Gazi University
Journal of Science
PART A: ENGINEERING AND INNOVATION
http://dergipark.org.tr/gujsa

Diagnosis of Pneumonia from Chest X-ray Images with Vision Transformer Approach

Emrah ASLAN[1*]

[1] Dicle Üniversitesi, Silvan Meslek Yüksekokulu, Diyarbakır, Türkiye

Keywords
ViT
Pneumonia
Artificial Intelligence
CNN

Abstract

People can get pneumonia, a dangerous infectious disease, at any time in their lives. Severe cases of pneumonia can be fatal. A doctor would usually examine chest x-rays to diagnose pneumonia. In this work, a pneumonia diagnosis system was developed using publicly available chest x-ray images. Vision Transformer (ViT) and other deep learning models were used to extract features from these images. Vision Transformer (ViT) is an attention-based model used for image processing and understanding as an alternative to the convolutional neural networks traditionally used for this purpose. ViT consists of a series of attention layers, where each attention layer models the relationships between input pixels to represent an image. These relationships are determined by a set of attention heads and then fed into a classifier. ViT performs effectively in a variety of visual tasks, especially when trained on large datasets. The study shows that the ViT model's classification procedure has a high success rate of 95.67%. These results highlight how deep learning models can be used to quickly and accurately diagnose dangerous diseases such as pneumonia in its early stages. The study also shows that the ViT model outperforms current approaches in the biomedical field.

Cite

Aslan, E. (2024). Diagnosis of Pneumonia from Chest X-ray Images with Vision Transformer Approach. *GU J Sci, Part A*, *11*(2), 324-334. doi:10.54287/gujsa.1464311

Author ID (ORCID Number)		**Article Process**	
0000-0002-0181-3658	Emrah ASLAN	**Submission Date**	04.04.2024
		Revision Date	19.04.2024
		Accepted Date	06.05.2024
		Published Date	14.06.2024

1. INTRODUCTION

Recently, significant progress has been made in computer vision thanks to the introduction of deep learning (DL) algorithms. Convolutional Neural Networks (CNNs) have long dominated image processing tasks (Aslan & Özüpak, 2024). They have demonstrated outstanding performance in numerous applications, including photo classification, object identification, and segmentation. Nevertheless, the reliance on CNNs for visual tasks has led researchers to investigate alternative architectures that could provide improvements in effectiveness and efficiency. One viable option that has received much interest is the Vision Transformer (ViT) model. In ViT, spatial features are extracted from images using a transformer-based architecture first introduced for natural language processing tasks, as opposed to typical CNNs that use convolutions. This departure from CNNs represents a paradigm shift in image understanding, providing a novel approach to capture long-range dependencies and semantic relationships within visual data (Berliner et al., 2016).

The basic idea of the ViT model is the self-attention mechanism, which allows the model to capture global contextual information while focusing on relevant areas of the input image. By decomposing the input image into patches and processing them through a sequence of transform blocks, ViT learns to extract hierarchical representations of visual content, effectively leveraging both local and global information for downstream tasks. To extract high-level information from the input image, ViT's architecture consists of multiple layers of self-attention blocks followed by feedforward neural networks. Most importantly, ViT eliminates the need for

*Corresponding Author, e-mail: emrah.aslan@dicle.edu.tr

manually constructed features, allowing the model to self-train on massive datasets and learn discriminative features directly from raw pixel values. Through this comprehensive review, we aim to provide researchers and practitioners with a comprehensive understanding of ViT's capabilities, potential areas for improvement, and future research directions in the evolving landscape of computer vision (Salehinejad et al., 2018).

Vision Transformer (ViT) has significant potential in medical imaging and can be applied to critical health issues such as the diagnosis of pneumonia. Pneumonia is a disease resulting from infection in the lungs and can lead to serious complications. Traditionally, pneumonia is diagnosed by examining chest X-rays. Within these images, specific clues are searched for. However, deep learning models like ViT can be employed to extract information from such images and diagnose the disease. When trained on large datasets, ViT can perceive complex patterns in images and detect symptoms of specific illnesses like pneumonia. For the diagnosis of pneumonia, ViT can extract features from chest X-rays. Based on these features, it can classify the disease. This approach could support the diagnosis process of human doctors or contribute to the development of automated diagnostic systems. However, before implementing such a system, the accuracy and reliability of ViT need to be thoroughly validated. Additionally, strict compliance with regulations and standards for medical devices and applications is necessary. As a result, deep learning models like ViT could play an important role in the early diagnosis and treatment of pneumonia, but careful evaluation is essential (Pacal, 2023).

Image processing refers to a technology that rapidly performs various tasks similar to those performed by the human eye within a computerized environment, using various interface software. A plethora of models have been developed within this domain, with accompanying scientific research contributing significantly. Recently, the predominant model emerging from the analytical results is deep learning, an integral component of machine learning. Deep learning has gained popularity due to its multi-layered design, which distinguishes it from typical machine learning techniques. Its inspiration comes from the complex functioning of the human brain (Koitka & Friedrich, 2016). The focus of deep learning models in the field of image processing includes biomedical applications, where their integration has heralded remarkable successes (Ravi et al., 2017). Throughout history, infectious diseases have emerged as a major threat to human well-being. Pneumonia, medically known as pneumonias, reigns supreme in the hierarchy of infectious diseases (Bakator & Radosav, 2018). Characterized by inflammation of lung tissue due to microbial invasion, pneumonia takes a heavy toll, affecting approximately 7% of the global population annually and resulting in an estimated 4 million deaths (Akter et al., 2015). Timely diagnosis is paramount to mitigating the impact of such diseases. Recognizable symptoms include chest pain, dyspnea, and cough, among others, with diagnostic modalities including sputum cultures and chest radiographs (Berliner et al., 2016).

Much research has attempted to apply various computer vision techniques to the interpretation of human chest X-rays to detect pneumonia. In 2017, Rajpurkar and colleagues presented a framework called "ChexNet" that demonstrated the ability to diagnose pneumonia from chest X-rays with a level of accuracy exceeding that of expert radiologists (Rajpurkar et al., 2017). With over 100,000 frontal view X-rays labeled with 14 diseases, the ChestXray14 dataset serves as the training set for ChexNet, a 121-layer CNN. This repository is currently the largest of its kind in the public domain. In another study, Tatiana Gabruseva and colleagues developed a computational method based on augmentation, multi-task learning, squeeze-and-excitation deep convolutional neural networks, single-shot detectors, and multi-task learning for pneumonia area detection (Gabruseva et al., 2020). When the suggested method was assessed in the context of the Radiological Society of North America Pneumonia Detection Project, it produced good results within the project framework.

For the purpose of analyzing abnormal and normal chest X-rays, Varshni et al. (2019) evaluated the effectiveness of pre-trained CNN models used as feature extractors followed by different classifiers. The best CNN model for this purpose was then analytically determined and the accuracy was 80.02% (Varshni et al., 2019). To diagnose pneumonia more accurately than individual models, Chouhan et al. (2020) proposed an ensemble model that incorporates results from multiple pre-trained models. Their ensemble model achieved excellent recall of 99.62% and accuracy of 96.4% on unknown data from the Guangzhou Women and Children's Medical Center dataset. Salehinejad et al. (2018) used GAN-generated images to improve the original chest X-ray images in order to overcome the absence of medical data. The dataset was subsequently subjected to DCNN, which produced considerable improvements in classification performance (Salehinejad et al., 2018). A text-image embedding network for feature extraction was proposed by Wang et al. (2017) Subsequently, an automatic annotation framework was created, which demonstrated an amazing accuracy of 0.9 (Wang et al., 2017). Toğaçar et al. (2022) used CNN as a feature extractor along with a variety of convolutional neural network models such as VGG-16, VGG-19, and AlexNet. They also used a feature selection approach (mRMR) to reduce combined features (Toğaçar et al., 2020). Using VGG16 with Bi-directional LSTM, Suganya G et al. extracted characteristics from chest X-ray pictures. This resulted in a notable accuracy of 97.76% in tuberculosis detection (Chowdary et al., 2021). With the help of cytological imaging, Guan et al. (2019) were able to discriminate between benign thyroid nodules and papillary thyroid cancer with a reasonable accuracy of 95% in their patient base. In the contemporary landscape of medical practice, reliance on manual interpretation of chest X-rays by clinicians presents a cumbersome process amidst the era of technological advancements. Leveraging extant technological resources and software applications to facilitate diagnosis represents a commendable stride in terms of efficiency and cost-effectiveness. By harnessing deep learning models, particularly in training with chest X-ray images sourced from pneumonia patients, superior diagnostic outcomes can be achieved compared to conventional methodologies. This study, employing publicly available chest X-ray imagery, has harnessed ViT, a prominent deep learning methodology, with ensuing results showcasing its efficacy and proficiency.

2. MATERIAL AND METHOD

A well-liked method in the deep learning family, transformers have shown remarkable results in natural language processing (NLP) applications. Transformers have been used in image processing activities recently, a significant development that was started by Dosovitskiy and associates. Demonstrating success in image processing, transformers have swiftly gained prominence across various domains. Characterized by a simple network architecture based on attention mechanisms, transformers enable focused processing of input elements. Similar to operations in natural language processing, inputs are divided into multiple patches, akin to words, and undergo a series of linear transformations (Dosovitskiy et al., 2021).

In the realm of image processing, inputs are fragmented into multiple patches, resembling words, which are then utilized as input elements. The Vision Transformer (ViT) stands as the pioneering attempt to apply a pure transformer architecture to process images. While the original transformer model encompasses both encoder and decoder components, the ViT model solely incorporates an encoder. The operational principles of image transformers closely parallel those observed in NLP. In the ViT architecture, the input image I is represented as R^{HxWxC} of dimensions. Subsequently, this image is partitioned into N patches of dimensions $PxPxC$. The value of N is mathematically expressed as shown in Equation 1.

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

Flattening and Embedding Process: In this context, H represents the height, P the patch size, W the width, and C the number of channels in the image. The fragmented image patches, divided into N parts, are flattened and subjected to a linear embedding process. Subsequently, a positional embedding process is applied to retain positional information of the patches. The functioning of vision transformers proceeds akin to natural language processing. Three layers typically make up the ViT architecture: the classifier, encoder, and embedding layers. This structure is illustrated in Figure 1.

Embedding Layer: Each patch is handled separately in this layer, and a learnable linear projection is used to map the patches to the embedding dimensions E and D. The embedded projections are combined with a learnable class token U_{class}, which serves as a trainable token completing the classification process. Positional embedding, E_{pos}, tracks the arrangement of each patch, ensuring their maintenance to facilitate the recognition of the actual image. The patch-encoded series, denoted as Z_0, is mathematically expressed as shown in Equation 2 below.

$$Z_0 = \left[U_{class}; X_p^1 E;\ X_p^2 E; \dots \dots X_p^N E\right] + E_{pos} \tag{2}$$

Encoder Layer: The transformer encoder is employed to process the previously obtained series of embedded patches Z_0. In vision transformers, the encoder unit comprises L identical layers. Each identical layer consists of a multi-head self-attention (MSA) block and a fully connected feed-forward dense block (MLP) structure (Equations 3 and 4). In the transformer encoder, the MSA block serves as the fundamental component, incorporating self-attention and merging layers. These blocks consist of the GeLU activation function after two dense layers. Skip connections are used in the encoder, and layer normalization (LN) is used prior to the output layer.

$$Z_1' = MSA(LN)(Z_1 - 1) + (Z_1 - 1), 1 = 1 \dots . L \tag{3}$$

$$Z_1' = MLP(LN)(Z_1') + Z_1', 1 = 1 \dots . L \tag{4}$$

Within the transformer encoder, the output of the multi-head self-attention (MSA) is obtained through the aggregation of several self-attention heads. Mathematically, self-attention is represented as shown in Equation 5.

$$H = Atten(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{D_K}}\right)V \tag{6}$$

In Equation 5, the query obtained after matrix multiplications is represented by Q, the key by K, and the value matrix by V. In vision transformers, the final output of the MSA is obtained by passing the concatenation of all self-attention heads through a linear layer. This linear layer is mathematically expressed in Equation 6.

$$MSA(Q, K, V) = [H_1, H_2, \dots \dots . . H_h]W_0 \tag{7}$$

Where, W_0 represents the learnable output transformation matrix, while H denotes the number of self-attention heads.

2.1. Vision Transformer Approach

The Vision Transformer (ViT) technique is now a helpful tool for work involving NLP. In the area of computer vision, ViT offers a pure transformer model devoid of any convolutional blocks (Vaswani et al., 2017). Convolutional Neural Networks (CNN) have historically dominated image recognition efforts. Nevertheless, CNNs have several drawbacks. Most notably, because of processes like max pooling, they process information more slowly, and large datasets are required for efficient processing and neural network training (Zhou et al., 2021). The suggested model uses a dataset of chest X-rays to classify pneumonia using the Vision Transformer (ViT) approach. Notably, when it comes to managing massive computer vision datasets, the Vision Transformer has lately gained favor over CNNs. Data integration across the full image is made possible by ViT's usage of a transformer architecture with self-attention. The operating principle of ViT is visualized in Figure 1.

As required by the algorithm, the image is split up into tokens, or patches, of similar size. These patches go through a 2D flattening process to become a vector format. The patch embedding is then combined with a position embedding to preserve positional information. Layers for self-attention and multi-head attention make up the transformer encoder. The multi-layer perceptron blocks are connected to a second layer normalization, and the embedded patches are connected to layer normalization within the multi-head arrangement. Each X-ray image was resized to a consistent pixel dimension of 250 by 250. Subsequently, each image was partitioned into 25 identical patches, each with dimensions of 50 by 50 pixels. These patches were vectorized and flattened in order to be input into the transformer encoder network, where positional encoding was applied to the picture vectors. Among the six transformer blocks, the multi-head attention layer utilized eight heads.

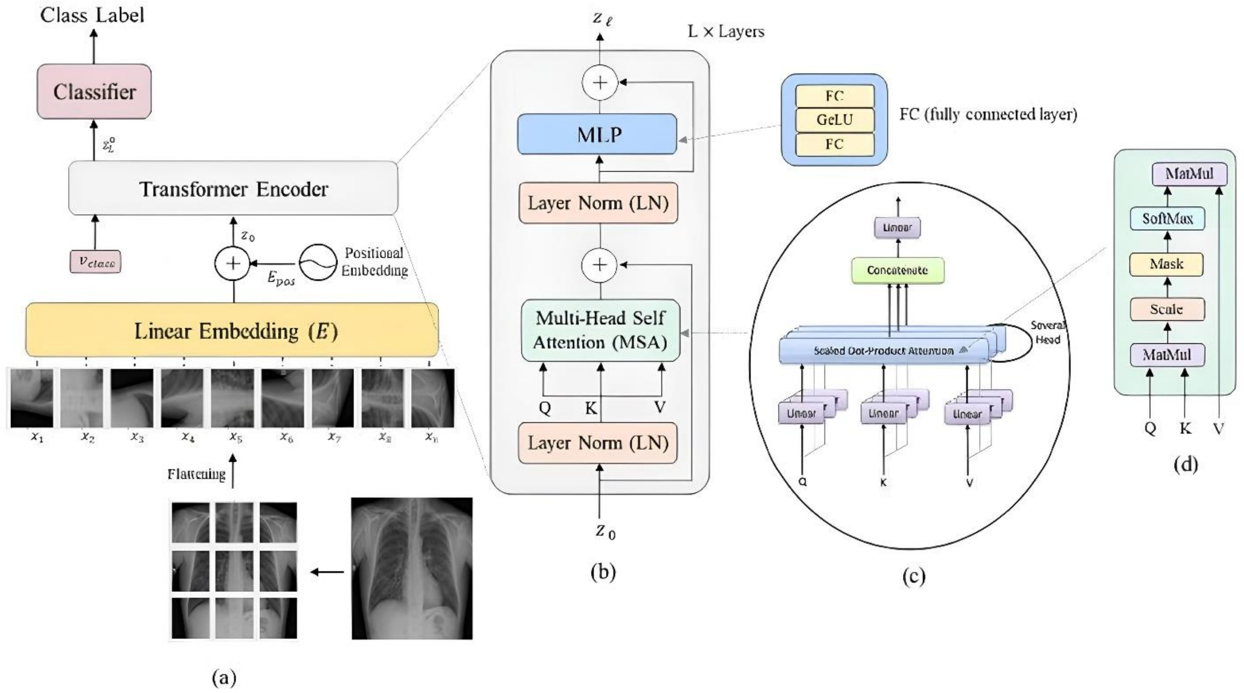

Figure 1. *Overview of the design of the vision transformer*

2.2. Classification layer

Within this unit dedicated to the classification process, the entity Z_l^0 is initially retrieved and subsequently input into an external auxiliary head classifier to forecast the ultimate layer of the encoder for classification. This procedure is formally delineated in Equation 7, wherein y symbolizes the model's output and Z_l^0 denotes the initial element utilized for decision-making (Pacal, 2023).

$$y = layer\_normalization\ (Z_l^0) \tag{8}$$

In this study, three different transformer-based models were applied for pneumonia diagnosis from chest X-ray images. Table 1 displays each vision transformer model's specifics. The ViT-B model consists of two different models, ViT-B16 and ViT-B32. Actually, the ViT-B model, or the base model, is obtained by changing the patch sizes to 16x16 or 32x32, resulting in two models, but there is no change in the number of layers, which remains at 12. Similarly, in the ViT-L, or large model, ViT-L16 and ViT-L32 models are obtained by changing the number of patches. However, the ViT-H model, or the larger model, was not used in the study due to its unavailability on the RTX-2080-TI graphics card. Despite selecting the lowest batch size of 1, the GPU proved inadequate for operation. As seen in Table 1, the MLP size increased from 3072 in the base model to 5120 in the high model. It is well known that larger models provide more effective results, especially with large-scale data.

2.3. Dataset

The dataset is divided into three subdirectories, one for each of the three photo categories (Pneumonia/Normal): train, test, and validation. It includes 5,863 X-ray images (JPEG) in total, divided into two categories: pneumonia and normal. Anterior-posterior chest X-ray pictures were selected from a retrospective cohort of pediatric patients (aged one to five) from Guangzhou Women and Children's Medical Center in Guangzhou. The patients' regular clinical care included the acquisition of these chest X-rays. Prior to commencing the analysis of chest X-ray images, all radiographs underwent meticulous quality control, wherein any scans deemed of subpar quality or unreadable were eliminated. Subsequently, two board-certified medical experts evaluated the images prior to training the AI system for diagnosis. A third specialist validated the assessment set to ensure absence of grading discrepancies. The dataset's specifics are depicted in Figure 2.

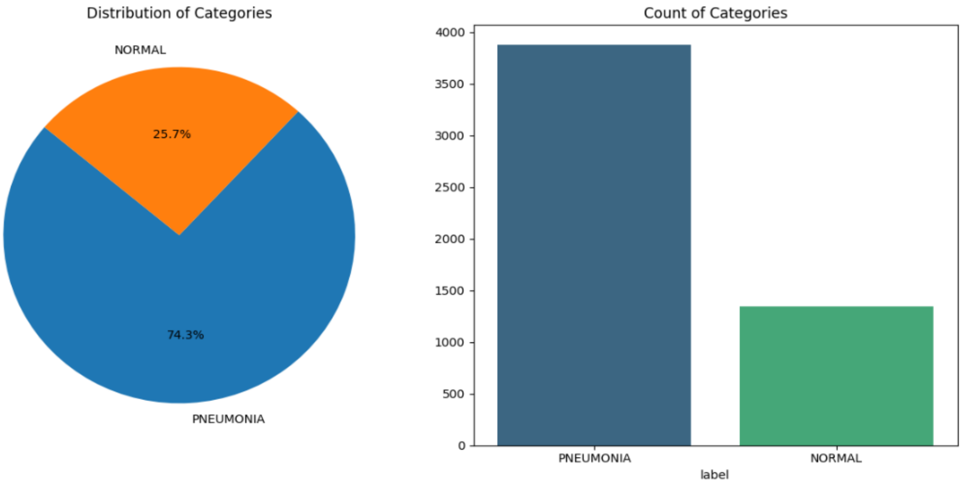

Figure 2. *Representation of values in the dataset*

In the pneumonia dataset, the training and evaluation images of the deep learning approaches come pre-separated. The main advantage of such a situation is that the study is easily comparable to other studies and the actual performance of the model can be measured. The dataset is an extended publicly available dataset consisting of images from different datasets. Train, Test, and Validation are the three primary folders into which the dataset is arranged. Pneumonia and Normal are two subfolders that correspond to different image categories in each of these directories. There are 5,863 JPEG-formatted X-ray images in total, divided into two groups: normal and pneumonia. Here, the class distributions in the training, validation and test data are well-adjusted and there is no data imbalance. Thus, the model will try to learn each class in the dataset in a better way, and the bias towards any one class will be reduced. Some random sample images from the pneumonia dataset are shown in Figure 3.

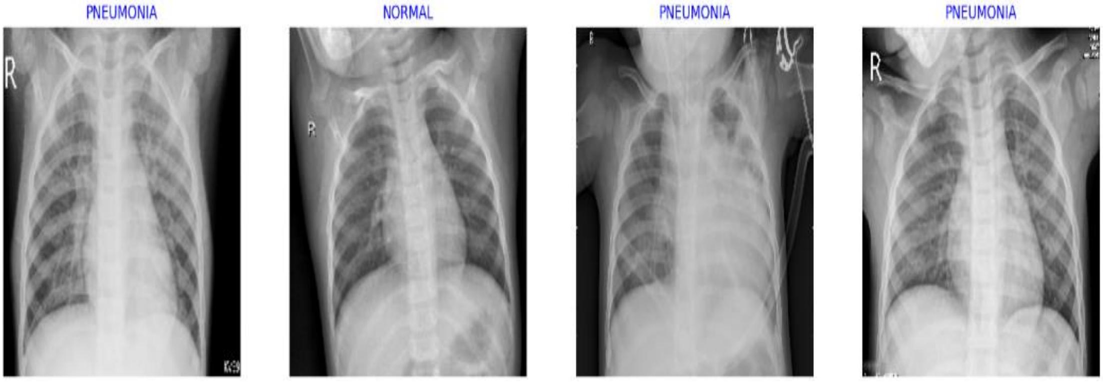

Figure 3. *Random sample images from the pneumonia dataset.*

2.4. Proposed Method

There are essentially three parts to the suggested technique for automatically detecting pneumonia. Figure 4 displays the primary parts of the suggested system. By using the achievements of deep learning techniques in medical image processing to chest X-ray pictures, a more efficient system is suggested. The first part of this deep learning based system, which consists of three parts, is the data set unit. The next unit after the dataset is the unit where data pre-processing and data augmentation techniques are combined. This unit's primary goal is to use some fundamental data augmentation techniques and shrink all of the dataset's photos to the same size. The most fundamental data augmentation methods, including translation, rotation, and panning, were used during training as there are enough images in the dataset. For large-scale datasets, data augmentation is not very effective, but it contributes to the performance, while for small-scale and less diverse datasets, data augmentation is very effective. The last unit of the proposed method is the unit where deep learning approaches are used. This unit includes learning transfer and a vision transformer for classification. The technique of moving the weights of a model trained in one domain to another is known as transfer learning or learning transfer. For the pneumonia dataset in this study, the weights of the vision transformers trained on the ImageNet dataset were utilized. Significant performance is obtained by learning transfer, particularly with small-scale datasets. In this dataset, it resulted in both better performance and faster convergence than training from scratch. After learning transfer, we used vision transformers for classification. This architecture is described in detail in the materials and methods section.

2.5. Evaluation Metrics

Important measures including accuracy, F1 score, precision, and recall are used to evaluate the suggested model. Four parameters are used in the calculation of these metrics: true positive (TP), false negative (FN), false positive (FP), and true negative (TN). The following definitions apply to the parameters in these equations: Examples of data that the model correctly categorized as positive are called True Positives (TP). Examples of negative outcomes that the model correctly detects are called True Negatives (TN). False positives (FP) are instances in which something was mistakenly categorized as positive by the model. Examples of negative numbers that the model misinterpreted are called False Negatives (FN) (Özüpak, 2024).

These indicators provide a comprehensive evaluation of the model's efficacy, accounting for both positive and negative classifications. The F1-score finds a compromise between precision and recall, even though both indicate details about the model's capacity to accurately identify positive and negative samples, respectively. The accuracy measure gives an overall evaluation of the model's prediction accuracy.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{9}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{10}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{11}$$

$$F_1 = 2 * \frac{Precision * Recall}{(Precision + Recall)} \tag{12}$$

3. RESULTS AND DISCUSSION

The model's initial values were used for training and evaluation in the experimental investigation. Figure 4 shows the confusion matrix for the study's model. The expected class is shown on the x-axis of the confusion matrix, and the actual class is shown on the y-axis. The confusion matrix allows us to see both true and false positives and negatives for each class, which improves our ability to observe. The model's experimental findings are shown in Table 1.

In the Figure 4, the confusion matrix shows that the model correctly classified 208 patients as normal (TP) and 389 patients as pneumonia (TN). The model misclassified 26 patients as pneumonia (FP) and 1 patient as not pneumonia (FN). The overall test accuracy of the model was 95.67%.

Table 1. *Metrics showing the performance of the trained model on the test data*

Model	$Accuracy$	$Recall$	$Precision$	F_1 $Score$
Proposed Model	0.9567	0.9952	0.8888	0.9390

This table shows that the accuracy rate of the suggested model is 95.67%. This indicates that the model makes accurate predictions in general. The recall rate is quite high at 99.52%, indicating that the model can effectively recognize positive class instances. However, the Precision rate is 88.88%, which indicates how many of the cases that the model predicts as positive are actually positive. If the precision is lower than the true positive rate, it may indicate that the model is making false positive predictions and classifying some negative samples as positive. The F1 score was calculated as 0.9390. In this instance, the model's success is gauged by the F1 score, which illustrates the trade-off between precision and the capture rate of true positives. These results show that the model performs well overall but could be improved in specific cases. A comparison of the suggested model with a few findings from the literature is presented in Table 2.

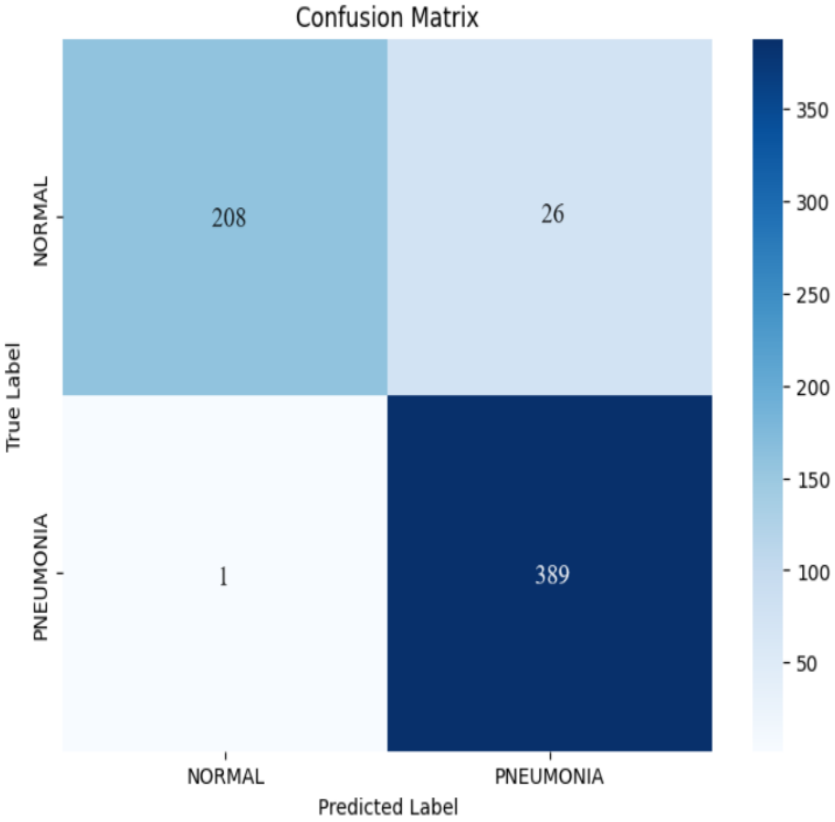

Figure 4. *Confusion matrix for the suggested architecture*

Table 2. *Metrics showing the performance of the trained model on the test data*

Model	$\boldsymbol{Accuracy}$	$\boldsymbol{F_1\ Score}$
(Hassan, 2018)	0.9214	0.9234
(Dey et al., 2021)	0.9012	0.8999
(Singh & Tripathi, 2022)	0.9375	**0.9405**
(Singh & Tripathi, 2022)	0.8814	0.8812
Proposed Model	**0.9567**	0.9390

According to Table 2, we analyzed the performance of five different models and obtained accuracy and F1 score values for each of them, including the proposed model. In this work, we assessed how well several machine learning models performed in diagnosing pneumonia. First, we analyzed the results obtained for Model (Hassan, 2018), where we obtained an accuracy of 92.14% and an F1 score of 92.34%. Although this model performed well overall, we observed that it had a lower F1 score compared to the other models. Model (Dey et al., 2021) performed slightly lower than Model (Hassan, 2018) with an accuracy of 90.12% and an F1 score of 89.99%. We observed that this model was less accurate in diagnosing pneumonia and tended to make more false positive and negative classifications. However, Model (Singh & Tripathi, 2022) did rather well, scoring an F1 score of 94.05% and an accuracy of 93.75%. It is noteworthy that this model is quite effective in diagnosing pneumonia, despite its high accuracy rate and F1 score. When we examined the results obtained for model (Singh et al., 2022), we obtained an accuracy rate of 88.14% and an F1 score of 88.12%. We observed that this model performs poorly compared to the other models and is less reliable in pneumonia diagnosis. Finally, our proposed model performed the best with an accuracy rate of 95.67% and an F1 score of 93.90%. This model provided more reliable results for pneumonia diagnosis with a higher accuracy rate and a balanced F1 score compared to the other models. These findings demonstrate how well the suggested model diagnoses pneumonia in comparison to the other models. Therefore, we suggest that the proposed model may be a more effective tool for pneumonia diagnosis in clinical applications. While this analysis evaluates the performance of each model separately, it highlights that the proposed model achieves a higher accuracy and F1 score than the others. This indicates that the proposed model is a more reliable option for diagnosing pneumonia.

Figure 5 displays the model's accuracy following testing and training. Figure 6 provides several instances comparing the values predicted by the model with the actual values in the dataset.

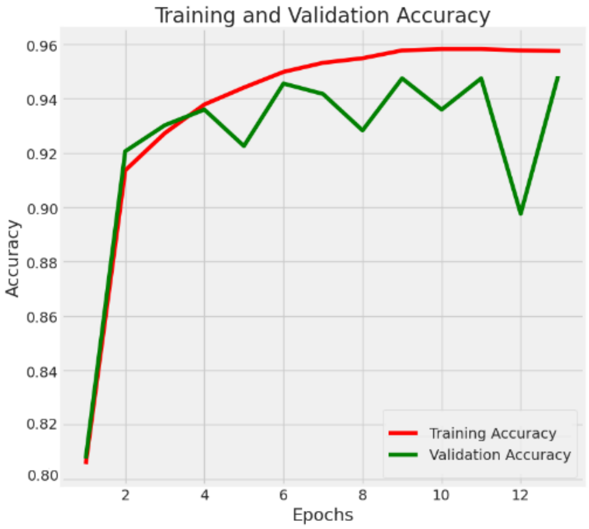

Figure 5. *Training and validation accuracy results*

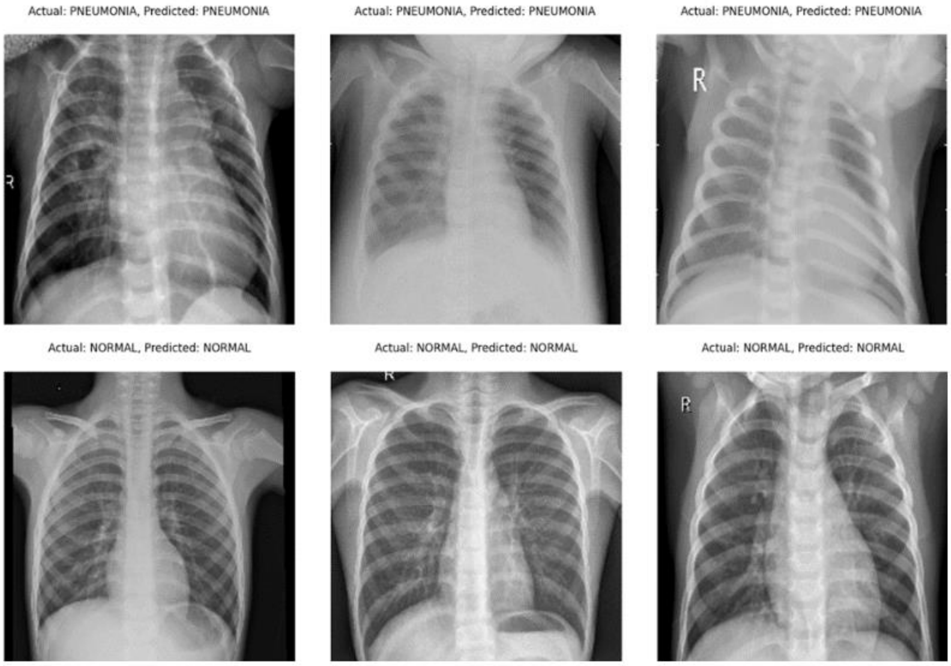
Actual: PNEUMONIA, Predicted: PNEUMONIA
Actual: PNEUMONIA, Predicted: PNEUMONIA
Actual: PNEUMONIA, Predicted: PNEUMONIA
Actual: NORMAL, Predicted: NORMAL
Actual: NORMAL, Predicted: NORMAL
Actual: NORMAL, Predicted: NORMAL

Figure 6. *Some examples showing actual and predicted values*

4. CONCLUSION

In this work, we evaluate the pneumonia diagnosis performance of a DL model. The tremendous potential of transfer learning in the field of medical imaging is demonstrated by the use of the ViT for the classification of chest X-rays. ViT, which has its origins in NLP, has successfully transferred to computer vision through self-supervised learning, offering a workable method for determining the presence of pneumonia from X-rays. The versatility of the Vision Transformer is shown by its distinct method of segmenting images for processing. Its fundamental design, which was first created for NLP tasks, has been skillfully modified for challenging computer vision tasks. A steady reduction in training loss is seen during the training period. This demonstrates how well the model learns from the data and adjusts its parameters to minimize errors. Not only could the model learn, but it also showed excellent generalization abilities. The 95.67% test accuracy indicates that the model performs well. A detailed examination of the confusion matrix demonstrated the model's excellent ability to distinguish between chest X-rays labeled as "Pneumonıa" and "Normal." The low rate of misclassifications strengthened trust in the model's predictions. Its excellent accuracy, versatility, and capacity for self-supervised learning bode well for its further applications in medical imaging.

CONFLICT OF INTEREST

The author declares no conflict of interest.

REFERENCES

Akter, S., Shamsuzzaman, & Jahan, F. (2015). Community Acquired Pneumonia. *International Journal of Respiratory and Pulmonary Medicine*, *2*(1). http://doi.org/10.23937/2378-3516/1410016

Aslan, E., & Özüpak, Y. (2024). Classification of Blood Cells with Convolutional Neural Network Model. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, 13(1), 314–326. https://doi.org/10.17798/bitlisfen.1401294

Bakator, M., & Radosav, D. (2018). Deep Learning and Medical Diagnosis: A Review of Literature. *Multimodal Technologies and Interaction*, *2*(3), 47. https://doi.org/10.3390/mti2030047

Berliner, D., Schneider, N., Welte, T., & Bauersachs, J. (2016). The differential diagnosis of dyspnoea. *Deutsches Ärzteblatt International*, *113*(49), 834–844. https://doi.org/10.3238%2Farztebl.2016.0834

Chouhan, V., Singh, S. K., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., Damaševičius, R., & de Albuquerque, V. H. C. (2020). A novel transfer learning based approach for pneumonia detection in chest X-ray images. *Applied Sciences*, *10*(2), 559. https://doi.org/10.3390/app10020559

Chowdary, G. J., Suganya, G., Premalatha, M., & Karunamurthy, K. (2021). Class dependency-based learning using Bi-LSTM coupled with the transfer learning of VGG16 for the diagnosis of tuberculosis from chest X-rays. In: M. Sabharwal, B. B. Balusamy, S. R. Kumar, N. Gayathri, & S. Suvanov (Eds.), *Applications of Artificial Intelligence in E-Healthcare Systems*, (pp. 37-54). https://doi.org/10.1049/PBHE040E_ch3

Dey, N., Zhang, Y.-D., Rajinikanth, V., Pugalenthi, R., & Raja, N. S. M. (2021). Customized VGG19 architecture for pneumonia detection in chest X-rays. *Pattern Recognition Letters*, *143*, 67–74. https://doi.org/10.1016/j.patrec.2020.12.010

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2021, May 3-7). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: Proceedings of the Ninth International Conference on Learning Representations (ICLR 2021), (pp. 1-21). https://iclr.cc/virtual/2021/poster/3013

Gabruseva, T., Poplavskiy, D., & Kalinin, A. (2020, June 14-19). *Deep Learning for Automatic Pneumonia Detection*. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (pp. 1436-1443), Seattle, WA, USA. https://doi.org/10.1109/CVPRW50498.2020.00183

Guan, Q., Wang, Y., Ping. B., Li, D., Du, J., Qin, Y., Lu, H., Wan, X., & Xiang, J. (2019). Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study. *Journal of Cancer*, *10*(20), 4876–4882. https://doi.org/10.7150/jca.28769

Hassan, M. (2018, November 20). VGG16–Convolutional Network for Classification and Detection. https://neurohive.io/en/popularnetworks/vgg16

Koitka, S., & Friedrich, C. M. (2016, September 5-8). *Traditional feature engineering and deep learning approaches at medical classification task of imageCLEF 2016*. In: K. Balog, L. Cappellato, N. Ferro, & C. Macdonald (Eds.), Proceedings of the Conference and Labs of the Evaluation Forum (vol. 1609, pp. 304-317), Évora, Portugal.

Özüpak, Y. (2024). Detection of Malaria with Convolutional Neural Network (CNN) Architectures Using Cell Images. *Cukurova University Journal of the Faculty of Engineering*, *39*(1), 197-210. https://doi.org/10.21605/cukurovaumfd.1460434

Pacal, I. (2023). A Vision Transformer-based Approach for Automatic COVID-19 Diagnosis on Chest X-ray Images. *Journal of the Institute of Science and Technology*, *13*(2), 778–791. https://doi.org/10.21597/jist.1225156

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Ball, R. L., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. https://doi.org/10.48550/arXiv.1711.05225

Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G.-Z. (2017). Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 4-21. https://doi.org/10.1109/jbhi.2016.2636665

Salehinejad, H., Valaee, S., Dowdell, T., Colak, E., & Barfett, J. (2018, April 15-20). *Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks*. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 990–994), Calgary, AB, Canada. https://doi.org/10.1109/ICASSP.2018.8461430

Singh, S., & Tripathi, B. K. (2022). Pneumonia classification using quaternion deep learning. *Multimedia Tools and Applications*, *81*, 1743–1764. https://doi.org/10.1007/s11042-021-11409-7

Toğaçar, M., Ergen, B., Cömert, Z., & Özyurt, F.(2020). A deep feature learning model for pneumonia detection applying a combination of mRMR feature selection and machine learning models. *IRBM*, *41*(4), 212–222. https://doi.org/10.1016/j.irbm.2019.10.006

Varshni, D., Thakral, K., Agarwal, L., Nijhawan, R., & Mittal, A. (2019, February 20-22). *Pneumonia Detection Using CNN based Feature Extraction*. In: Proceedings of the IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1–7), Coimbatore, India. https://doi.org/10.1109/ICECCT.2019.8869364

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł, Polosukhin, I.(2017, December 4-9). *Attention is all you need*. In: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach CA. https://doi.org/10.48550/arXiv.1706.03762

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017, July 21-26). *ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3462-3471). https://doi.org/10.1109/CVPR.2017.369

Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., & Feng, J. (2021). Deepvit: Towards deeper vision transformer. https://doi.org/10.48550/arXiv.2103.11886