IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ENHANCING PNEUMONIA DETECTION USING VISION TRANSFORMERS

[1]Greeshma N, [2]Leela P, [3]Niraja K, [4]Sai Mohana R, [5]Amar Tej G

[1]Student, [2]Student, [3]Student, [4]Student,[5]Assistant Professor

[1]Electronics and Communication Engineering,
[1]Vasireddy Venkatadri Institute of Technology, Guntur, India

Abstract: Pneumonia is a leading cause of illness and mortality worldwide, especially among elderly populations and children aged below five. Every year, an estimated two million deaths were caused by pneumonia, with children aged below five years contributing twenty-three percent of these deaths. India accounts twenty-three percent of all pneumonia cases worldwide. In order to lower the death rate, early and correct diagnosis is important, especially in situations where there is limited access to healthcare services. In low-resource settings, interpretation of chest X-rays, a standard diagnostic tool for pneumonia, is mostly left to radiologists. We present a Vision Transformer system that uses Adam optimizer to enhance efficiency and accuracy for the automatic detection of pneumonia from chest X-ray images.

Keywords - pneumonia, Vision Transformers, chest X-rays, CT scans, Adam optimizer.

I. INTRODUCTION

Pneumonia is a respiratory infection that inflames the air sacs in the lungs. Pneumonia can be caused by bacteria, viruses, fungi, or inhaling foreign substances [1]. Bacterial pneumonia, caused by bacteria like *Streptococcus pneumoniae*, is often severe and sudden. Viral pneumonia, caused by viruses like flu or RSV, typically results in milder symptoms but can become severe. Fungal pneumonia, caused by fungi, mainly affects people with weak immune systems[2]. Pneumonia begins with mild symptoms like a cough, sore throat, and fatigue. As the condition progresses, the lungs become inflamed, and the air sacs fill with fluid or pus, leading to a worsening cough, fever, and chest pain. This inflammation reduces the lungs' ability to exchange oxygen, resulting in low blood oxygen levels and causing extreme fatigue, frequent coughing, and difficulty breathing. Without timely treatment, pneumonia can lead to severe complications like lung failure.

To address these challenges, there has been a growing interest in using artificial intelligence to automate the analysis of chest X-rays for pneumonia detection. Among several deep learning models, Convolutional Neural Networks (CNNs) have been widely applied with success. These models excel at identifying features in medical images, making them effective for pneumonia detection. However, CNNs have limitations when dealing with long-range dependencies in image data. Recent advancements in Vision Transformers (ViTs) offer an innovative solution to these limitations, providing enhanced performance in image classification tasks by capturing global context through self-attention mechanisms [3].

This study proposes a ViT-based system for automating pneumonia detection from chest X-ray images. The system utilizes the Adam optimizer for better accuracy. By automating the process, our system aims to provide a fast, more accurate, and scalable alternative to traditional manual image interpretation, particularly in regions with limited access to trained radiologists.

II. LITERATURE REVIEW

The paper "Pneumonia Detection using Deep Learning" by Swapnil Singh [4] proposes an automated method for detectig pneumonia via chest X-rays.The author comapres two deep learning models, Convolutional Neural Networks (CNN) and Multi-Layer Perceptron (MLP),to find out which one works better. Using a Kaggle dataset, the researchers trained both models and developed an easy-to-use GUI where users can input chest X-ray images to predict pneumonia. The findings indicated that CNN significantly performed better than MLP, reaching upto 92.63% accuracy as opposed to MLP's 77.56%. Though the work illustrates the use of deep learning in the identification of pneumonia, it is limited in certain aspects, that is, failure to explore further into more complex models that might increase performance further.

The paper "Pneumonia Detection Using Enhanced Convolutional Neural Network Model on Chest X-Ray Images" by Shadi A. Aljawarneh and Romesaa Al-Quraan [5] investigates deep learning models for detecting pneumonia through chest X-rays. The researchers used an Enhanced Covolutional Neural Networks (CNN), VGG-19, ResNet-50 models trained and tested on 5,863 images from Kaggle. Results indicated that Enhanced CNN had the highest accuracy (92.4%), whereas ResNet-50 had the lowest (82.8%), making Enhanced CNN the best model to detect pneumonia. Additionally, while the dataset size is substantial, it is crucial to consider the potential impact of its diversity and representativeness across different demographic groups.

The paper "Pneumonia Detection Using Convolutional Neural Networks (CNNs)" by V. Sirish Kaushik, Anand Nayyar, Gaurav Kataria, Rachna Jain, and Bhagwan Parshuram Institute of Technology [6] proposes CNN-based models for detecting pneumonia in children under five using chest X-ray images. The authors experimented with four different CNN models, each model consists of one to four convolutional layers, achieving accuracies ranging from 85.26% to 92.31%. The study's evaluation metrics, including recall and F1 scores, provide a comprehensive understanding of the model's performance in addition to accuracy. The limitation of this research is the relatively low accuracy achieved by the simpler models (89.74% and 85.26%), which suggests that a more complex network architecture may be necessary for optimal results.

The study "Pneumonia Detection by Analyzing X-ray Images Using MobileNet, ResNet Architecture, and Long Short-Term Memory (LSTM)" by Md. Sabbir Ahmed, Rafeed Rahman, and Shahriar Hossain [7] explores deep learning-based automation for pneumonia detection using chest X-rays. The research highlights the inconsistencies in manual diagnosis due to human error and proposes ResNet 101, MobileNet, and LSTM to enhance accuracy. Using 5,856 X-ray images, the study achieved a maximum accuracy of 93.2% with the LSTM model, demonstrating its effectiveness. Additionally, while LSTM showed superior performance, the computational cost of training deep learning models such as ResNet and LSTM may pose challenges for practical implementation in resource-limited healthcare environments.

Siddiqi and Javaid's [8] thorough analysis of DL methods emphasises the increasing significance of CNNs for CXR based pneumonia identification. CNNs work well because they can automatically extract features from picture data, which eliminates the need for human input. Pre trained models like DenseNet121 and ResNet are widely used, however, research is increasingly concentrating on more recent designs, such as Vision Transformers (ViTs), which show promise in spite of drawbacks including adversarial assaults and dataset biases.

III. OBJECTIVE

Detecting pneumonia is crucial for early diagnosis, but current techniques have limitations in accuracy, often leaving a performance gap of approximately 7%. These challenges stem from the inability of existing models to capture the full context of the image, leading to errors in detection, particularly in resource- constrained settings. Furthermore, the reliance on traditional CNNs limit model's ability to effectively process patterns which are complex in medical imaging. This gap in accuracy highlights the need for improvised methods to aid in fast, more accurate pneumonia detection, ensuring reliable and accessible diagnostic support for healthcare professionals.

IV. PROPOSED SYSTEM

To address the limitations of existing pneumonia detection methods, our project leverages the Vision Transformer (ViT), an advanced deep learning architecture known for its superior performance in image classification. The model will be trained on a labeled dataset of chest X-ray images, enabling it to learn intricate patterns and contextual features which are needed for accurate diagnosis. After training, the model will be capable of predicting whether a given X-ray image indicates a normal or pneumonia-affected condition, potentially improving diagnostic accuracy and efficiency.

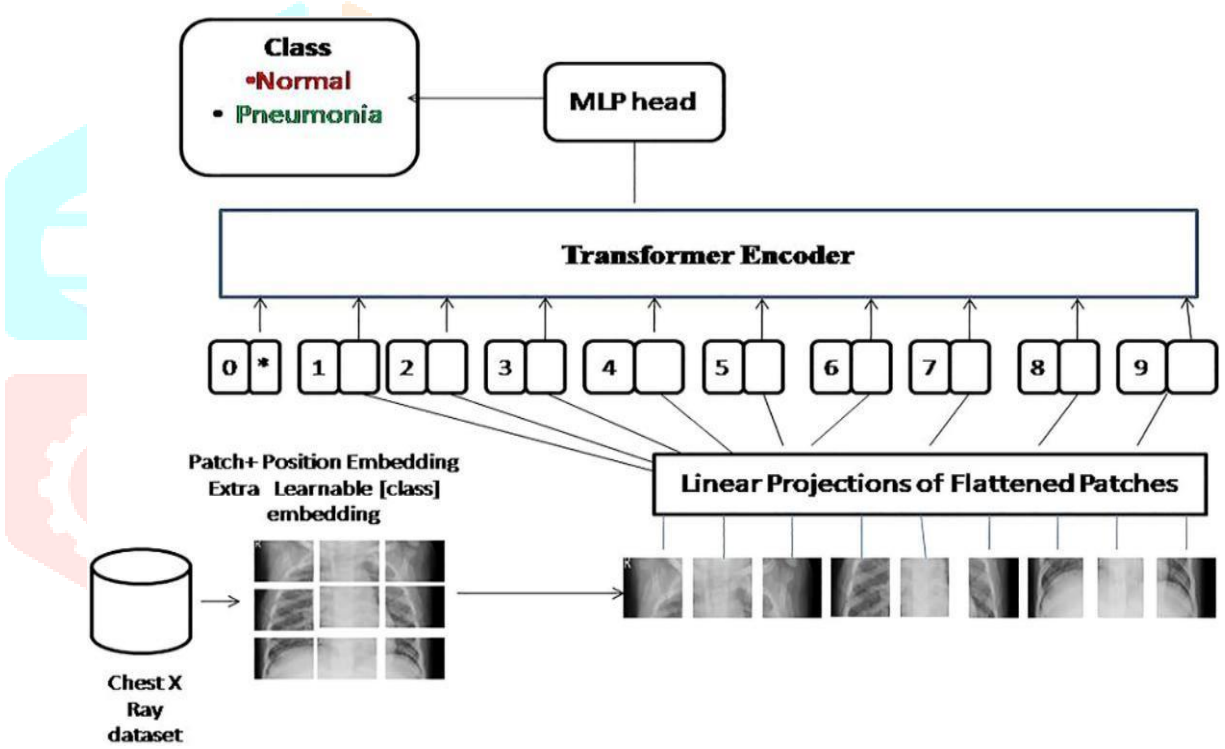


Fig 1. Proposed Architecture [9]

A. Dataset Used

For our research, we conducted a thorough search online to find a suitable dataset of chest X-rays related to pneumonia. After careful evaluation, we selected the 'Chest X-ray Images (Pneumonia)' dataset from Kaggle [10], as it best fit our requirements. This dataset specifically consists of chest X-rays of children, making it highly relevant since 23% of pneumonia deaths are of children under the age five.

The dataset is organized into two main categories: one containing chest X-ray images of healthy children and the other consisting of X-rays of lungs affected by pneumonia.F or training, we used 1,341 images of healthy lungs and 3,875 images of pneumonia-affected lungs. Similarly, the test dataset follows the same structure, with 200 images of healthy individuals and 250 images of pneumonia cases.

TABLE 1 : Dataset Distribution

	Normal Images	**Pneumonia Images**
Training	1343	3875
Testing	200	250

B. Methodology

The methodology of our project is to systematically enhance the pneumonia detection using Vision Transformers (ViT). To begin with, we utilized publicly available chest X-ray dataset from kaggle, which contains labeled images categorized as either Normal or Pneumonia. The dataset was divided into training and testing sets and prior to model training, all images were resized to a uniform resolution of 512x512 pixels. Image augmentation techniques like horizontal flipping, zooming and rotations were used for improving the generalization of model.

The size of each image patch is 32x32 so we will get (512/32)*(512/32)=256 patches for a single image. These patches are then converted into 1D vectors where the size of each patch is calculated by Patch_Height×Patch_Width×Colour_Channels and in our project the patch height and width both are 32 and the x-rays are gray in colour. So, the size of 1D vector is 32x32x1=1024. The final shape of the image is 256 patches x 1024=[256,1024] and these patches are fed to linear projection layer where they are converted into low dimensional vectors, which consists of important features and removes noise. The positions of the patches are also being added to their corresponding low dimensional vectors. These positional embeddings improve the accuracy of the model by 3%.

Vision Transformers consists of Encoder part of the transformer model and there can be multiple encoders stacked up in encoder block and every encoder consists of Multi Head Self Attention Layer and Feed Forward Neural Network. For our project we used 4 encoder layers. Self-Attention is the heart of vision Transformers. Self-attention helps the model to look at each patch and figure out how each patch is related to other patches. Self-attention processes the entire image globally, making it better at identifying patterns spread across different regions. The Multi-head attention (MHA) uses multiple self-attention heads, where each head is looking the input from different angles at the same time and when all heads are combined, we can get a richer understanding.

The Feedforward Neural Network (FFN) in ViT is a fully connected layer that operates on the output of the self-attention mechanism. After self-attention, the representation of each patch goes through a position-wise feedforward network to learn non-linear combinations of the features. Without non-linearity, the network would only be able to learn linear patterns, such as simple thresholds or basic separations between classes which is not sufficient for classification with non-linearity, the network can learn more complex patterns like the specific shape of pneumonia-related lesions, subtle differences in lung structure,and more, which are critical for accurate diagnosis. These two sub-layers are applied parallel to the input sequence and then combined to generate the encoder layer's output. The process is repeated multiple times to form a stack of encoder layers, where each encoder layer builds upon the representation learned by the preceding encoder layer, enabling the model to learn increasingly complex and generalized representations of the input sequence.

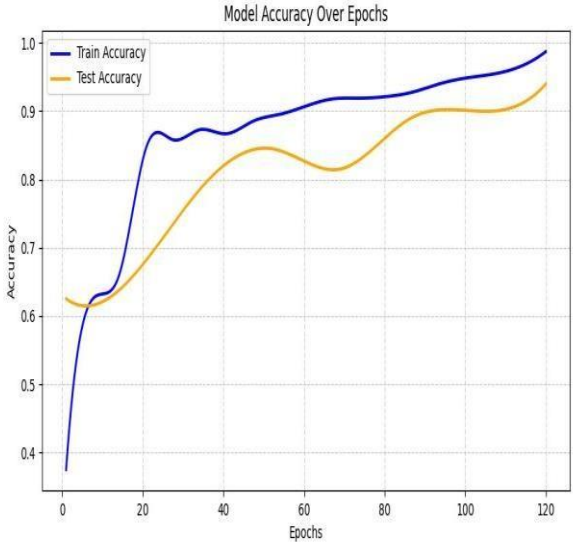
To train the model efficiently, an optimizer is required to minimize the loss function by adjusting the model's internal weights during backpropagation. In our project, we used the Adam optimizer (Adaptive Moment Estimation), which is well-suited for training deep networks like Vision Transformers. Adam combines the advantages of two other popular optimizers momentum and RMSProp by computing adaptive learning rates for each parameter. It maintains moving averages of both the gradients and their squared values, enabling faster and more stable convergence. This makes Adam especially effective for handling sparse gradients and large parameter spaces, ensuring that the model learns efficiently and generalizes well on the pneumonia classification task.

An MLP (Multilayer Perceptron) head is used at the end to make the final classification. The features extracted by the model will be stored in MLP head and it acts as the summary of the model training and is sufficient to make the final classification of images and it passes through SoftMax classifier which outputs a vector p with probabilities, whose size is equal to the number of output classes and the class with highest probability is taken as the final output.

V. RESULTS AND DISCUSSION

- ***Accuracy and Loss plots:***

The accuracy and loss plots demonstrate the effectiveness of our model trained over 120 epochs. In Fig. 2 both training and testing accuracy show a steady increase after some lows, indicating the model learnt properly without overfitting. In Fig. 3 (Loss vs. Epochs), the loss starts at a high value and gradually decreases, confirming that the model is learning effectively. The smooth decline in both training and testing loss further supports the stability and convergence of the model.



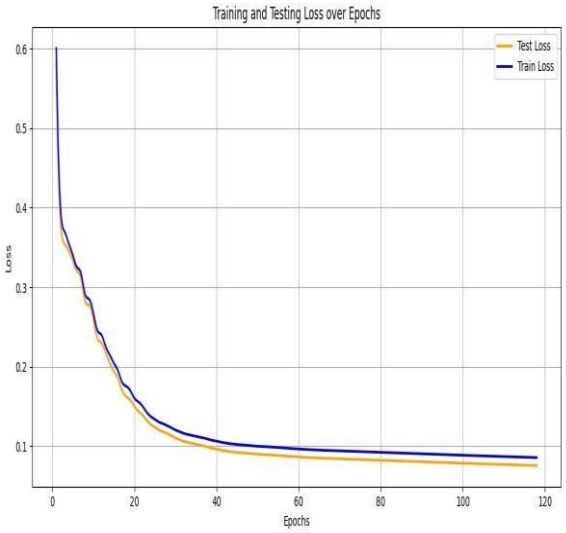


Fig. 2. Accuracy plot

Fig. 3. Loss plot

- ***Confusion Matrix***

A confusion matrix is a table used to evaluate the performance of a classification model by comparing actual and predicted values. It consists of four key metrics: True Positives (TP), where the model correctly predicts pneumonia cases (244 in this case); True Negatives (TN), where healthy patients are correctly classified as normal (186); False Positives (FP), where healthy patients are incorrectly classified as having pneumonia (14); and False Negatives (FN), where pneumonia cases are mistakenly predicted as normal (6). This matrix helps assess the model's accuracy, precision, recall, and overall effectiveness in distinguishing between pneumonia and healthy cases.

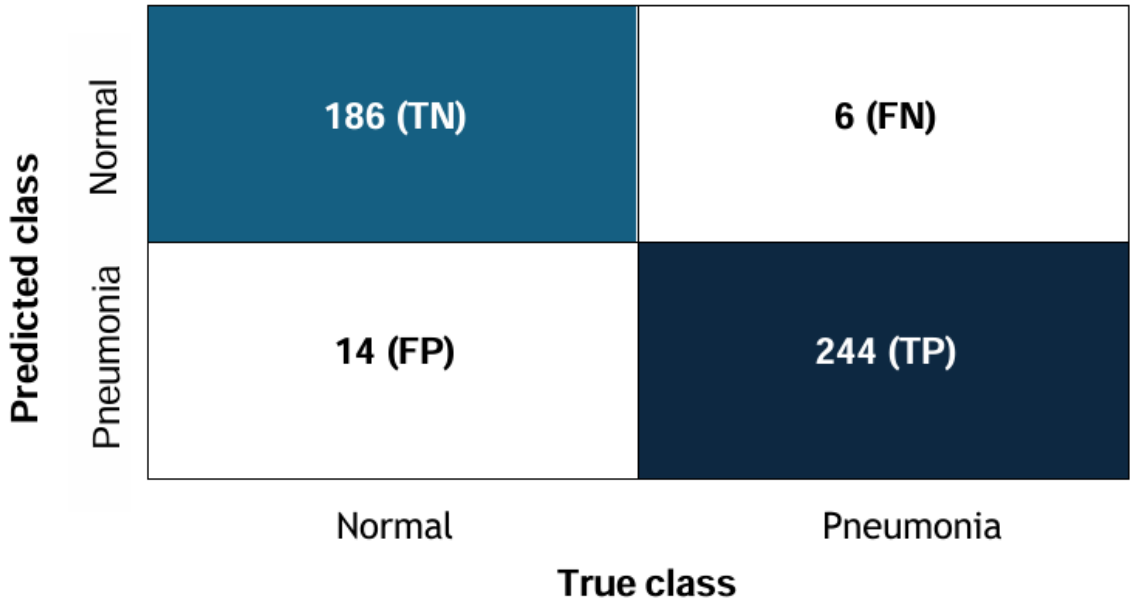


Fig 4. Confusion Matrix

- ***Evaluation Metrics:***

From confusion matrix, we know that TP=244, TN=186, FP=14, FN=6. From this we can calculate:

1) Accuracy: Accuracy measures the overall correctness of the model's predictions. It is the ratio of correctly classified instances to the total number of instances.

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

Accuracy: 0.955 or 95.5%

2) Precision: Precision indicates how many of the predicted pneumonia cases are actually pneumonia

$$\text{Precision} = \frac{TP}{TP+FP}$$

Precision: 0.93 or 93%

3) Recall: Recall measures how well the model identifies pneumonia cases.

$$\text{Recall} = \frac{TP}{TP+FN}$$

Recall: 0.979 or 97.9%

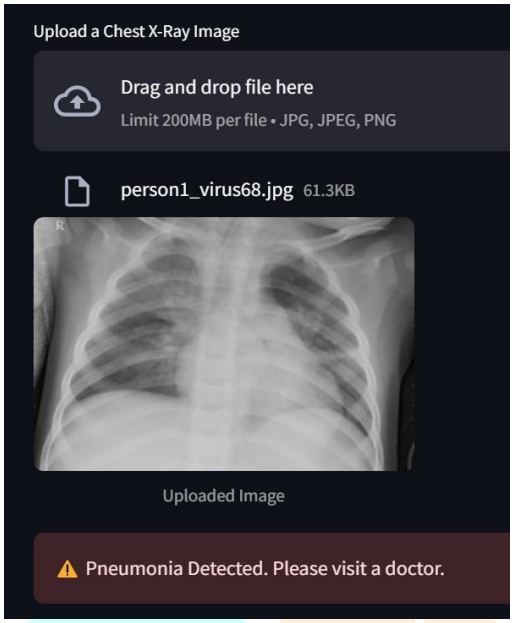
4) F1 Score: The F1-score is the harmonic mean of precision and recall, balancing false positives and false negatives.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$$

F1 Score: 0.958 or 95.8%

- ***GUI Interfacing:***

A GUI provides a user-friendly way to interact with the model, making it accessible even to those with basic computer knowledge. Streamlit app is used for front end, an option is provided to drag and drop the file as well as browsing the files from our computer. After selecting the image it loads the model that is stored in the computer and predicts the output as Normal or Pneumonia detected.



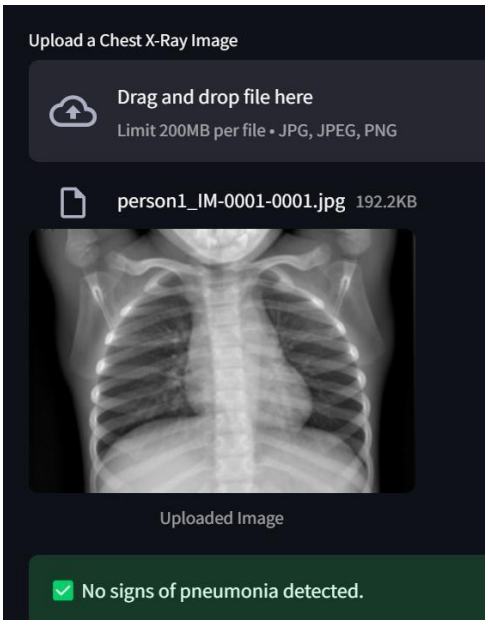


Fig 5a. Prediction of Pneumonia through GUI(Pneumonia) Fig 5b. Prediction of Pneumonia through GUI(Normal)

VI. CONCLUSION

In conclusion our project demonstrates the effectiveness of Vision Transformers (ViTs) in detecting pneumonia from chest X-ray images. Vision Transformers excel in capturing global context, spatial relations, and handling variable image resolutions, leading to accurate pneumonia detection (95.38%) over traditional CNNs. This model provides a reliable and automated diagnosis, helping in the early detection of pneumonia and assist healthcare professionals in making faster and more reliable decisions, ultimately improving patient outcomes.

REFERENCES

[1] American Lung Association. "What Causes Pneumonia?" Last updated December 11, 2024. https://www.lung.org/lung-health-diseases/lung-disease-lookup/pneumonia/what-causes-pneumonia.

[2] National Heart, Lung, and Blood Institute. "Pneumonia Causes and Risk Factors." Accessed April 2, 2025. https://www.nhlbi.nih.gov/health/pneumonia/causes.

[3] Khan, S. H., et al. "Efficient Pneumonia Detection Using Vision Transformers on Chest X-Rays."
Scientific Reports 15, no. 1 (2025): 11309. https://doi.org/10.1038/s41598-022-15341-0.

[4] Singh, Swapnil. "Pneumonia detection using deep learning." In 2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE), pp. 1-6. IEEE, 2021.

[5] Aljawarneh, Shadi A., and Romesaa Al-Quraan. "Pneumonia detection using enhanced convolutional neural network model on chest x-ray images." *Big Data* (2023).

[6] Kaushik, V. & Nayyar, Anand & Kataria, Gaurav & Jain, Rachna. (2020). Pneumonia Detection Using Convolutional Neural Networks (CNNs). Lecture Notes in Networks and Systems. 471-483. 10.1007/978-981-15-3369-3_36.

[7] Ahmed, Md. Sabbir & Rahman, Rafeed & Hossain, Shahriar. (2021). Pneumonia Detection by Analyzing Xray Images Using MobileNET, ResNET Architecture and Long Short Term Memory. 10.1109/ICCTA52020.2020.9477664.

[8] Siddiqi, Raheel & Javaid, Sameena. (2024). Deep Learning for Pneumonia Detection in Chest X-ray Images: A Comprehensive Survey. Journal of Imaging. 10. 176. 10.3390/jimaging10080176.

[9] https://www.nature.com/articles/s41598-024-52703-2

[10] P. Mooney, "Chest x-ray images (pneumonia)," 2018, accessed:2024-09-28. [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia