Research Paper
https://doi.org/10.55525/tjst.1411197

PneumoNet: Automated Detection of Pneumonia using Deep Neural Networks from Chest X-Ray Images

Zehra KADIROGLU[1*], Erkan DENIZ[2], Mazhar KAYAOGLU[3], Hanifi GULDEMIR[4], Abdurrahman SENYIGIT[5], Abdulkadir SENGUR[6]

[1,2,4,6] Department of Electrical and Electronics Engineering, Faculty of Technology, Fırat University, Elazıg, Turkey
[3] Department of Informatics, Bingol University, Bingol, Turkey
[5] Department of Chest Diseases and Tuberculosis, Faculty of Medicine, Dicle University, Diyarbakır, Turkey
[*1] zehrakad@gmail.com, [2] edeniz@firat.edu.tr, [3] mkayaoglu@bingol.edu.tr, [4] hguldemir@firat.edu.tr,
[5] drasenyigit@gmail.com, [6] ksengur@firat.edu.tr

(Geliş/Received: 28/12/2023; Kabul/Accepted: 18/04/2024)

Abstract: Pneumonia is a dangerous disease that causes severe inflammation of the air sacs in the lungs. It is one of the infectious diseases with high morbidity and mortality in all age groups worldwide. Chest X-ray (CXR) is a diagnostic and imaging modality widely used in diagnosing pneumonia due to its low dose of ionizing radiation, low cost, and easy accessibility. Many deep learning methods have been proposed in various medical applications to assist clinicians in detecting and diagnosing pneumonia from CXR images. We have proposed a novel PneumoNet using a convolutional neural network (CNN) to detect pneumonia using CXR images accurately. Transformer-based deep learning methods, which have yielded high performance in natural language processing (NLP) problems, have recently attracted the attention of researchers. In this work, we have compared our results obtained using the CNN model with transformer-based architectures. These transformer architectures are vision transformer (ViT), gated multilayer perceptron (gMLP), MLP-mixer, and FNet. In this study, we have used the healthy and pneumonia CXR images from public and private databases to develop the model. Our developed PneumoNet model has yielded the highest accuracy of 96.50% and 94.29% for private and public databases, respectively, in detecting pneumonia accurately from healthy subjects.

Key words: Pneumonia detection, medical image classification, chest x-ray imaging, transformer, deep neural networks.

PneumoNet: Göğüs Röntgeni Görüntülerinden Derin Sinir Ağları Kullanarak Pnömoninin Otomatik Tespiti

Öz: Pnömoni, akciğerlerdeki hava keseciklerinin şiddetli iltihaplanmasına neden olan tehlikeli bir hastalıktır. Dünya genelinde tüm yaş gruplarında yüksek morbidite ve mortaliteye sahip bulaşıcı hastalıklardan biridir. Göğüs röntgeni (CXR), düşük iyonize radyasyon dozu, düşük maliyeti ve kolay erişilebilirliği nedeniyle pnömoni teşhisinde yaygın olarak kullanılan bir teşhis ve görüntüleme yöntemidir. Çeşitli tıbbi uygulamalarda klinisyenlere CXR görüntülerinden pnömoni tespit ve teşhisinde yardımcı olmak için birçok derin öğrenme yöntemi önerilmiştir. CXR görüntülerini kullanarak pnömoniyi doğru bir şekilde tespit etmek için evrişimsel sinir ağı (ESA) kullanan yeni bir PneumoNet önerilmiştir. Doğal dil işleme (NLP) problemlerinde yüksek performans sağlayan dönüştürücü tabanlı derin öğrenme yöntemleri son zamanlarda araştırmacıların ilgisini çekmektedir. Bu çalışmada, CNN modelini kullanarak elde ettiğimiz sonuçlar dönüştürücü tabanlı mimarilerle karşılaştırılmıştır. Bu dönüştürücü mimariler görüntü dönüştürücü (ViT), kapılı çok katmanlı algılayıcı (gMLP), MLP-mikser ve FNet'tir. Bu çalışmada, modeli geliştirmek için kamu ve özel veri tabanlarından sağlıklı ve pnömoni CXR görüntüleri kullanılmıştır. Geliştirdiğimiz PneumoNet modeli, sağlıklı bireylerden pnömoniyi doğru bir şekilde tespit etmede özel ve kamu veri tabanları için sırasıyla %96,50 ve %94,29'luk en yüksek doğruluğu sağlamıştır.

Anahtar kelimeler: Pnömoni tespiti, tıbbi görüntü sınıflandırma, göğüs röntgeni görüntüleme, transformatör, derin sinir ağları.

1. Introduction

Pneumonia is inflammation of the lung parenchyma caused by infective or non-infective causes [1]. People of all ages can have mild to severe diseases. The prevalence and mortality rates are highest in children younger than five years and persons over 70 years worldwide [2]. Depending on the severity of the disease: cough, shortness of breath, fever, sweating and chills, chest pain, nausea, vomiting and diarrhea can be seen. Factors such as unhealthy climatic conditions, pollution, passive smoking, and malnutrition increase the risk of this disease [3]. Other lung diseases also present similar symptoms like pneumonia. Hence it is challenging to diagnose it by symptoms and a medical examination [4]. Various medical imaging modalities and diagnostic tools are used in clinics to diagnose pneumonia. Due to its low cost, faster imaging time, and ease of access, CXR is the most common medical imaging technique used to detect pneumonia worldwide [5]. Chest radiographs contain

[*] Corresponding author: zehrakad@gmail.com. ORCID Number of authors: [1] 0000-0002-2696-8138, [2] 0000-0002-9048-6547, [3] 0000-0002-5807-9781, [4] 0000-0003-0491-8348, [5] 0000-0001-9603-2231, [6] 0000-0003-1614-2639

significant amount of information about the patient's condition. However, even for very experienced radiologists, distinguishing similar lesions or detecting and interpreting very indistinct nodules can be difficult [6]. Therefore, manual reading of CXR radiographs is tedious, time-consuming, and prone to human errors. Hence, accurate processing and analysis of medical images are crucial for a faster and more reliable diagnosis. In recent years, studies on the precise detection of diseases using deep learning techniques have been employed for computer-aided diagnosis (CAD) [3-6]. As a decision support system, deep learning-based CAD systems assist clinicians and radiologists in making accurate and rapid diagnoses, reducing the rate of misdiagnosis, and improving healthcare quality at lower costs [7]. Convolutional neural network (CNN), one of the deep learning architectures is considered as the most advanced architecture in image classification and computer vision problems, as it effectively extracts features from low to deep convolution layers in the network [8]. In further studies, researchers sought successful recognition and performance improvement without convolution functions in deep learning models for computer vision applications [9]. Transformer models, which are attention mechanism-based architectures and widely used in natural language processing (NLP), have been adapted to solve this problem, and remarkable results have been obtained [10, 11]. Vision transformer (ViT), the image model of the transformer architecture, represents an input image as a fixed-size array of image patches and predicts image class tags, similar to the sequence of word embedding used in NLP applications [12]. In the literature, pneumonia is currently detected using transformer-based architectures.

Ukwuoma et al. [15] designed a hybrid deep learning framework to diagnose pneumonia. The designed hybrid framework is developed by combining both convolutional networks and transformer encoder mechanisms. The method was trained and evaluated using Mendeley [13] and chest x-ray [14] datasets for binary and multiple classification tasks. Their hybrid framework produced over 95% accuracy and F1 score values for classification tasks. Cha et al. [7] developed a transfer learning framework based on the attention mechanism for effective pneumonia detection on CXR images. First, as a feature extractor, features were collected from three pre-trained models, namely ResNet152, DenseNet121, and ResNet18. Then, the attention mechanism was implemented as a feature selection operation. The proposed approach achieved a 96.63% accuracy score, a 97.3% F1-score, a 96.03% area under the curve (AUC), a 96.23% precision and a 98.46% sensitivity. Singh et al. [16] combined a deep neural network (DNN) with an attention mechanism to detect pneumonia disease using CXR images. The proposed network is built to generate attention-sensitive features by combining channel and spatial attention modules in the DNN architecture. They studied a public CXR dataset with the proposed network [13]. The proposed network yielded a classification accuracy of 95.47% and an F1-score of 0.92. Tyagi et al. [17] proposed an auto-detection model for the early detection of pneumonia. Three models, namely CNN, VGG16, and Vision Transformer, were used in their study. In addition, they studied a publicly available CXR dataset to develop their models [13]. The ViT method used in experimental studies showed an accuracy performance of 96.45% in identifying pneumonia.

Ma et al. [18] used the Swin transformer as the pneumonia recognition model in CXR images and optimized them according to the characteristics of CXR images. After comparative experiments on two different datasets [13, 14], the experimental results showed that the model's accuracy increased from 76.3% to 87.3% and from 92.8% to 97.2%, respectively. Jiang et al. [19] proposed the multi-level patch merger vision transformer (MP-ViT) to automatically diagnose pneumonia from CXR images. They performed their experiments on a publicly available dataset [13]. Their proposed model achieved 0.91 accuracy, 0.92 precision, 0.89 sensitivity, and 0.90 F1-score. Wei et al. [20] proposed a deep learning architecture called Deep Pneumonia to recognize pneumonia lesions. The authors proposed a feature learning mechanism based on the mask-driven intense attention and comparative editing strategies. These strategies were applied to the attention map and the extracted features to draw more attention to the lesion area with more distinctive features and guide the model. Their proposed model achieved an accuracy of 83.85%. Okolo et al. [21] evaluated the performance of a new deep-learning model based on a transformer network for the CXR image classification task. They examined the ViT performance and then proposed and evaluated the input-enhanced vision transformer (IEViT). Experiments on four CXR image datasets containing various pathologies (pediatric pneumonia, tuberculosis, viral pneumonia, COVID-19) showed that the developed IEViT architecture produced a higher accuracy score than the ViT model for all datasets. Mabrouk et al. [22] proposed a computer-assisted tool for detecting pneumonia, using ensemble learning using CXR images. Their suggested ensemble learning model comprised three well-known CNN models, namely DenseNet169, MobileNetV2, and ViT. These models were trained on the CXR dataset using fine-tuning [13]. The proposed ensemble learning model outperformed other state-of-the-art methods, achieving 93.91% accuracy and a 93.88% F1 score. Ar et al. [23] proposed an ensemble framework for pneumonia detection using chest X-ray images. The proposed framework has been tested and evaluated using the CXR dataset [13]. VDSNet and CAPSNet models were used for the proposed pneumonia detection framework. Their proposed ensemble (combined) model reached 98% accuracy.

The aim of this study is to produce a new CXR dataset as an alternative to the pneumonia datasets in the literature. The new data set was created for binary classification tasks as pneumonia and healthy classes. Our study uses CXR images of healthy and pneumonia patients obtained from a private and public database. We also developed a new CNN-based PneumoNet model in this study. It is about designing a network that can be compared to pre-trained networks. Additionally, we compared the model developed in this study with four transformer-based models, namely ViT, FNet, gMLP and MLP mixers, for pneumonia detection using CXR images. The proposed models were compared in terms of classification performances such as recall, precision, F1-score, and accuracy. Our developed PneumoNet model yielded 96.50% and 94.29% accuracy for private and public databases, respectively, detecting pneumonia accurately from healthy subjects. The organization of this paper is as follows. In this section, medical information about pneumonia is given in detail. Literature studies on pneumonia detection using transformer-based models are mentioned and the purpose of this study is also explained. Section 2 provides detailed information about the datasets used for pneumonia detection in this study, the developed deep learning-based pneumonia detection model, transformer-based models, and performance evaluation metrics. In Section 3, the steps followed in experimental studies and the results of experimental studies are mentioned. The discussion is described in Section 4. In Section 5, the results obtained in the study are evaluated.

2. Material and Method

2.1. Private database

The private database used in this study consists of pneumonia and healthy CXR images obtained from Dicle University Medical Faculty Chest Diseases and Tuberculosis clinic, intensive care unit, and chest polyclinic. The study on 2000 subjects, 1000 subjects were diagnosed with pneumonia on Poster Anterior (PA) chest X-ray and 1000 subjects had normal PA chest X-ray. All images were in different sizes and RGB. Figure 1 shows typical pneumonia and healthy CXR images obtained from the private database.

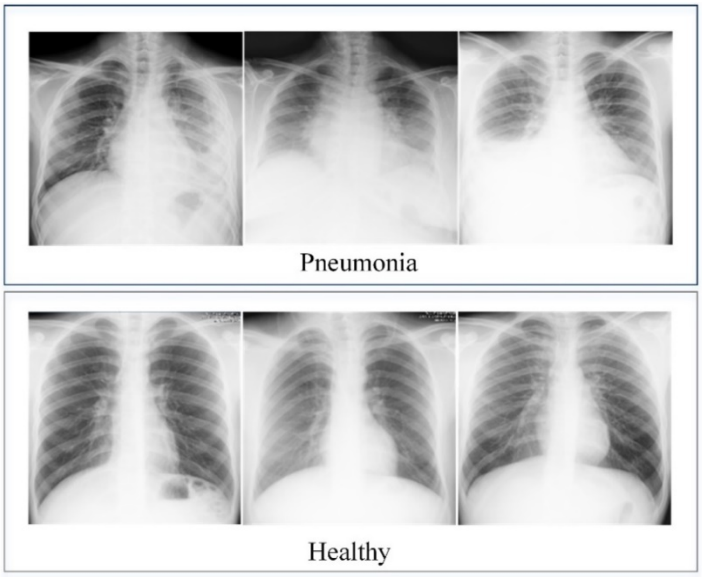

Figure 1. Typical healthy and pneumonia CXR images collected from our private dataset.

2.2. Public database

In order to test the effectiveness of the developed pneumonia detection network, experimental studies were carried out with CXR image datasets, which are the most used in the literature for pneumonia detection [13, 14]. To avoid class imbalance problems, equal numbers of CXR images from both classes in the datasets were randomly selected and used. In this study, 3500 images were used, 1750 of which were from the healthy class and the other 1750 from the pneumonia class.

2.3. PneumoNet model

A new PneumoNet model is developed and trained end-to-end for pneumonia classification. The developed PneumoNet model consists of 21 layers, as shown in Figure 2. The new CNN architecture starts with an input layer. It continues with the batch normalization and ReLU layer, which then follows each convolution layer. Among the pooling operations, the max operator function was used in the pooling layers. There are five convolution layers in the 21-layer CNN architecture. Two pool layers, pool_1, and pool_2, come after the ReLu_1 and ReLu_2 layers. The last three layers of the proposed new network are used for classification purposes. The proposed PneumoNet architecture for pneumonia detection is shown in Figure 2. Besides, Table 1 shows the various setting parameters of the developed PneumoNet model.

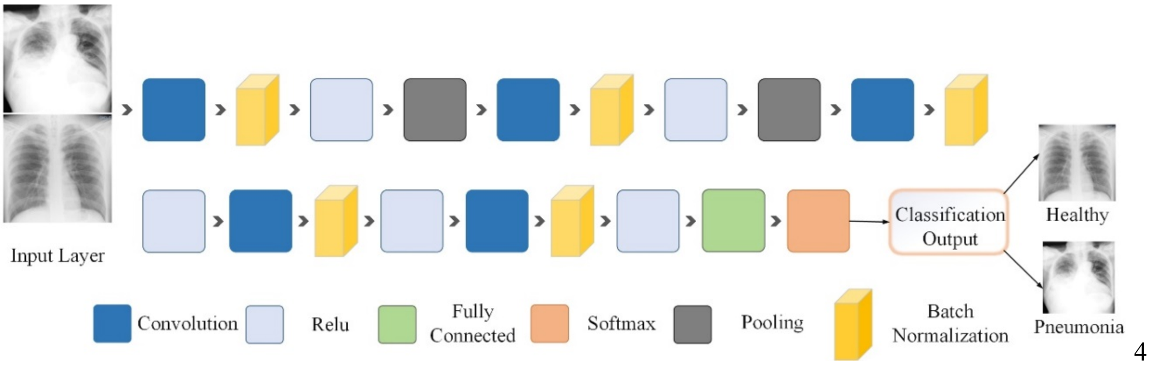

Figure 2. Architecture of PneumoNet for pneumonia detection.

Table 1. Hyperparameters used to develop the PneumoNet model.

Type	Activations	Learnable
Image input	224×224×3	-
Convolution	224×224×64	Weights 3×3×3×64 Bias 1×1×64
Batch Normalization	224×224×64	Offset 1×1×64 Scale 1×1×64
ReLu	224×224×64	-
Max Pooling	112×112×64	-
Convolution	112×112×32	Weights 3×3×3×64 Bias 1×1×64
Batch Normalization	112×112×32	Offset 1×1×32 Scale 1×1×32
ReLu	112×112×32	-
Max Pooling	56×56×32	-
Convolution	56×56×16	Weights 3×3×3×64 Bias 1×1×64
Batch Normalization	56×56×16	Offset 1×1×16 Scale 1×1×16
ReLu	56×56×16	-
Convolution	56×56×8	Weights 3×3×3×64 Bias 1×1×64
Batch Normalization	56×56×8	Offset 1×1×8 Scale 1×1×8
ReLu	56×56×8	-
Convolution	56×56×4	Weights 3×3×3×64 Bias 1×1×64
Batch Normalization	56×56×4	Offset 1×1×4 Scale 1×1×4
ReLu	56×56×4	-
Fully Connected	1×1×2	Weights 2×12544 Bias 2×1
Softmax	1×1×2	-
Classification Output	-	-

2.4. Vision Transformer (ViT)

ViT [12], is a computer vision-adapted version of attention mechanism-based transformer models that have been successfully performed in NLP applications. When classifying an image, ViT treats it as a patch sequence, similar to a sequence of word embedding generated by the NLP transformer. The working principle of the ViT model consists of the following steps. The ViT splits an input image into sequences of patches or visual tokens. Each 2D image patches are flattened. Then, the flattened patches are embedded linearly, called patch embedding.

Learnable position embeddings are added to each image patch. An embedded image patch is combined with extra learnable class tokens to predict the class of the new image. Finally, the resulting sequence is given to the transformer encoder block. While the standard transformer takes 1D token embedding sequences as input to handle a 2D image as a 1D sequence in ViT; $x \in \mathbb{R}^{H\times W\times C}$ input image is reshaped $x_p \in \mathbb{R}^{N\times(P^2.C)}$ as a sequence of flattened 2D patches. The number of patches is calculated using Equation (1).

$$N = HW/P^2 \tag{1}$$

Where (H, W) is the size of each original image, i.e., the height and width values, C is the number of image's channels, (P, P) is the size of each square image patch. The following equations explain the handling of images in the ViT. Flattening the patches and mapping them to the D dimension with a trainable linear projection is calculated as shown in Equation (2).

$$z_0 = \left[x_{class}; x_p^1\mathrm{E}; x_p^2\mathrm{E}; \dots; x_p^N\mathrm{E}\right] + \mathrm{E}_{pos}, \qquad \mathrm{E}\epsilon\mathbb{R}^{(P^2.C)\times D}, \mathrm{E}_{pos}\epsilon\mathbb{R}^{(N+1)\times D} \tag{2}$$

The sum size of the flattened patches (x_p) can be determined using Equation (3).

$$x_p = N \times (P^2.C) \tag{3}$$

In Equation (4,5), there are MLP (Multilayer Perceptron) and MSA (Multi-head self-attention) blocks of the transformer encoder layer (L). The outputs of the ℓ.th layer are calculated as in the equations. In Equation (6), the image representation of the encoder's output (z_L^0) is denoted by y [12]. The illustration of the developed ViT architecture is given in Figure 3.

$$z'_\ell = MSA\big(LN(z_{\ell-1})\big) + z_{\ell-1}, \qquad \ell = 1 \dots L \tag{4}$$
$$z_\ell = MLP\big(LN(z'_\ell)\big) + z'_\ell, \qquad \ell = 1 \dots L \tag{5}$$
$$y = LN(z_L^0) \tag{6}$$

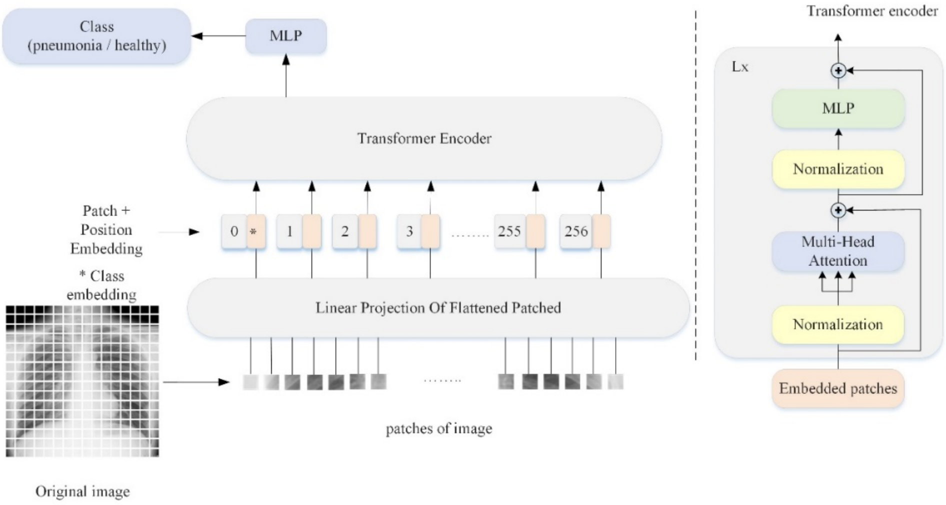

Figure 3. Illustration of the ViT model for pneumonia detection.

2.5. MLP mixer

The basic architecture of the developed MLP-mixer is given in Figure 4. As can be seen, the input colour image has a size of 512×512, and a non-overlapping window of size 32×32 is used for patch extraction. After all, patches are extracted, each patch is initially flattened, and a linear projection layer is employed to reduce the number of samples in each flattened patch. The linear projection layer is formulated using Equation (7).

$$y = x\omega^T + b \tag{7}$$

Where x is the input vector, ω, and b are the weights vector and bias value, respectively. The length of a flattened patch is 3072; after linear projection, the length of each patch vector becomes 2048. Similarly, after the linear projection layer, a patch embedding matrix of size 256×1024 is obtained for all patches. Next, these patches embedding of shape 256×1024 of an input image goes through mixer layers before being fed to the MLP Head for classification [24].

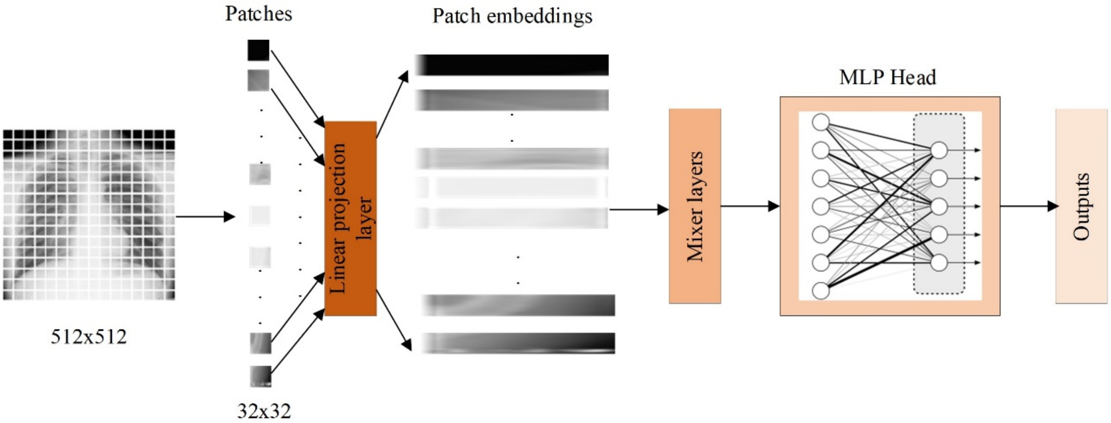

Figure 4. Illustration of the MLP-mixer architecture.

2.6. FNet

In the model introduced by Lee-Thorp et al., the Fourier transform is used as the token mixing mechanism [25]. With this model, simple linear transforms, including Fourier transforms, have been shown to be competent in modeling various relationships in text data. The advantage of FNet hybrid models, which contain only two self-attention sublayers, over other transformer models, is that they run faster and perform better at longer input lengths. In addition, this model has shown that the attention mechanism affects increasing accuracy, but it is not necessary to use it in every layer. As shown in Equation (8), based on FNet, the attention sublayer has been replaced with a Fourier sublayer that implements 2D DFT. Here $\mathcal{F}_{seq}$ is the sequence length, $\mathcal{F}_h$ is the hidden dimension. The developed FNet architecture for this study is given in Figure 5.

$$y = \Re(\mathcal{F}_{seq}(\mathcal{F}_h(x))) \tag{8}$$

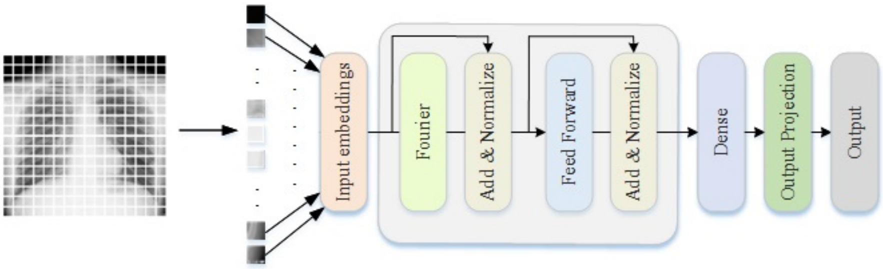

Figure 5. An illustration of the FNet architecture.

2.7. GMLP

The gMLP proposed by Liu et al. is a without self-attention MLP-based transformer model [26]. It was designed by combining basic MLP layers with gating to research the necessity of attention mechanisms in transformer architectures. The innovation in gMLP uses a spatial gating unit (SGU), which learns the complex spatial interactions between sequence elements, as an alternative to attention mechanisms. SGU does not require encoding for element positions because such information is held in the $s(\cdot)$ layer, which captures spatial interactions. gMLP uses fewer trainable parameters than other transformer models. The basic working principle of gMLP for the initial matrix $X \in \mathbb{R}^{n\times d}$ with the length of the token sequence n and the size of the token, d are shown in the following equations. Where σ is an activation function like Gaussian Error Linear Unit (GeLU), U

and V are matrices that describe linear projections along the channel dimension, $s(\cdot)$ is the layer that captures spatial interactions. Y is the output of the block. The illustration of the proposed gMLP architecture is given in Figure 6. Each block can be defined as in the equations (9) to (11).

$$\mathrm{Z} = \sigma(XU) \tag{9}$$
$$\check{\mathrm{Z}} = s(\mathrm{Z}) \tag{10}$$
$$Y = \check{\mathrm{Z}}V \tag{11}$$

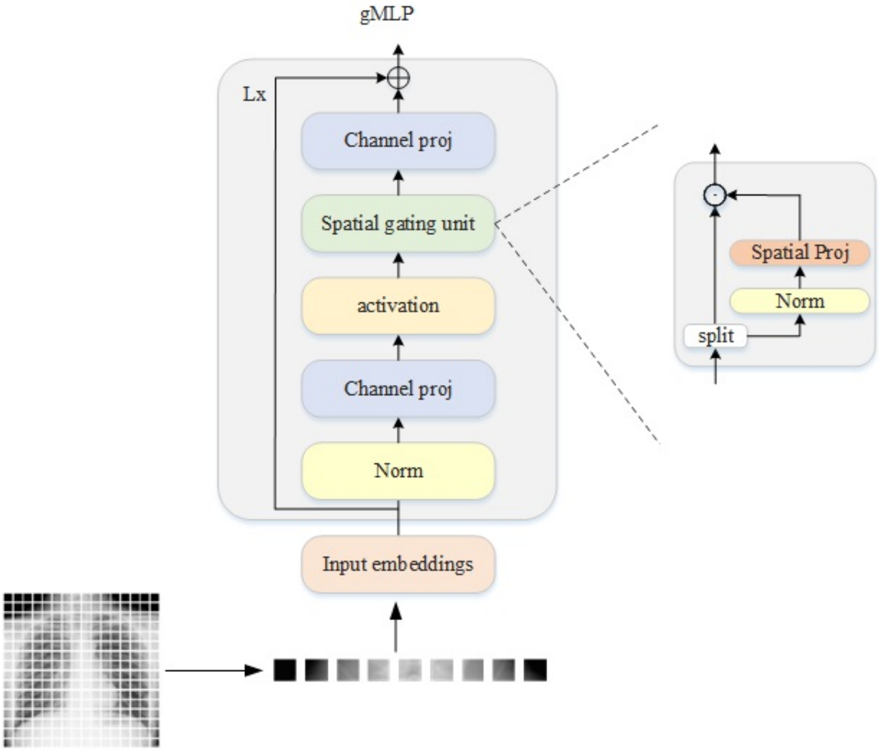

Figure 6. Illustration of the gMLP architecture.

2.8. Performance Metrics

Various evaluation metrics, i.e., accuracy, precision, recall, and F1 score, were used to assess the performance of the proposed methods in the study. Accuracy, as in Equation 12, refers to the number of data samples correctly identified out of the total number of data samples given [27,28].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

Precision/positive predictive value, is the ratio of correctly detected positive cases to all expected positive cases, as shown in Equation 13.

$$Precision = \frac{TP}{TP+FP} \tag{13}$$

The sensitivity/recall/true positive ratio, is the number of samples correctly identified as positive cases compared to all true positive cases, as shown in Equation 14.

$$Recall = \frac{TP}{TP+FN} \tag{14}$$

F1 Score is defined as the harmonic mean of sensitivity and precision, also indicates the overall accuracy of the method and is calculated as shown in Equation 15.

$$F_1 = \frac{2*(Precision*Recall)}{Precision+Recall} \tag{15}$$

True Positive (TP) is the pneumonia image being classified as pneumonia by the model. True Negative (TN) labels the non-pneumonia image, the healthy image, as healthy by the post-classification model. False Positive (FP) is when the model labels the healthy image as pneumonia. False Negative (FN) is the model's classification of the pneumonia image as a healthy image.

3. Experimental Results

This section presents experimental evaluations of proposed pneumonia detection methods. All experiments were conducted on Python and MATLAB, using a computer with an A5000 GPU with 45 GB RAM. During the pre-processing stage of the study, since the CXR images in the dataset are of different sizes, they are resized to 512×512 to meet the minimum input dimensions of the transformer-based models. The CXR images are resized to 224×224 for the PneumoNet model additionally. Then each image was divided into a total of 256 patches of size 32×32 pixels. Adam optimizer was considered in training the mentioned methods. Besides, hold-out validation was used to split the dataset into the training and test data, where the division ratio was 90% for training and 10% for testing [29, 30]. For the developed PneumoNet model, the learning rate was 0.0001, the batch size was 10, and the number of epochs was 8. Stochastic gradient descent with momentum (SGDM) optimizer was considered in training the mentioned methods.

The transformer-based models' parameters were assigned heuristically while running each code. For the ViT model, the learning rate was set to 0.001, weight decay was set to 0.0001, and batch size, number of epochs, projection dimension, number of heads and number of transformer layers were set to 64, 400, 64, 4 and 8, respectively. For FNet, MLP-mixer and gMLP models, weight decay was set to 0.0001, batch size, number of epochs, embedding dimension, number of MLP-mixer layers and dropout rate were set to 64, 400, 256, 4, and 0.2, respectively. The learning rate was set to 0.005 for the MLP mixer, 0.001 for FNet and 0.003 for gMLP. The flowchart of the proposed method is given in Figure 7. The flowchart of the proposed method is given in Figure 7. Classification of pneumonia and healthy CXR images with transformer-based models and 21-layer CNN. Performance metrics such as accuracy, precision, sensitivity, and F1-measure were used to evaluate the effectiveness of the proposed methods.

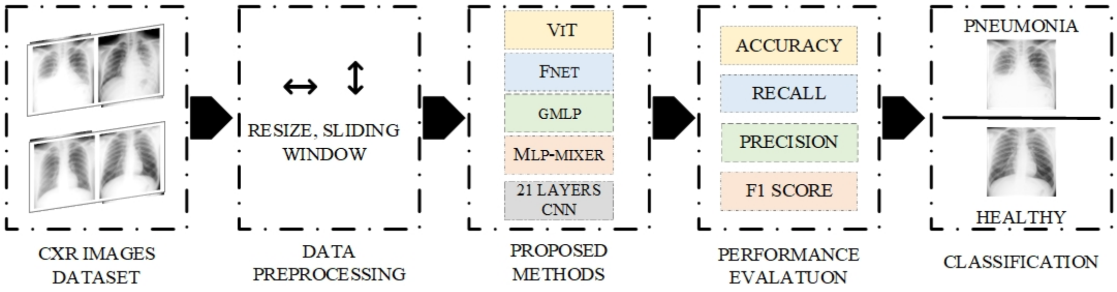

Figure 7. Flowchart of the proposed method.

Table 2 gives the evaluation metrics of the classification obtained for each method. The rows in Table 2 show the type of model, and the columns show the classification performance parameters obtained. As seen in Table 2, the ViT and FNet models produced identical accuracy, precision, recall and F1-scores values. The produced evaluation scores were 96.43%, 97.98%, 95.10% and 96.52%, respectively. Besides, MLP-mixer produced a 95.41% accuracy score, 96.04% precision value, 95.10% recall value and 95.57% F1-score, respectively. Finally, 94.39% accuracy score, 97.89% precision, 91.18% recall and 94.42% F1-score values were obtained by the gMLP approach. As the calculated evaluation metrics were observed, it was seen that ViT and FNet achievements were the highest among the examined transformer-based approaches. The MLP-mixer model produced the second-best evaluation scores (F1-scores, recall, and accuracy). gMLP produced the second-best precision score, where the calculated precision was 97.89%. However, the proposed PneumoNet outperformed all the transformer-based models and reported an accuracy of 96.50%, precision rate of 100%, recall of 93.46 and F1 score of 96.62% using the private database with a hold-out strategy. Considering the small amount of data in this study and the use of transformer-based models as the basic architecture during experimental studies, it is considered reasonable that their performance is relatively low.

Table 2. Summary of classification performances obtained for used models.

Model	F1 score (%)	Recall (%)	Precision (%)	Accuracy (%)
ViT	96.52	**95.10**	97.98	96.43
FNet	96.52	95.10	97.98	96.43
MLP-mixer	95.57	95.10	96.04	95.41
gMLP	94.42	91.18	97.89	94.39
PneumoNet	**96.62**	93.46	**100.00**	**96.50**

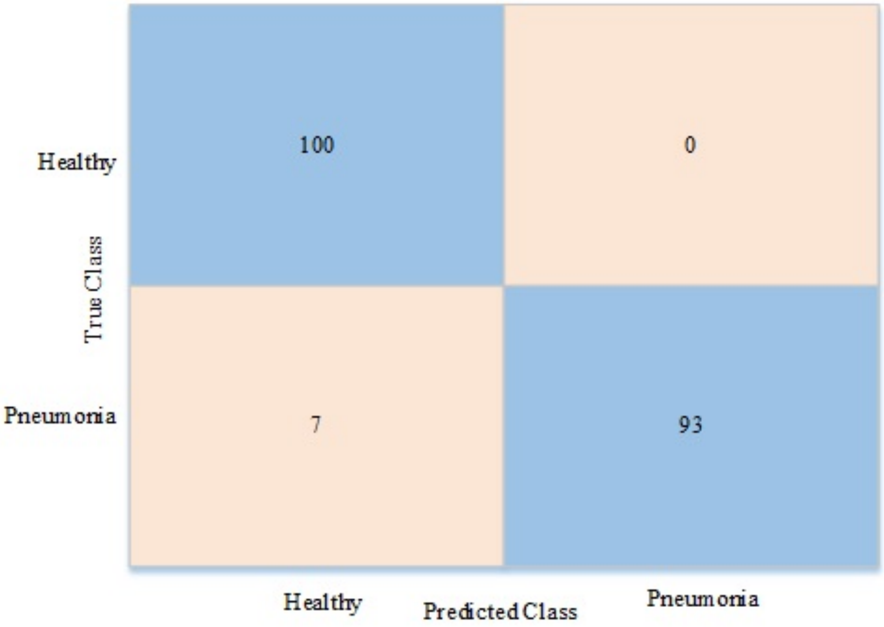

Figure 8. Confusion matrix obtained using our PneumoNet model.

The confusion matrix acquired by the developed PneumoNet model is illustrated in Figure 8. While the rows of the confusion matrix show the number of instances of the true classes, the columns show the number of instances of the predicted classes. As can be seen from Figure 8, the healthy class was 100% correctly classified, and 7 pneumonia samples were misclassified. The summary of performance obtained using the PneumoNet model is included in Table 3. As indicated in the Table 3, the developed PneumoNet model produced 96.50% accuracy score, 100% precision, 93.46% recall, and 96.62% F1 score values for the private database with a hold-out validation strategy. In addition, the PneumoNet model yielded precision, F1-score, recall and accuracy values were 93.33%, 92.15%, 91.00% and 92.25%, respectively, using a private database with a ten-fold cross-validation strategy.

Table 3. Summary of performance obtained using the PneumoNet model for the private database.

Model	Data Partition	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
PneumoNet	Hold-out Validation (90% training and 10% test)	96.50	100.00	93.46	96.62
PneumoNet	Ten-fold cross validation	92.25	93.33	91.00	92.15

3.1. Explainable Transformer models with the Grad-CAM technique

Figure 9 shows the heat maps obtained for healthy and pneumonia classes employing the Grad-CAM method for the PneumoNet. It helps to delineate the problematic region of the input image for the clinicians and helps to detect pneumonia from healthy classes. In this work, we have used the well-known Grad-CAM approach [31,32]. Grad-CAM enables viewing every model layer and examining each feature map layer, both of which are required to understand how input values influence model categorization. In this work, we also used Grad-CAM on the output of the PneumoNet model. Figure 9 shows the Grad-CAM illustrations for pneumonia and healthy classes. As seen in Figure 9, the images in the first-row show pneumonia, and the images in the second row show the healthy subjects.

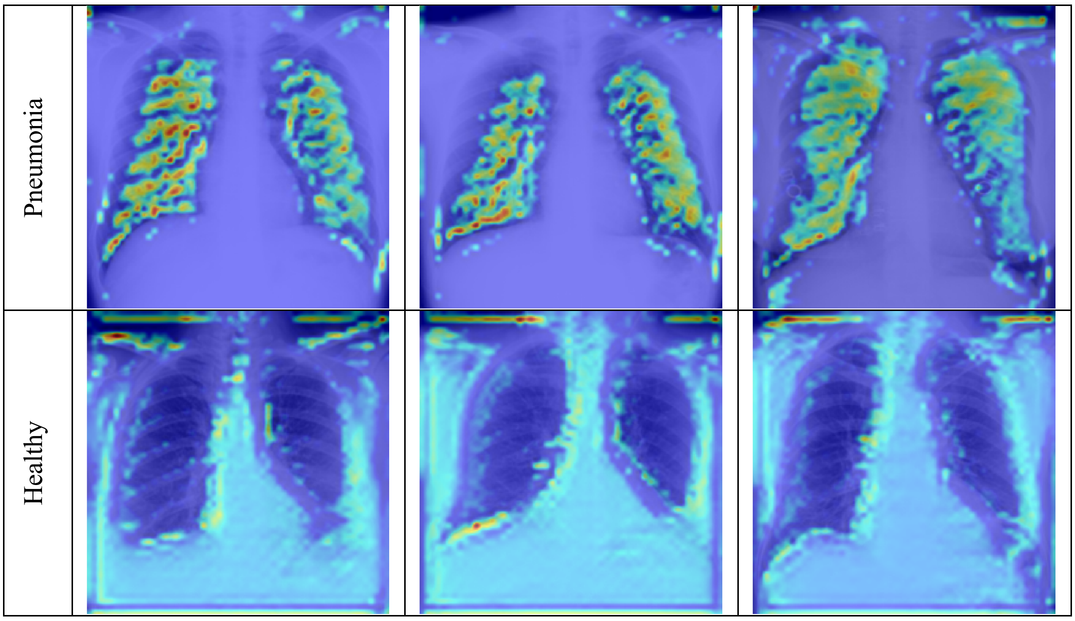

Figure 9. Heat maps obtained for healthy and pneumonia classes using the Grad-CAM technique for the PneumoNet model.

As illustrated in Figure 9, for the pneumonia class, the PneumoNet model mostly considered the left lung blob for pneumonia detection and has not concentrated on lung blobs for the health class for the public database.

The performance evaluation obtained using a 10-fold cross-validation test for the public databases is given in Table 4. We have obtained a precision value of 94.46%, an accuracy of 94.29 %, an F1 score of 94.27, and a recall rate of 94.10%.

Table 4. Summary of performance matrices obtained using our PneumoNet model on CXR image datasets with ten-fold cross-validation strategy.

Model	Data Partition	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
PneumoNet	Cross Validation (10-fold)	94.29	94.46	94.10	94.27

4. Discussion

This study has proposed a novel PneumoNet model and compared its performance with various transformer-based approaches to classify CXR images to detect pneumonia. In this study, the novel model was developed using private and public databases. The obtained results showed that the proposed CNN-based model is effective and accurate in detecting pneumonia using the CXR dataset collected for this study. As presented in Table 2, classification accuracy of the developed model is 96.5% and 94.29% for private and public databases, respectively in detecting pneumonia. Table 5 lists previous studies published to detect pneumonia using publicly available CXR datasets. It can be noted that most of the studies used different CNN techniques and transformer-based ensemble CNN frameworks for pneumonia detection. As shown in Table 5, when studies in the literature were reviewed for the detection of pneumonia, a dataset containing 5856 CXR images was widely used in the literature in binary classification tasks [13]. The Kaggle Pneumonia dataset is divided into subfolders for each image category (Pneumonia/Normal) and is arranged into three folders: training, testing, and validation. However, we did not use the data set in this form in our study. We combined the same class images in the training and test folders to ensure similarity to the data set we created. To make a two-class classification as normal and pneumonia. Additionally, an extra data set was used, and the CXR images were selected randomly.

As seen in Table 5, the best accuracy score of 99.21% in binary classification was obtained by Ukwuoma et al. [15] using a hybrid transformer encoder-based deep learning model. Authors in [21] obtained a 98.08% accuracy score using the input-enhanced vision transformer model, and Ar et al. [23] obtained 98.0% using a transformer-based ensemble framework on the same CXR pneumonia image database. Our study is the first work to use private and public databases for pneumonia detection using CXR images accurately (Table 5).

Table 5. Comparison of the classification performance of our work with the latest techniques developed using CXR images for pneumonia detection.

Authors	Dataset	Method	Compared techniques	Acc %
Ukwuoma et al. [15]	Kermany [13] Chest X-ray [14]	Hybrid Deep Learning Framework	Ensemble A (DenseNet201, VGG16, GoogleNet) Ensemble B(DenseNet201, InceptionResNetV2, Xception)	[13] 99.21 [14] 98.19
Cha et al. [7]	[13]	Attention-Based Transfer Learning Framework	ResNet152 ResNet18 DenseNet	96.63
Singh et al. [16]	[13]	Deep Attention Network	ResNet, ResNet with attention	95.47
Tyagi et al. [17]	[13]	CNN Based Framework	CNN, VGG16, ViT	96.45
Ma et al. [18]	[13] [14]	Transformer Backbone Network	Swin Transformer	[13] 97.2 [14] 87.3
Jiang et al. [19]	[13]	Multisemantic Level Patch Merger Vision Transformer	Baseline (ResNet50), ViT, ViT + Patch Merger	91.18
Wei et al. [20]	20,012 CXR images	Attention Based Contrastive Learning	ResNet18 (backbone)	83.85
Okolo et al. [21]	[13]	Input Enhanced Vision Transformer	IEViT variants	98.08
Mabrouk et al. [22]	[13]	Ensemble Of Deep CNN	DenseNet169, MobileNetV2, ViT	93.91
Ar et al. [23]	[13]	Ensembling Framework	CAPSNet, VDSNet, Ensemble Scheme	98.0
This work	Own dataset	CNN-Based	PneumoNet	96.50
This work	Public dataset [13,14]	CNN-Based	PneumoNet and XAI	94.29

In this study, a 21-layer CNN architecture is proposed to separate and classify input images into two classes. This network was created and trained from scratch using the end-to-end training method. Its purpose is to compare its performance with other pre-trained networks. The reason why our model gives low results when compared to studies in the literature is that the models we compare in the literature are generally in ensemble structures. Both the transformer models and the new 21-layer CNN are backbones.

The advantages of this study are given below:

1- A new dataset was created for the automated detection of pneumonia.
2- Our novel PneumoNet model has yielded an accuracy of 96.5% and 94.29% for private and public databases, respectively, in detecting pneumonia accurately from healthy subjects.
3- Our proposed model yielded higher classification performance than the recently developed transformer-based models for pneumonia detection.
4- Shown typical heatmaps for the healthy and pneumonia classes to develop confidence in clinicians by showing the regions of interest.
5- A new PneumoNet was proposed for pneumonia detection.

5. Conclusion

Chest X-ray is a promising imaging modality for diagnosing pneumonia as it is economical, fast, and readily available. The adoption of deep learning-based methods for pneumonia detection provides significant benefits for improving the interpretation, usability, accuracy, and consistency of CXR images. This study proposed a new model called PneumoNet for detecting pneumonia using CXR images and compared classification performance with recently transformer networks. Our developed model has yielded 96.5% and 94.29% accuracy for private and public databases, respectively, in detecting pneumonia accurately from healthy subjects. The limitation of this study is that we only used two datasets for this work. We plan to evaluate the model's performance using more chest X-ray images from diverse races. Also, we intend to explore the possibility of using this developed model to detect other pulmonary disorders using CXR images.

Acknowledgements

Ethics approval

This article is derived from the PhD thesis of the corresponding author Zehra Kadiroğlu. This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of University of Dicle. Date: 13.10.2021, Number: 421. We would like to thank Dicle University Faculty of Medicine, Department of Chest Diseases and Tuberculosis for their contribution to the study. The dataset used for pneumonia detection during this study is available from the corresponding author upon reasonable request.

Author Contribution

Conceptualization Abdulkadir Sengur, Zehra Kadiroglu, and Erkan Deniz; investigation, Abdulkadir Sengur, Zehra Kadiroglu, Mazhar Kayaoglu and Erkan Deniz; methodology, Abdulkadir Sengur, Zehra Kadiroglu, Mazhar Kayaoglu and Hanifi Guldemir; software, Zehra Kadiroglu, and Mazhar Kayaoglu; project administration, Abdurrahman Senyigit, and Hanifi Guldemir ; resources, Abdurrahman Senyigit, and Hanifi Guldemir; supervision; Hanifi Guldemir, and Abdurrahman Senyigit; visualization, Abdulkadir Sengur, Zehra Kadiroglu, and Mazhar Kayaoglu; writing—original draft, Abdulkadir Sengur, Zehra Kadiroglu, Mazhar Kayaoglu; All authors have read and agreed to the published version of the manuscript.

References

[1] World Health Organization. “Pneumonia.” Erişim: 9 Aralık 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ pneumonia

[2] Torres A, Cilloniz C, Niederman MS, Menéndez R, Chalmers JD, Wunderink RG, van der Poll T. Pneumonia. Nat Rev Dis Primers 2021;7(1):25.

[3] Kumar S, Singh P, Ranjan M. A review on deep learning based pneumonia detection systems. In: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, Coimbatore, India: IEEE.pp. 289-296.

[4] Kwon T, Lee SP, Kim D, Jang J, Lee M, Kang SU, Kim H, Oh K, On J, Kim YJ, Yun SJ, Jin KN, Kim EY, Kim KG. Diagnostic performance of artificial intelligence model for pneumonia from chest radiography. PLoS One 2021;16(4):e0249399.

[5] Mujahid M, Rustam F, Álvarez R, Luis Vidal Mazón J, Díez IT, Ashraf I. Pneumonia Classification from X-ray Images with Inception-V3 and Convolutional Neural Network. Diagnostics 2022;12(5):1280.

[6] Govindarajan A, Govindarajan A, Tanamala S, Chattoraj S, Reddy B, Agrawal R, Iyer D, Srivastava A, Kumar P, Putha P. Role of an Automated Deep Learning Algorithm for Reliable Screening of Abnormality in Chest Radiographs: A Prospective Multicenter Quality Improvement Study. Diagnostics 2022; 12(11):2724.

[7] Cha S-M, Lee S-S, Ko B. Attention-Based Transfer Learning for Efficient Pneumonia Detection in Chest X-ray Images. Appl Sci 2021; 11(3):1242.

[8] Al Mamlook RE, Chen S, Bzizi, HF. Investigation of the performance of machine learning classifiers for pneumonia detection in chest X-ray images. In: 2020 IEEE International Conference on Electro Information Technology (EIT); 2020, Chicago, IL, USA, IEEE: pp. 98-104.

[9] Bai Y, Mei J, Yuille A L, Xie C. Are transformers more robust than cnns?. Advances in Neural Information Processing Systems, 2021; 34:26831-26843.

[10] Usman M, Zia T, Tariq A. Analyzing Transfer Learning of Vision Transformers for Interpreting Chest Radiography. J Digit Imaging 2022;35(6):1445-1462.

[11] Vaswani A, Shazeer NM, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is All you Need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, USA, 2017, pp. 5998-6008.

[12] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations (ICLR); 2021, Virtual Event, Austria, *arXiv:2010.11929.*

[13] Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172(5):1122-1131.e9.

[14] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospitalscale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; Honolulu, HI, USA: IEEE, pp. 3462-3471

[15] Ukwuoma CC, Qin Z, Belal Bin Heyat M, Akhtar F, Bamisile O, Muaad AY, Addo D, Al-Antari MA. A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. J Adv Res 2023; 48: 191-211.

[16] Singh S, Rawat S, Gupta M, Tripathi B, Alanzi F, Majumdar A, Khuwuthyakorn P, Thinnukool O. Deep attention network for pneumonia detection using chest x-ray ımages. Comput Mater Contin 2023; 74(1): 1673-1691.

[17] Tyagi K, Pathak G, Nijhawan R, Mittal A. Detecting pneumonia using vision transformer and comparing with other techniques. In: 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2021; Coimbatore, India, IEEE: pp. 12-16.

[18] Ma Y, Lv W. Identification of Pneumonia in Chest X-Ray Image Based on Transformer. Int J Antennas Propag 2022; 2022(1): 5072666.

[19] Jıang Z, Chen L. Multisemantic level patch merger vision transformer for diagnosis of pneumonia. Comput Math Methods Med 2022; 2022(1): 7852958.

[20] Wei X, Niu X, Zhang X, Li Y. Deep Pneumonia: Attention-Based Contrastive Learning for Class-Imbalanced Pneumonia Lesion Recognition in Chest X-rays. In: 2022 IEEE International Conference on Big Data (Big Data); 2022; IEEE. pp. 5361-5369.

[21] Okolo GI, Katsigiannis S, Ramzan N. IEViT: An Enhanced Vision Transformer Architecture for Chest X-ray Image Classification. Comput Methods Programs Biomed 2022; 226:107141.

[22] Mabrouk A, Díaz Redondo RP, Dahou A, Abd Elaziz M, Kayed M. Pneumonia Detection on Chest X-ray Images Using Ensemble of Deep Convolutional Neural Networks. Appl Sci 2022; 12(13):6448.

[23] Gokul AG, Kumaratharan N, Rani PL, Devi N. Ensembling Framework for Pneumonia Detection in Chest X-ray images. In: 2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN); 2022; Villupuram, India: IEEE, pp. 1-5.

[24] Tolstikhin IO, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, Yung J, Keysers D, Uszkoreit J, Lucic M, & Dosovitskiy A. MLP-Mixer: An all-MLP Architecture for Vision. In NeurIPS 2021, 34: 24261-24272.

[25] Lee-Thorp J, Ainslie J, Eckstein I, Ontanon S. FNet: Mixing tokens with Fourier transforms. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2022; Seattle, United States: Association for Computational Linguistics, pp. 4296-4313.

[26] Liu H, Dai Z, So D, Le QV. Pay attention to mlps. In NeurIPS 2021; 34:9204-9215.

[27] Visuña L, Yang D, Garcia-Blas J, Carretero J. Computer-aided diagnostic for classifying chest X-ray images using deep ensemble learning. BMC Med Imag 2022; 22(1): 1-16.

[28] Jain DK, Singh T, Saurabh P, Bisen D, Sahu N, Mishra J, Rahman H. Deep Learning-Aided Automated Pneumonia Detection and Classification Using CXR Scans. Comput Intell Neurosci 2022;2022(1): 7474304.

[29] Şengür D. Investigation of the relationships of the students' academic level and gender with Covid-19 based anxiety and protective behaviors: A data mining approach. Turkish J Sci Technol 2020; 15(2): 93-99.

[30] Şengür D, Siuly S. Efficient approach for EEG-based emotion recognition. Electron Lett 2020; 56(25): 1361-1364.

[31] Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Comput Methods Programs Biomed 2022; 226:107161.

[32] Sobahi N, Atila O, Deniz E, Sengur A, Acharya UR. Explainable COVID-19 detection using fractal dimension and vision transformer with Grad-CAM on cough sounds. Biocybern Biomed Eng 2022; 42(3):1066-1080.