attention, its final attention output is then summed over the channel dimensions according to the results of (9).

The computation complexity of Multi-Head Self-Attention can be inferred according the (4)–(9):

$$\Omega(\text{MHSA}) = 2ch^2 w^2 + 3hwc^2.$$  

The space complexity includes the term $O(h^2 w^2)$, which means that the memory consumption could become very large when the input is high-resolution ($h \times w$). This limits the application of the MHSA module to CXR image recognition tasks that place demands on fast training and inference. Based on this problem, we propose MMSA, a multilevel multi-head self-attention mechanism with low computational complexity. It reduces the complexity while obtaining an excellent information understanding capability under different levels without compromising performance.

2) Multilevel Multi-Head Self-Attention (MMSA): The architecture of the proposed MMSA is shown in the Right of Fig. 5. Compared with the conventional MHSA, MMSA does not directly calculate the attention of the whole input image but pre-divides the input into different levels, which means that only a limited number of tokens need to be calculated in the attention calculation of each level. This form of attention calculating can encourage the model to comprehensively focus on more fine-grind feature regions and capture the relation of global-local information instead of paying all attention to overall input. Given the input feature map $I_F \in \mathbb{R}^{c \times h \times w}$, it will be divided into several small grid according to its original resolution before calculating the attention. The small grid can be considered as $G \in \mathbb{R}^{g \times g}$, and the input feature map be reshaped according to the small grid as:

$$I_F \in \mathbb{R}^{c \times h \times w} \rightarrow I_{F1} \in \mathbb{R}^{c \times \left(\frac{h}{g}\times\frac{w}{g}\right)}$$

$$\rightarrow I_{F1} \in \mathbb{R}^{c \times \left(\frac{g}{7} \times \frac{g}{3}\right) \times (g \times g)}.$$
The query, key, and value of this level can be obtained by

\[
Q_1 = I_{F1} W_{q1}, K_1 = I_{F1} W_{k1}, V_1 = I_{F1} W_{v1},
\]

where \(W_{q1}, W_{k1},\) and \(W_{v1}\) are trainable weight matrices that is \(W_{q1} \in \mathbb{R}^{c \times c}, W_{k1} \in \mathbb{R}^{c \times c},\) and \(W_{v1} \in \mathbb{R}^{c \times c},\) respectively. Then, the local attention of this level can be generated by

\[
Att_1 = \text{Softmax}(Q_1 K_1^T / \sqrt{d}) V_1,
\]

\((15)\)

\(Att_1\) is reshaped to the shape of the input image \(I_F\) before computing the next level of local attention, and a residual connection applied between this local attention and the input image, which can be formulated as

\[
\begin{align*}
&Att_1 \in \mathbb{R}^{c \times (\frac{h}{g} \times \frac{w}{g}) \times (g \times g)} \rightarrow Att_1 \in \mathbb{R}^{c \times (h \times g) \times (g \times g)} \rightarrow Att_1 \in \mathbb{R}^{c \times h \times w}, \\
&Att_1 = Att_1 + I_F.
\end{align*}
\]

\((16)\)

\((17)\)

Compared with the conventional Multi-Head Self-Attention that operates directly on \(h\) or \(w, Att_1\) directly performs the attention operation on the small grid \(g \times g\), which reduces its space complexity significantly.

After entering the next computational level, \(Att_1\) is downsampled to \(g'\) by pooling. Consistent with the calculation of the first level, each small grid \(G' \in \mathbb{R}^{g' \times g'}\) will be regarded as a token in the calculation, and the process can be expressed as

\[
\begin{align*}
&Att_1 = \frac{1}{2} \left[ \alpha \times \text{MaxPool}_{g'}(Att_1) + \beta \times \text{AvePool}_{g'}(Att_1) \right], \\
&\text{Here, the } \text{MaxPool}_{g'}(x) \text{ and } \text{AvePool}_{g'}(x) \text{ denotes to downsample the input feature map by } g' \text{ times using max pooling and average pooling with the filter size and stride of } g', \text{ respectively, and the } \alpha \text{ and } \beta \text{ are control coefficients, they are both positive greater than 0 and add to 1. From this, it can be obtained that the } \\
&\text{Att}_1' \text{ after downsampling is } \text{Att}_1' \in \mathbb{R}^{c \times \frac{h}{g'} \times \frac{w}{g'}}. \text{ Furthermore, the token size } (\frac{h}{g} \times \frac{w}{g}) \text{ can be obtained form the reshaped } \text{Att}_1' \text{ as:} \\
&\text{Att}_1' \in \mathbb{R}^{c \times \frac{h}{g} \times \frac{w}{g} \times \frac{g}{g'}} \rightarrow \text{Att}_1' \in \mathbb{R}^{c \times (\frac{h}{g'} \times \frac{w}{g')}}, \\
&\text{The query, key, and value of this level can be computed by } \text{Att}_1' \text{ as:} \\
&Q_2 = \text{Att}_1' W_{q2}, K_2 = \text{Att}_1' W_{k2}, V_2 = \text{Att}_1' W_{v2}, \\
&\text{where } W_{q2}, W_{k2}, \text{ and } W_{v2} \text{ are trainable weight matrices that is } W_{q2} \in \mathbb{R}^{c \times c}, W_{k2} \in \mathbb{R}^{c \times c}, \text{ and } W_{v2} \in \mathbb{R}^{c \times c}, \text{ respectively. Next, the local attention } \text{Att}_2 \text{ can be obtained by} \\
&\text{Att}_2 = \text{Softmax}(Q_2 K_2^T / \sqrt{d}) V_2,
\end{align*}
\]

\((18)\)

\((19)\)

\((20)\)

\((21)\)

The final output of MMSA can be computed by the reshaping operation and a residual connection:

\[
\text{MMSA}(I_F) = (\text{Att}_1 + \text{Att}_2) W^m W^n + I_F,
\]

\((23)\)

where \(W^m, W^n\) are trainable weight matrices. The proposed MMSA has strong local and global feature modeling capabilities by fusing different levels of attention outputs. Next, the computational complexity of MMSA can be computed by

\[
\Omega(\text{MMSA}) = chw \left( 2 g'^2 + 4c \right) + \frac{2chw}{g'^2}(c + hw),
\]

\((24)\)

which means that the computation complexity from \(O(h^2 w^2)\) reduce to \(O \left( hwg'^2 + \frac{h^2 w^2}{g'^2} \right).\)

**D. Computational Complexity Analysis**

Eqs. (12) and (24) show the computational complexity of the conventional MHSA and our MMSA, respectively. Suppose a CXR image is \(360 \times 360 \times 3,\) and \(g'\) in MMSA is 2, the computational parameters of conventional MHSA and our MMSA are:

\[
P_{\text{MHSA}} = 2 \times 3 \times 360^4 + 3 \times 360^2 \times 3^2 = 100, 780, 459, 200,
\]

\[
P_{\text{MMSA}}^{g'=2} = 3 \times 360^2 \times (2^3 + 4 \times 3) + \frac{2 \times 3 \times 360^2}{2^2} \times (3 + 360^2) = 25, 202, 599, 200,
\]

\[
P_{\text{MMSA}}^{g'=5} = 3 \times 360^2 \times (5^3 + 4 \times 3) + \frac{2 \times 3 \times 360^2}{5^2} \times (3 + 360^2) = 4, 084, 437, 312,
\]

\[
P_{\text{MMSA}}^{g'=10} = 3 \times 360^2 \times (10^3 + 4 \times 3) + \frac{2 \times 3 \times 360^2}{10^2} \times (3 + 360^2) = 1, 401, 258, 528.
\]

\((25)\)

This means that the proposed MMSA pays only a quarter of the computational cost of the conventional MHSA when \(g' = 2.\) In addition, the \(P_{\text{MMSA}}\) is far smaller than the \(P_{\text{MHSA}}.\) The relationship between the conventional MHSA and our MMSA with different parameters \(g'\) can be formulated as:

\[
P_{\text{MHSA}} \approx 4 \times P_{\text{MMSA}}^{g'=2} \approx 25 \times P_{\text{MMSA}}^{g'=5} \approx 72 \times P_{\text{MMSA}}^{g'=10}.
\]

According to (26), when \(g' = 2,\) the conventional MHSA has at least four times more parameter operations than the MHS. In addition, when \(g' = 5, 10,\) this number reaches a staggering 25 and 72 times.

**IV. Experiments**

This experiments section shows the details of dataset, the experimental setup, evaluation metrics, and analyzes and discusses the experimental results, and model interpretability exploration.

**A. Dataset**

The pneumonia dataset used in our work consists of CXR images corresponding to different symptoms: Normal, COVID-19, Lung Opacity, and Viral pneumonia [48, 49]. The dataset has 21,175 CXR images with 299 × 299 resolution. These CXR images were collected from five different medical institutions such as Banco digital de Imagen Médica de la Comunidad Valenciana (BIMCV), the German Medical School (GMS), the Società Italiana di Radiologia Medica e Interventistica (SIRM), Radiological Society of North America (RSNA), and Guangzhou Women’s and Children’s Medical Center (GWCMC). The detailed information of the dataset is presented in Table I, and some example images are shown in Fig. 6.
Table I

| Disease          | Data Source | Amount | Resolution     | Total |
|------------------|-------------|--------|----------------|-------|
| COVID-19         | BMVC        | 2473   | 299 x 299      | 3615  |
| Normal           | RSNA        | 8851   | 299 x 299      | 10192 |
| Lung Opacity     | RSNA        | 6012   | 299 x 299      | 6023  |
| Viral pneumonia  | GWCMC       | 1345   | 299 x 299      | 1345  |

Fig. 6. Examples of CXR image datasets, from top to bottom, are COVID-19 positive, Normal, Lung Opacity, and Viral pneumonia.

B. Experimental Setup

All models and algorithms in this work are built on the Pytorch. The first four layers of ResNet101, pre-trained by ImageNet, are considered the CIFE backbone network for accelerating the convergence of the overall model. The MMSA Transformer is trained from scratch by receiving feature embeddings from CIFE to enter the stage. In the pre-training FFA module, the patch size (p) is set to 2, α to 0.4, and the position parameters m, n are set to the permutations obtained from (m = 0, 1; n = 0, 1), such as m = 1, n = 1 or m = 0, n = 1, etc. During training, N in the MMSA Transformer structure is set to N = 4 and g and g’ in the MMSA are set to [2, 2], and α and β in (18) are set to 0.3 and 0.7, respectively. The first and second momentum is 0.9, and 0.999 of Adam [50] were considered as optimizers of the network to minimize the loss function with a learning rate of $10^{-5}$. During the training process, the CXR image dataset is randomly disrupted and then divided into training, test, and validation sets in the ratio of 8 (16,940) : 1 (2,117) : 1 (2,117), without any overlap. Note that all results generated by the above model are based on twenty iterations of model training, and the early stopping strategy is adopted. The learning rate Warmup mechanism [27] and Cosine Annealing strategy [51] is employed to avoid model oscillation and accelerate convergence, which can be expressed as

$$\eta_t = \eta_{\text{min}} + \frac{1}{2} (\eta_{\text{max}} - \eta_{\text{min}}) \left( 1 + \cos \left( \frac{T_{\text{cur}}}{T_i} \pi \right) \right), \quad (27)$$

where $\eta_{\text{max}}$ and $\eta_{\text{min}}$ is the max and min learning rate that define the range of learning rate, $i$ denotes the index of iteration times, and $T_{\text{cur}}$ and $T_i$ is the current iteration times and total iteration times. The loss function is BCE Loss, which can be formulated as

$$L(o, t) = - \frac{1}{n} \sum_{i} (t[i] * \log(o[i])) + (1 - t[i]) * \log(1 - o[i])), \quad (28)$$

Here, the $o$ and $t$ are the output and actual target, respectively. In addition, in each FFN within the MMSA-Transformer, a Dropout operation with probability 0.2 is applied to avoid overfitting the model. All images are adjusted to 360 x 360 and converted to RGB images before input into the model, and each mini-batch contains 16 randomly scrambled images.

All the algorithms and models run on the Tesla V100 GPU (16 GB), Intel Xeon W-2245 (3.90 GHz).

C. Evaluation Metrics

The recall ($R$) and precision ($P$) of each disease label were selected as evaluation metrics. The results were first classified into the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) according to the classification, as shown in Table II. The $R$ and $P$ can be formulated as

$$R = \frac{TP}{TP + FN}, \quad P = \frac{TP}{TP + FP}, \quad (29)$$

Then, the average class precision (CP) and the average class recall (CR), which denotes the average of each class precision and recall, are calculated by

$$CR = \frac{\sum_{n=1}^{N} \sum_{i=1}^{K} R_{n,i}}{NK}, \quad CP = \frac{\sum_{n=1}^{N} \sum_{i=1}^{K} P_{n,i}}{NK}, \quad (30)$$

where $N$ and $K$ is the scale of the test dataset and the number of disease labels, respectively. Similarity, the overall precision (OP) and overall recall (OR) that means the actual prediction of all images in all disease labels, respectively, can be formulated as

$$OR = \frac{\sum_{n=1}^{N} \sum_{i=1}^{K} f(p_{n,i}, \hat{p}_{n,i})}{\sum_{n=1}^{N} \sum_{i=1}^{K} p_{n,i}}, \quad OP = \frac{\sum_{n=1}^{N} \sum_{i=1}^{K} f(p_{n,i}, \hat{p}_{n,i})}{NK}, \quad (31)$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE III

| Model                | COVID-19 | Normal | Lung Opacity | Viral pneumonia | CP  | CR  | CPI | OP  | OR  | OPI |
|----------------------|----------|--------|--------------|-----------------|-----|-----|-----|-----|-----|-----|
| MobileNet [52]       | 0.8934   | 0.9426 | 0.8833       |                  |     |     |     |     |     |     |
| AlexNet [29]         | 0.9727   | 0.8802 | 0.9729       |                  |     |     |     |     |     |     |
| ResNet [27]          | 0.9211   | 0.8366 | 0.9148       |                  |     |     |     |     |     |     |
| VGGNet [28]          | 0.9466   | 0.8558 | 0.8537       |                  |     |     |     |     |     |     |
| ResNetXt [53]        | 0.9418   | 0.8885 | 0.8409       |                  |     |     |     |     |     |     |
| Wide ResNet [54]     | 0.9190   | 0.9037 | 0.8806       |                  |     |     |     |     |     |     |
| Inception V3 [35]    | 0.9552   | 0.8912 | 0.9018       |                  |     |     |     |     |     |     |
| ConvNetXt [56]       | 0.9677   | 0.9090 | 0.8563       |                  |     |     |     |     |     |     |
| RegNet [37]          | 0.9721   | 0.9624 | 0.8515       |                  |     |     |     |     |     |     |
| CNN-LSTM             | 0.8121   | 0.9088 | 0.9038       |                  |     |     |     |     |     |     |
| CNN-ConvLSTM         | 0.8702   | 0.9058 | 0.9431       |                  |     |     |     |     |     |     |
| CNN-Transformer      | 0.9806   | 0.8613 | 0.8992       |                  |     |     |     |     |     |     |
| CMT (Inception V3)   | 0.9841   | 0.8662 | 0.9293       |                  |     |     |     |     |     |     |
| CMT (VGGNet)         | 0.9624   | 0.9002 | 0.9700       |                  |     |     |     |     |     |     |
| CMT (ResNetXt)       | 0.9787   | 0.9628 | 0.9203       |                  |     |     |     |     |     |     |
| CMT (Wide ResNet)    | 0.9625   | 0.9029 | 0.9497       |                  |     |     |     |     |     |     |
| CMT (ConvNetXt)      | 0.9724   | 0.9212 | 0.9421       |                  |     |     |     |     |     |     |
| CMT (RegNet)         | 0.9797   | 0.9620 | 0.9354       |                  |     |     |     |     |     |     |
| CMT (Ours, ResNet101) | 0.9972  | 0.9628 | 0.9714       |                  |     |     |     |     |     |     |

TABLE IV

| Model                 | Attention | Param# | Training | Inference |
|-----------------------|-----------|--------|----------|-----------|
| CNN-Transformer       | MHSAs     | 195,728 M | 738 s    | 102 s     |
| CNN-Transformer       | W-MSAs    | 195,728 M | 584 s    | 86 s      |
| CNN-Transformer       | PS        | 195,728 M | 542 s    | 80 s      |
| CNN-Transformer       | GS        | 195,728 M | 487 s    | 75 s      |
| CMT (Ours)            | MMSAs     | 195,728 M | 456 s    | 72 s      |

D. Baseline

The baseline in our work is divided into two distinct parts: Nonlinear Classifiers and Deep Neural Networks.

1) Nonlinear Classifiers: Three representative nonlinear classifiers, such as Decision Tree (DT), Support Vector Machine (SVM), and Multilayer Perceptron (MLP), are selected as baselines of the nonlinear classifiers part. The Principal Component Analysis (PCA) reduces input CXR images’ dimensions before these classifiers operate. Moreover, the kernel function of the SVM is RBF, and all hidden layers’ unit is 512 in MLP.

2) Deep Neural Networks: Currently popular CNN-based models such as MobileNet [52], AlexNet [29], VGGNet [28], ResNet [27], ResNetXt [53], Wide ResNet [54], Inception V3 [35], ConvNetXt [56], and RegNe [57], pre-trained by ImageNet [58], were selected as baselines for comparison. Specialy, there is VGGNet16, ResNet101, ResNetXt101-32×8d, Wide ResNet101, Inception V3, ConvNetXt-base, and RegNetY-400MF. In addition, CNN-RNN-based models (e.g., CNN-LSTM and CNN-ConvLSTM) are considered baselines in which the CNN is ResNet152 pre-trained by ImageNet, and the LSTM/ConvLSTM is a four-layers architecture. Note that the final classification layer of these CNN-based backbones was changed to four when those acts separately to fit the recognition task in our work. Moreover, the CNN-Transformer with the conventional MHSA is considered the baseline to demonstrate the advantage of MMSA, the number of attention heads and Transformer layers is consistent with the proposed CMT.

E. Experimental Results

Three machine learning-based nonlinear classifiers and several deep neural network-based classifiers were utilized to validate the superiority of our proposed CMT. The experimental results section is illustrated in three parts: Nonlinear Classifiers, Deep Neural Network-based Classifiers, and Training and Inference Time.

1) Comparison with Nonlinear Classifiers: The performance comparison of the proposed CMT with nonlinear classifiers is presented in Table II. The proposed CMT outperforms a series of nonlinear classifiers that act separately to achieve recognition. Moreover, to validate the superiority of CMT in a more fair way, the pre-trained ResNet101 is adopted as the dimensional
RESULTS OF THE ABLATION EXPERIMENTS, WHICH CONTAIN TEST RESULTS FOR DIFFERENT PATCH SIZES (p) IN FFA AND VARIOUS COMBINATIONS OF LEVELS (g, f) IN THE PROPOSED MMSA. (√) MEANS THAT THE MODEL CARRIES A SPECIFIC MODULE (FFA OR MMSA), WHILE (×) IMPLIES THAT THE MODULE DOES NOT EXIST. (+) REPRESENTS THE ABSENCE OF THIS CONFIGURATION, (∗) MMSA IS ONLY WHEN CMT CONTAINS MMSA INTERNALLY BY DEFAULT, THE RESULTS FOR EACH CLASS REPRESENT THE RECOGNITION ACCURACY, WITH < 100% FOR CP, CR, CF1, OP, OR, AND OF1 RELATIVE TO THE INITIAL RESULTS.

| Model       | PPA | PPA Conguration | MMSA | MMSA Conguration | COVID-19 | Normal | Long Opyety | Viral prevalence | CP | CR | CF1 | OP | OR | OF1 |
|-------------|-----|-----------------|------|------------------|---------|--------|-------------|-----------------|----|----|-----|----|----|-----|
| CNN-Transformer | √   | -               | -    | -                | 0.9306  | 0.8611 | 0.8912     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             |    | -               | -    | -                | 0.9306  | 0.8611 | 0.8912     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             | √   | -               | -    | -                | 0.9306  | 0.8611 | 0.8912     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             |    | -               | -    | -                | 0.9306  | 0.8611 | 0.8912     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             | √   | -               | -    | -                | 0.9306  | 0.8611 | 0.8912     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             |    | -               | -    | -                | 0.9306  | 0.8611 | 0.8912     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
| CMT         | √   | -               | -    | -                | 0.9779  | 0.9530 | 0.9812     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             |    | -               | -    | -                | 0.9779  | 0.9530 | 0.9812     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             | √   | -               | -    | -                | 0.9779  | 0.9530 | 0.9812     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             |    | -               | -    | -                | 0.9779  | 0.9530 | 0.9812     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             | √   | -               | -    | -                | 0.9779  | 0.9530 | 0.9812     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             |    | -               | -    | -                | 0.9779  | 0.9530 | 0.9812     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
| ResNet      | √   | -               | -    | -                | 0.9779  | 0.9530 | 0.9812     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             |    | -               | -    | -                | 0.9779  | 0.9530 | 0.9812     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             | √   | -               | -    | -                | 0.9779  | 0.9530 | 0.9812     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             |    | -               | -    | -                | 0.9779  | 0.9530 | 0.9812     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             | √   | -               | -    | -                | 0.9779  | 0.9530 | 0.9812     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |
|             |    | -               | -    | -                | 0.9779  | 0.9530 | 0.9812     | 0.3368           | 91.1 | 93.5 | 91.5 | 93.8 | 96.6 | 90.0 |

2) Comparison with Deep Neural Network-based Classifiers: The experiment results of performance comparison between CMT and deep neural network-based baseline are shown in Table III. We can observe that i) The overall recognition accuracy and generalization performance of the proposed CMT outperforms all baselines; ii) The F1 scores (CF1 and OF1) of the CMT are 6.4% and 7.9% higher than the CNN-Transformer, respectively, which means that the proposed MMSA mechanism has a significant advantage over conventional MHSA. iii) our proposed CMT, with ResNet101 as the backbone on default, has better recognition performance than its variants (e.g., the backbone is replaced with Inception V3/VGGNet/ResNetXt/Wide Resnet/ConvNeXt/RegNet), in terms of recognition accuracy and evaluation metrics’ scores. The proposed CMT’s average recognition accuracy is 2.4%, 3.1%, 1.3%, 2.1%, 3.3%, 2.4% higher than the ones that backbone is Inception V3/VGGNet/ResNetXt/Wide Resnet/ConvNeXt/RegNet. The average F1 score is 5.1%, 4.1%, 7.1%, 5.8%, 5.4%, 4.0% higher than the CMT with the Inception V3/VGGNet/ResNetXt/Wide Resnet/ConvNeXt/RegNet back- bone. In addition, the performance of CMT’s variants with different backbones is far beyond these pre-trained backbone act independently, which means the proposed MMSA Transformer within CMT can enhance CNN-based models. The superiority of MMSA and CMT, the advantage of our MMSA over MHSA, and the necessity of ResNet101 as the default backbone is validated.

3) Training and Inference Time: In addition, to verify the superiority of the proposed MMSA mechanism in terms of computational complexity, we experiment with comparing training and inference timing between the CNN-Transformer with conventional MHSA, Windows Multi-Head Self Attention (W-MSA) [41], Pale-Shaped Attention (PS) [39], and Global Self-Attention (GSA) [40] and CMT (with proposed MMSA). The result is shown in Table IV. When keeping the same model parameters, the proposed CMT consumes significantly less time for computing than the CNN-Transformer, is 38.2%
28.1%/18.9%/6.8% and 29.4%/19.4%/11.1%/4.2% faster than the CNN-Transformer with MHS/W-MSA/PS/GSA in terms of training and inference, respectively. This means that the proposed MMSA mechanism reduces the computational complexity while maintaining an excellent recognition performance.

F. Ablation Experiments

Extensive ablation studies are conducted to validate the effectiveness and necessity of the proposed FFA and MMSA. The model’s hyperparameters setting is kept the same as the previous experiments in the ablation experiments. In addition, the results below are all calculated based on the effects of twenty iterations of these models. The proposed FFA’s effectiveness is validated by adding it to CNN-Transformer and ResNet and using different parameters $p$ to illustrate the effect of FFA’s weight matrix scales. The results of the ablation experiments are shown in Table V, and to more obviously represent the impact of each module on the recognition performance, the amount of variation of metric was shown in Table VI. The absolute value of the variation bigger, the recognition performance better. We can observe that i) The average F1 score of CNN-Transformer with FFA is higher 0.9, 1.3, 2.0 than the original one when the parameters $p = 2, 4, 8$, respectively. Similarly, the ResNet with FFA compare to the original one has the same trend as CNN-Transformer. The improvements in the average F1 score are 2.35, 3.65, 7.2, respectively. This means that the FFA can boost the model’s recognition. The $p$ higher, the enhancement more significant. ii) When considering the effect of MMSA independently rather than the combination of FFA and MMSA, the recognition performance of CMT still outperforms CNN-Transformer (average F1 score is 4.5, 2.85, 1.2 higher under $g'/g = 2, 5, 10$). In addition, the smaller the MMSA parameters $g, g'$ (the finer the level of MMSA calculation), the better the CMT’s performance. The superiority and effectiveness of the proposed MMSA over the conventional MHS are validated. iii) Compared to the model without FFA and MMSA, the state-of-the-art (SOTA) performance was obtained for the whole CMT, and the best results were obtained for the $p = 8$, $g = 2, g' = 2$ configuration, the average F1 score is 7.5 higher than the conventional MHS. All these imply that the proposed FFA can effectively eliminate the negative impact of the scarcity of medical image data and improve recognition accuracy. The absence of either the FFA or MMSA module negatively affects CMT, demonstrating their effectiveness and necessity.

G. Interpretability Exploration

DL models, long regarded as ‘black boxes’ due to their feature extraction and attention mechanism, are unclear and lack credibility in real medical applications. To explore the interpretability of the proposed CMT, we use Grad-CAM [59] to extract weights from the model in CXR image examples with lung disease and visualize the weighing matrix and feature activation maps for presenting the region of the model’s attention. The visualization results are shown in Fig. 7. The region covered by darker color represents that the proposed CMT pays more attention to it and contains lots of task-related information. The proposed CMT focuses on the lung region, which has more influence on the recognition results than other regions. This demonstrates that our proposed CMT can focus more on the effective information across different regions in CXR images and ignore other noisy and minimally important features to obtain excellent recognition performance.

V. CONCLUSION, LIMITATION, AND FUTURE WORK

In conclusion, we proposed a novel pneumonia recognition framework with interpretability and high efficiency for providing reliable, accurate analytic support to computer-aided medical clinic diagnosing systems. The proposed CNN-MMSA-Transformer (CMT) consists of CNN, considered a feature extractor, and a Transformer with low computational complexity self-attention mechanism. Moreover, a low-cost CXR image data augmentation strategy has been introduced to reduce the negative impact on recognition accuracy because of the scarcity of medical image data. Experiments on the CXR image dataset fused from five distinct medical institutions have demonstrated the effectiveness of the proposed recognition framework. In addition, both the necessity and efficacy of the novel self-attention mechanism and data augmentation have been validated in further ablation studies.

Our work also has some limitations. First, as in most medical image recognition works that adopt the data-centric learning strategy, our works also run on the fusing CXR images dataset from different institutions. However, data interaction across various medical institutions is sometimes limited due to data exposure, and privacy concerns remain. For example, some hospital doesn’t share their data with others due to confidentiality. Secondly, it still relies on the per-labeled data that requires extensive expert experience intervention to understand knowledge from CXR images during training. Finally, the proposed model still needs to consider all parameters’ computations, although it is a high-performance and low computational complexity solution, which sometimes is unfriendly to satisfy the low-resourced constraint devices in hospitals.

In future work, aiming to solve the above three limitations, we will develop a general medical image recognition framework across medical institutions without direct data interaction to ensure data privacy. In the meantime, we will develop a new learning mechanism to obtain sufficiently superior recognition performance in understanding a few labeled CXR image datasets while reducing the model’s parameter utilization.

REFERENCES

[1] D. Cucinotta and M. Vanelli, “WHO declares COVID-19 a pandemic,” Acta Bio Medica: Aetnei Parmensis, vol. 91, no. 1, pp. 157–160, 2020.
[2] S. A. Lauer et al., “The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application,” Ann. Intern. Med., vol. 172, no. 9, pp. 577–582, 2020.
[3] J. Phua et al., “Intensive care management of coronavirus disease 2019 (COVID-19): Challenges and recommendations,” The Lancet Respir. Med., vol. 8, no. 5, pp. 506–517, 2020.
[4] C. Ieracitano et al., “A fuzzy-enhanced deep learning approach for early detection of COVID-19 pneumonia from portable chest X-ray images,” Neurocomputing, vol. 481, pp. 202–215, 2022.
[5] H. Y. F. Wong et al., “Frequency and distribution of chest radiographic findings in patients positive for COVID-19,” Radiology, vol. 296, no. 2, pp. E72–E78, 2020.
[6] L. Wang, Z. Q. Lin, and A. Wong, “Covid-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images,” Sci. Rep., vol. 10, no. 1, pp. 1–12, 2020.
[7] M. Heidari, S. Mirniaharikandehei, A. Z. Khuzani, G. Danala, Y. Qiu, and B. Zheng, “Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms,” Int. J. Med. Inform., vol. 144, 2020, Art. no. 104284.
[8] I. D. Apostolopoulos and T. A. Mapesiana, "COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," Phys. Eng. Sci. Med., vol. 43, no. 2, pp. 635–640, 2020.

[9] M. Z. Islam, M. M. Islam, and A. Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images," Inform. Med. unsealed, vol. 20, 2020, Art. no. 100412.

[10] R. Karthik, R. Menaka, and M. Harirahan, “Learning distinctive filters for COVID-19 detection from chest X-ray using shuffled residual CNN,” Appl. Soft. Comput., vol. 99, 2021, Art. no. 106744.

[11] A. I. Khan, J. L. Shah, and M. M. Bhat, "Coronet: A deep neural network for detection and diagnosis of COVID-19 from chest X-ray images," Comput. Methods Prog. Biomed., vol. 196, 2020, Art. no. 105581.

[12] D. Ezzat, A. E. Hassanien, and H. A. Ella, “An optimized deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization,” Appl. Soft. Comput., vol. 98, 2021, Art. no. 106742.

[13] M. Umer, I. Ashraf, S. Ullah, A. Mehmood, and G. S. Choi, “COVINet: A convolutional neural network approach for predicting COVID-19 from chest X-ray images,” J. Ambient. Intell. Humanized Comput., vol. 13, no. 1, pp. 535–547, 2022.

[14] A. Jacobi, M. Chung, A. Bernheim, and C. Eber, “Portable chest X-ray for COVID-19 classification,” In Proc. Int. Conf. Learn. Representations, 2021.

[15] V. Mnih et al., “Recurrent models of visual attention,” in Proc. 27th Int. Neural Inform. Process. Syst., 2014, pp. 2204–2212.

[16] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7794–7803.

[17] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7132–7141.

[18] J. Fu et al., “Dual attention network for scene segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3146–3154.

[19] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, “OCNet: Object context for semantic segmentation,” Int. J. Comput. Vis., vol. 129, no. 8, pp. 2375–2398, 2022.

[20] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “CCNet: Criss-cross attention for semantic segmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 603–612.

[21] A. Ng et al., “Sparse autoencoder,” CS294 A Lecture notes, vol. 72, pp. 1–19, 2011.

[22] I. Goodfellow et al., “Generative adversarial networks,” Commun. ACM, vol. 63, no. 11, pp. 139–144, 2020.

[23] G. Mariariu et al., “BAGAN: Data augmentation with balancing GAN,” in Proc. Int. Conf. Mach. Learn., 2018.

[24] W.-H. Hsu, Y. Zhang, and J. Glass, “Unsupervised domain adaptation for COVID-19 detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 16–25.

[25] T. Liu, E. Siegel, and D. Shen, “Deep learning and medical image analysis,” in Proc. Int. Conf. Learn. Representations, 2019.

[26] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 129, pp. 7132–7141, 2021.

[27] A. G. Howard et al., “Mobilenets: Efficient convolutional neural networks for small-scale image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 232, 2017, pp. 1–11, 2014.

[28] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inform. Process. Syst., 2017, vol. 30.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019.

[30] Z. Shen, I. Bello, R. Vemulapalli, X. Jia, and C.-H. Chen, “Global self-attention for image recognition,” in Proc. Adv. Neural Inform. Process. Syst., 2019, pp. 10012–10022.

[31] A. Farnani et al., “Attention is all you need,” in Proc. Adv. Neural Inform. Process. Syst., 2020, pp. 6824–6835.

[32] D. Zhou et al., “DeepViT: Towards deeper vision transformer,” arXiv:2103.11886, 2021.

[33] M. E. Chowdhury et al., “Can AI help in screening viral and COVID-19 pneumonia?” IEEE Access, vol. 8, pp. 132665–132676, 2020.

[34] J. Rahman et al., “Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images,” Comput. Biol. Med., vol. 132, 2021, Art. no. 104319.

[35] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980v1.

[36] I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” in Proc. Int. Conf. Learn. Representations, 2017.

[37] A. G. Howard et al., “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, arXiv:1704.04861.

[38] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1492–1500.

[39] S. Zagoruyko and N. Komodakis, “Wide residual networks,” 2016, arXiv:1605.07146.