Automated Pneumonia Classification In Chest Radiographs Using Dual-Stage Ensemble Learning And LIME Interpretability

1st Pushpita Biswas[1], 2nd Aum Dubey[2], 3rd Anirudh Vyas M[3] ,4th Prof. Kalaavathi B

[1]School of Computer Science and Engineering, VIT, Vellore, India [2]School of Computer Science and Engineering, VIT, Vellore, India [3]School of Computer Science and Engineering, VIT, Vellore, India
[4]Professor, School of Computer Science and Engineering, VIT, Vellore, India

Pneumonia is an inflammatory condition of the lung primarily affecting the small air sacs known as alveoli caused by microorganism infection which can be viral or bacterial. It remains a critical public health concern worldwide, particularly in low resource settings, affecting severely specific age groups such as newborns & infants under 2 years old and adults above 65, due to their weak immune system. While Chest radiograph imaging is the most well-known screening approach used for detecting pneumonia in the early stages, its blurry and low illumination nature may call forth human error in manual diagnosis. Hence, the contribution of this work is the development of a two-stage pneumonia detection Expert System fusing the capabilities of both ensemble convolutional networks and the Transformer mechanism. In the first stage, a binary classification ensemble model is employed to detect whether a given chest X-ray indicates pneumonia or not. Upon a positive detection, the second stage activates a multi-class classification ensemble model that further categorizes the pneumonia into viral or bacterial, thus providing a finer level of diagnostic detail. The ensemble learning extracts strong features from the raw input X-ray images in two different scenarios: ensemble A (i.e., DenseNet201, Xception and InceptionResNet V2) and ensemble B (i.e., DenseNet201, Xception and VGG-16).

The proposed ensemble deep learning model recorded 95.95% classification performance in terms of overall accuracy and F1-score for the binary classification task, while it achieved 88.57% for multi-classification task. To ensure that the diagnosis is not only automated but also interpretable to end-users, including healthcare professionals, the model is fed to an expert system where users can upload X-ray images, get a classification result and see the highlighted region which supports the diagnosis through Local Interpretable Model-Agnostic Explanations (LIME), a black box testing strategy. The proposed framework could provide promising and encouraging explainable identification performance compared to the individual or existing ensemble models building trust of users and healthcare professionals on the result.

Keywords: Pneumonia, Chest X-Ray, Deep Learning, Explainable AI, Ensemble Model, LIME

I. INTRODUCTION

Pneumonia is a serious respiratory condition characterized by inflammation of the air sacs in one or both lungs, commonly caused by bacterial, viral, or fungal infections. It continues to be a leading cause of morbidity and mortality worldwide, especially among children under five and the elderly. Early and accurate diagnosis is crucial for effective treatment and improved patient outcomes. Traditionally, pneumonia is diagnosed using clinical evaluation and imaging techniques, particularly chest X-rays, which allow physicians to visualize lung infiltrates or opacities indicative of infection. However, interpretation of radiographic images is often subjective, reliant on the expertise of radiologists, and susceptible to inter-observer variability. In resource-constrained environments where access to skilled professionals is limited, this dependency can hinder timely and accurate diagnosis. In recent years, advancements in artificial intelligence (AI) and deep learning have shown great potential in automating the detection of medical conditions from imaging data. AI-driven tools can analyze medical images with high precision, reducing the workload on clinicians and enhancing diagnostic consistency. Nevertheless, challenges such as dataset imbalance, subtle inter-class differences, and the need to differentiate between various pneumonia types persist. Furthermore, most existing AI systems are either focused solely on binary classification—

distinguishing pneumonia from healthy lungs—or limited in their ability to generalize across different forms of pneumonia. There is a growing need for comprehensive systems that not only detect the presence of pneumonia but also classify its type, enabling more targeted clinical decision-making. This paper addresses these challenges by developing an end-to-end automated system that integrates advanced image analysis techniques for both binary and multi-class pneumonia classification. The objective is to provide an intelligent, accessible, and efficient diagnostic support tool that can aid healthcare professionals in identifying pneumonia accurately and rapidly through chest X-ray interpretation.

The motivation behind this paper stems from the urgent need for accessible, accurate, and scalable diagnostic tools for pneumonia, a disease that continues to claim countless lives globally despite being treatable. In many parts of the world, particularly in low- resource and rural settings, there is a critical shortage of radiologists and diagnostic infrastructure, leading to delays or misdiagnoses that can have fatal consequences. With the increasing availability of medical imaging data and advancements in artificial intelligence, there is a powerful opportunity to bridge this gap through automated diagnosis. The ability of deep learning models to recognize complex patterns in chest X-ray images presents a promising solution for early detection. However, most existing AI solutions either focus only on detecting the presence of pneumonia or struggle to differentiate between its bacterial and viral forms, which are clinically significant distinctions for treatment planning. This paper is driven by the ambition to create a more comprehensive and intelligent system that not only detects pneumonia but also classifies its type with high accuracy. By combining multiple high-performing models through ensemble learning, and integrating the solution into an interactive web application, this work aims to deliver a practical tool that can assist healthcare professionals and potentially be deployed in real-world clinical settings to support faster, more accurate, and informed medical decisions.

This paper focuses on the development of an automated deep learning-based system for the detection and classification of pneumonia using chest X-ray images. The scope includes both binary classification (pneumonia vs. no pneumonia) and multi- class classification (bacterial pneumonia, viral pneumonia, and no pneumonia. The paper involves training and evaluating multiple state-of-the-art convolutional neural network (CNN) architectures on a labelled dataset of chest X-rays. From these, the top- performing models are selected based on accuracy and other performance metrics and are combined using an ensemble learning technique to improve overall prediction reliability and generalization. In addition to model development, the paper includes the deployment of a user-friendly web-based application that allows users to upload chest X-ray images and receive diagnostic predictions in real-time. This platform also intelligently decides whether to use the binary or multi-class classifier depending on the nature of the case. The paper is limited to pneumonia detection through posterior- anterior (PA) view chest X- rays and does not include diagnosis through other imaging modalities like scans. The system is designed as a diagnostic aid and not a replacement for professional medical judgment. The ultimate aim is to contribute to the advancement of accessible, accurate, and scalable diagnostic tools that can support medical practitioners.

II. LITERATURE REVIEW

Pneumonia detection has been a focus of significant research due to its critical impact on public health. Traditional diagnostic techniques primarily involve clinical evaluation and chest X-ray interpretation by radiologists. These methods, although effective, are time-consuming and heavily dependent on the expertise of medical professionals. To overcome these limitations, various automated and semi-automated techniques have been proposed over the years.

Machine learning approaches, including Support Vector Machines (SVMs), Random Forests, and k-Nearest Neighbors (k-NN), have been employed to classify medical images based on engineered features such as texture, shape, and intensity. These methods rely on manual feature extraction, which can be subjective and less adaptable to complex datasets. Advances in image processing have also introduced edge- detection and region-growing algorithms to identify abnormalities in chest X-rays. These techniques, while offering some level of automation, lack the robustness needed for widespread clinical adoption.

The advent of deep learning has significantly advanced the field, enabling end-to-end learning from raw data without the need for manual feature extraction. Pre-trained Convolutional Neural Networks (CNNs) such as InceptionResNetV2, Xception, Dense- Net, and VGG16 have demonstrated superior performance in medical imaging tasks, including pneumonia detection. These models leverage large datasets and hierarchical feature extraction to achieve high accuracy. Furthermore, studies have explored ensemble

methods to improve diagnostic reliability by combining predictions from multiple models. However, these studies often focus narrowly on model performance without addressing practical deployment challenges.

In addition to algorithmic developments, research has also highlighted the importance of system interpretability and user accessibility. Techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping) have been used to visualize the decision-making process of deep learning models, aiding in trust-building among healthcare professionals. Yet, few studies extend their findings to create holistic, deployable systems that integrate diagnostic tools with real-time usability and web- based interfaces. This paper builds on these advancements by addressing both technical and practical aspects, aiming to deliver a comprehensive solution for pneumonia detection and classification.

TABLE I LITERATURE REVIEW

Title	Source	Attributes	Methodology	Metrics
Deep learning-based approach for automatic X-ray image detection of pneumonia and COVID-19 patients	Guangzhou Women and Children's Medical Center, Persahabatan Hospital in Jakarta	Viral pneumonia and COVID-positive, Bacterial pneumonia, viral pneumonia and COVID-negative, Normal	CNN	99.6% Accuracy, 99.7% Precision, 99.7% Sensitivity, 99.1% Specificity, 99.74% F1 Score
BND-VGG-19: A deep learning algorithm for COVID-19 identification utilizing X-ray images	Italian Society of Medical, Interventional Radiology (SIRM) and the Novel Coronavirus 2019 dataset	COVID-19, Normal, Viral pneumonia	BND-VGG-19	95.48% Accuracy
Automated lung-related pneumonia and COVID-19 detection based on novel feature extraction framework and Vision Transformer approaches using chest X-ray images	COVID-19 Radiography Dataset, Covid-ChestX- ray-15k dataset	COVID-19, Normal, Lung Opacity, Viral pneumonia	Ensemble Techniques	97.84% Accuracy, 96.76% Sensitivity, 96.80% Precision
Attention-enhanced architecture for improved pneumonia detection in chest X-ray images	Guangzhou Women and Children's Medical Center	Normal, Bacterial pneumonia, Viral pneumonia	ResNet + Attention-enhanced model	98% Accuracy
Pneumonia detection in chest X- ray images using an ensemble of deep learning models	Radiological Society of North America (RSNA)	Normal, Pneumonia, Lung Opacity, No Lung Opacity	GoogleNet, ResNet-18, DenseNet-121	98.81% Accuracy, 87.02% Sensitivity

III. METHODOLOGY

This study proposes a two-phase ensemble-based deep learning framework for accurate pneumonia detection and classification from chest X-ray images. The methodology is organized into several critical stages, including data collection and preprocessing, model architecture, training configuration, ensemble strategy and performance evaluation.

A. Data Collection and Pre-processing

Data collection was the first phase of our methodology. The primary dataset comprises chest X-ray images categorized into four distinct classes: Normal, Pneumonia, Bacterial Pneumonia, and Viral Pneumonia collected from reliable Kaggle and Mendeley multiple repositories. These are split into two datasets—one for binary classification (Normal vs. Pneumonia) and another for multi-class classification (Normal, Bacterial, and Viral Pneumonia). The datasets are carefully curated to maintain a balanced representation and are subsequently partitioned into training (70%), validation (15%), and testing (15%) sets. Before feeding the images into the models, data preprocessing techniques are applied. This includes resizing the images to a uniform dimension (224x224), normalization of pixel values (1/255), and data augmentation methods such as rotation (15), horizontal flipping, and zooming (0.2) to enhance model generalizability. Exploratory data analysis (EDA) is conducted to understand the distribution of classes, detect potential class imbalances, and evaluate image quality. Insights from EDA inform the model selection and augmentation strategies.

B. Model Framework

The system employs an ensemble of pre-trained convolutional neural networks (CNNs) to leverage transfer learning. The models were selected after comparative analysis among 5 models in particular: Dense Net, Xception, InceptionV3, VGG16 and InceptionResNetV2. For binary classification during model training, DenseNet201, InceptionResNetV2, and Xception Net showed better results and hence were selected for the ensemble framework. For multi-class classification, the ensemble includes DenseNet201, VGG16, and Xception Net. These models are fine-tuned on the respective datasets. The ensemble approach aggregates predictions from each model to produce a more robust and accurate outcome by majority voting techniques.

1) Xception

Xception (Extreme Inception) is a deep learning architecture that improves upon the Inception model by using depth wise separable convolutions instead of traditional convolutions. In this approach, a depth wise convolution is first applied to each input channel separately, followed by a pointwise convolution (1x1 convolution) to combine the outputs, significantly reducing computational cost and parameters. The Xception network consists of three main components: Entry Flow, Middle Flow, and Exit Flow, where the middle flow focuses on depthwise separable convolutions. This design improves the model's efficiency and performance, especially in image classification tasks. The formula for a depthwise separable convolution is as follows where X is the input, Wfc is the fully connected layer weights, and the final output is passed through softmax for classification.

$$Y = \text{Softmax}\,(W_{fc} \cdot (\text{Pointwise Conv}\,(\text{Depthwise Conv}(X))))$$

2) Dense Net

DenseNet (Densely Connected Convolutional Networks) is a deep learning architecture where each layer is connected to every other layer in a dense block, allowing for efficient feature reuse and improved gradient flow. Instead of relying on traditional convolutions, Dense Net uses a growth rate, where each layer generates a fixed number of feature maps that are concatenated with previous layers' outputs. This reduces the number of parameters and prevents overfitting. DenseNet-201, for instance, has 201 layers, organized into dense blocks separated by transition layers for down sampling. The formula for each layer output is as follows where Hi represents the convolution and activation applied to the concatenated outputs of all previous layers:

$$Y_i = H_i([X_0, X_1, \ldots, X_{i-1}])$$

3) Inception-Ressnet V2

InceptionResNetV2 combines the benefits of the Inception architecture with the residual connections from ResNet. It uses Inception modules to apply various filter sizes and pooling operations in parallel, while residual connections help with gradient flow and faster convergence by skipping over certain layers. This combination enables the model to maintain the depth and flexibility of Inception while benefiting from residual learning. The formula involves applying a series of convolutions followed by a residual skip connection as follows where X is the input and the Inception module applies convolutions, and the output is added to the input via residual connections.

$$Y = X + \text{Inception Module}(X)$$

4) VGG 16

VGG16 is a simple and deep convolutional neural network architecture characterized by its stacked convolutional layers with small 3x3 filters and max-pooling layers. The network has 16 layers, including 13 convolutional layers and 3 fully connected layers. VGG16 is known for its uniform architecture, where each block contains a series of convolutions followed by max-pooling. The formula for a convolutional layer is as follows where X is the input, W is the convolution filter, b is the bias, and Y is the output. The fully connected layers are applied after flattening the feature maps, and the final output passes through a softmax function for classification.

$$Y = (W * X) + b$$

5) Inception V3

GoogleNet introduces the Inception module, which applies multiple convolutions with different filter sizes (1x1, 3x3, 5x5) and pooling operations in parallel, followed by concatenation of the outputs. This allows the model to capture features at different scales while maintaining computational efficiency. GoogLeNet also uses auxiliary classifiers to improve gradient flow during training. The formula for an Inception module is as follows where the outputs of different convolutional and pooling operations are concatenated to form the final output:

Y = Concat(Conv(1x1),Conv(3x3),Conv(5x5),MaxPool(3x3))

Inception V3 is an enhanced version of the GoogLeNet, further optimizing the Inception module by introducing techniques such as factorization of convolutions and asymmetric convolutions to reduce computational complexity while maintaining accuracy. It also uses auxiliary classifiers and applies batch normalization to improve training efficiency. Inception V3's architecture is composed of several Inception modules, and its formula is similar to the one in GoogleNet, but with more refined convolutional operations for efficiency.

.

Y = Concat(Conv(1x1),Conv(3x3),Conv(5x5),MaxPool(3x3))

C. Training Configuration

Each model is trained individually using the training dataset while hyperparameters such as learning rate(0.0001), batch size(4), and number of epochs(30) are tuned based on validation performance. The ensemble models are created by combining
the outputs of the individual CNNs. Training is performed using cross-entropy loss (binary and categorical) and the Adam optimizer. Early stopping and learning rate schedulers are used to avoid overfitting and improve convergence. Each model was modified at the top layers to accommodate three output classes using a softmax activation. Intermediate layers were frozen to preserve the generalizable feature extraction learned from large-scale datasets (e.g., ImageNet). Custom dense layers were appended for domain-specific classification, with dropout applied to prevent overfitting.

D. Ensemble Strategies

To enhance predictive performance and robustness, a soft voting ensemble approach was adopted. Probabilistic outputs from the individual models were averaged to determine the final class prediction. This method allows the ensemble to consider the confidence levels of all base learners rather than relying on a single model's output. The ensemble strategy aims to reduce variance and capture complementary strengths across the constituent models.

E. Performance and Evaluation

The trained models are evaluated on the test dataset using a comprehensive set of metrics. These include accuracy, precision, recall, specificity, sensitivity, and F1- score. Furthermore, confusion matrices are generated to analyze class- wise performance. The ROC (Receiver Operating Characteristic) curve and PR (Precision-Recall) curve are plotted to visualize the model's diagnostic ability across different thresholds. They were plotted for each class to visualize sensitivity versus specificity trade-offs. The Area Under the ROC Curve (AUC) provided a scalar summary of model discrimination ability. Finally, confusion matrices were generated to interpret classification outcomes and identify common misclassifications.

To ensure clinical reliability and transparency, the system incorporates both black- box and white-box testing methods. Local Interpretable Model-agnostic Explanations (LIME) is used for black-box testing to explain individual predictions. GradCAM (Gradient-weighted Class Activation Mapping) is utilized for white-box testing to highlight the regions in the X-ray images that influenced the model's decision. This step is crucial for gaining clinician trust and validating the model's reasoning.

The final ensemble models are integrated into an expert decision support system. This system takes a chest X-ray image as input, performs classification through the ensemble models, and outputs a diagnosis along with interpretability visuals. If pneumonia is detected, the system further classifies it into bacterial or viral types. The expert system thus serves as a comprehensive diagnostic tool for early and reliable pneumonia detection.

The testing phase of the proposed framework involves both black-box and white- box evaluation techniques to ensure the reliability and interpretability of the model predictions. For black-box testing, the Local Interpretable Model- Agnostic Explanations (LIME) method is employed to provide human-understandable explanations for individual predictions, helping to validate the model's decision- making process without accessing internal parameters. In parallel, white-box testing is conducted using Gradient-weighted Class Activation Mapping (Grad-CAM), which visualizes the specific regions of the chest X-ray images that influenced the model's output. These interpretability methods allow for enhanced transparency, clinical trust, and a deeper understanding of how the model distinguishes between normal and pneumonia-affected cases.

IV. RESULTS AND DISCUSSION

The Pneumonia Detector is an AI-powered web application developed to streamline and enhance the process of diagnosing pneumonia from chest X-ray images. Designed with a focus on accessibility and ease of use, the platform enables users to simply drag and drop or browse and upload chest X-rays in JPG, PNG, or JPEG formats, with support for files up to 200MB. Once an image is uploaded, the application utilizes a deep learning model—specifically an Ensemble model trained on a large dataset of annotated chest radiographs to analyze the image and detect signs of pneumonia. The diagnostic process is performed in real time, providing users with immediate feedback on the likelihood of infection. This tool not only demonstrates the practical integration of artificial intelligence in medical imaging but also highlights the potential for AI to assist healthcare professionals in making faster and more accurate decisions. The clean and modern interface, developed using Streamlit, ensures smooth user experience and makes it easy to interact with the underlying machine learning model. Whether for educational purposes, clinical assistance, or research demonstrations, the Pneumonia Detector serves as a powerful example of how machine learning can be effectively applied to solve real-world problems in the healthcare domain.

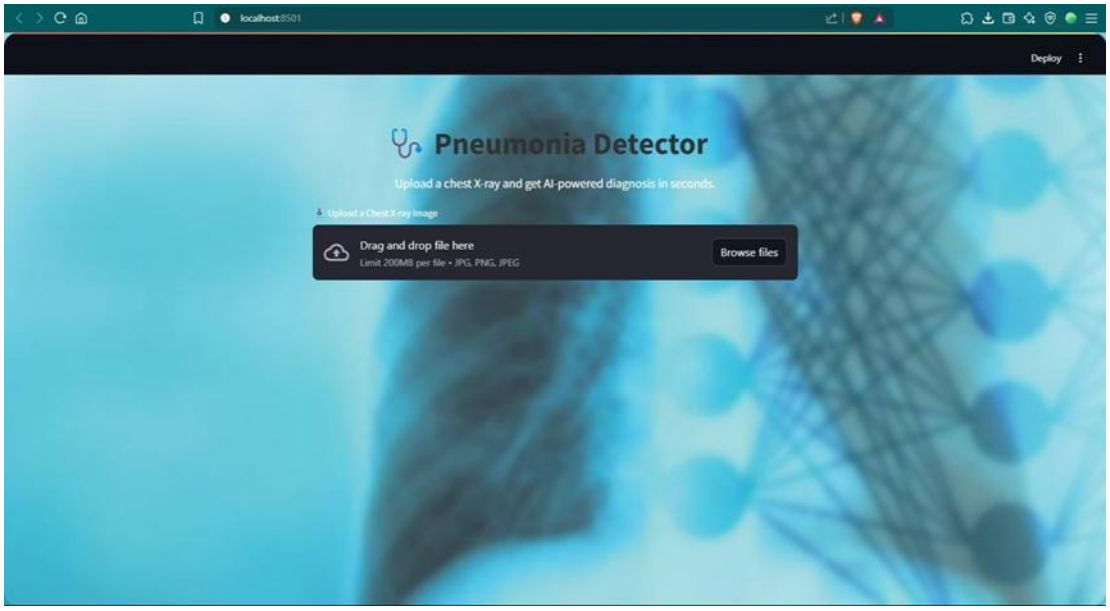

Fig. 1. WEBSITE FOR PREDICTION

To use the Pneumonia Detector application, the user begins by accessing the web interface, where they are prompted to upload a chest X-ray image either by dragging and dropping the file or by browsing their device to select an image in JPG, PNG, or JPEG format. Once the image is uploaded, the backend system processes the file and feeds it into a pre-trained deep learning model specifically designed to detect pneumonia. This model, built using convolutional neural networks (CNNs), analyzes the X-ray to identify patterns and anomalies typically associated with pneumonia. After processing, the model outputs a diagnostic result indicating whether the image shows signs of pneumonia or appears normal. The result is then displayed directly on the interface, providing users with an immediate, AI-generated interpretation of the X-ray. This seamless process is designed to be intuitive, fast, and reliable, allowing users to obtain a preliminary medical insight in just a few seconds.

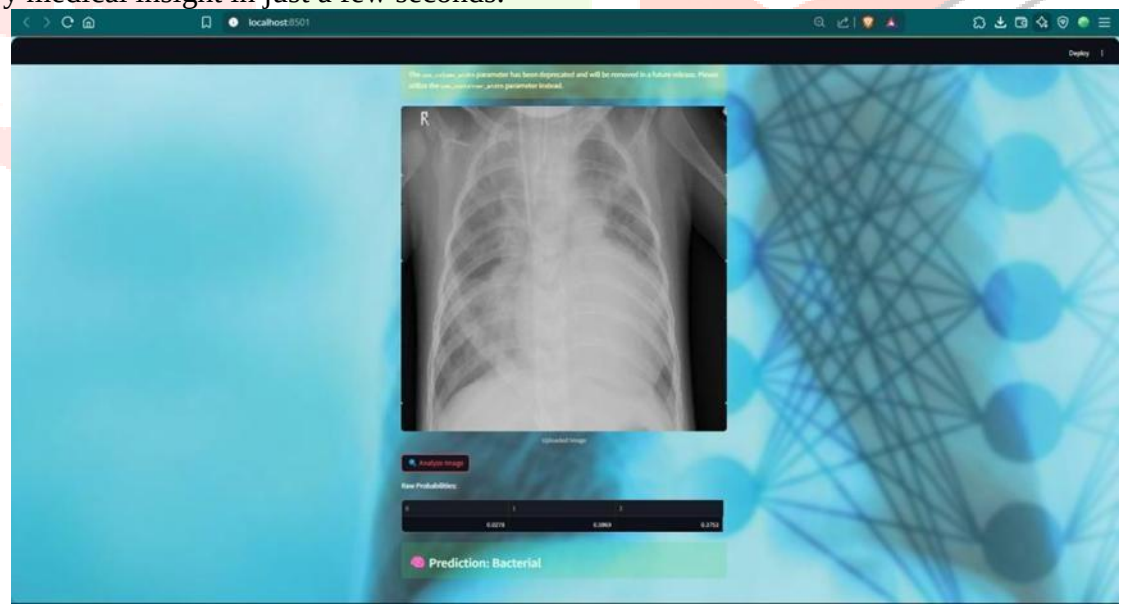

Fig. 2. BACTERAL PNEUMONIA CLASSIFIED BY WEBSITE

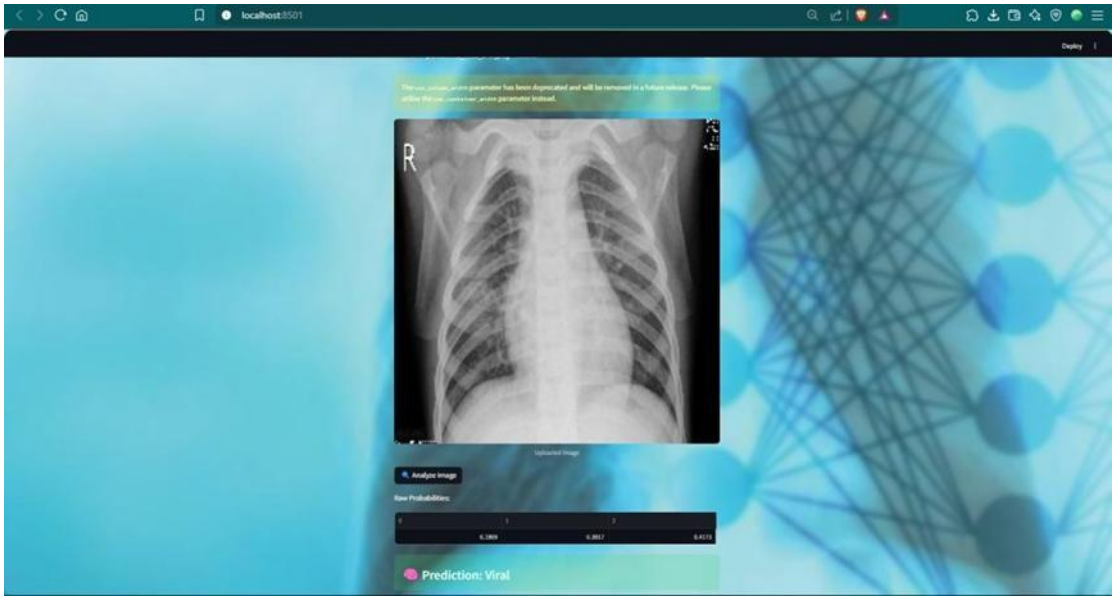

Fig. 3. VIRAL PNEUMONIA CLASSIFIED BY WEBSITE

Through extensive experimentation and evaluation, various models were tested on both binary classification (Normal vs. Pneumonia) and multi-class classification (Normal, Viral Pneumonia, and Bacterial Pneumonia) tasks. For binary classification, the best- performing models — Xception Net, DenseNet, and InceptionResNet — each achieved an impressive accuracy of 95.84%, while the ensemble learning approach slightly outperformed them all with a peak accuracy of 95.95%, showcasing the power of model fusion in boosting performance and reducing error as seen in Table 7.1. On the other hand, Xception, VGG16 and DenseNet were used to build the Ensemble_Multi model which scored significantly higher accuracy of 80.57% outperforming the rest. VGG16 and GoogleNet also performed strongly, with accuracies of 94.71% and 92.35% respectively.

TABLE II
ACCURACY OF BINARY CLASSIFICATION

Binary Classification	**Accuracy**
Xception Net	95.84%
VGG 16	94.71%
GoogleNet	92.35%
DenseNet	95.84%
InceptionResNet	95.84%
Ensemble learning	95.95%

TABLE III
ACCURACY OF MULTI CLASSIFICATION

Multi Classification	**Accuracy**
Xception Net	77.24%
VGG 16	77.21%
GoogleNet	73.45%
DenseNet	79.00%
InceptionResNet	72.50%
Ensemble learning	80.57%

In the multi-class classification scenario, which is inherently more challenging due to the finer distinctions required between different types of pneumonia, DenseNet led the way with 79.00% accuracy, followed closely by Xception Net and VGG16, both hovering around 77.2%. GoogleNet and InceptionResNet demonstrated slightly lower accuracies of 73.45% and 72.50%, respectively, which still reflects competent performance given the complexity of the task. While ensemble learning in the multi- class setting gives 80.57% testing accuracy, it holds promise for further accuracy improvements by leveraging the complementary strengths of individual models.

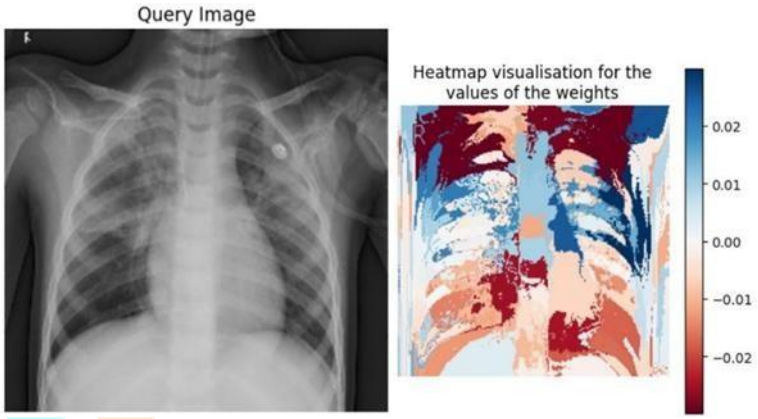

Fig. 4. LIME INTERPRETATION

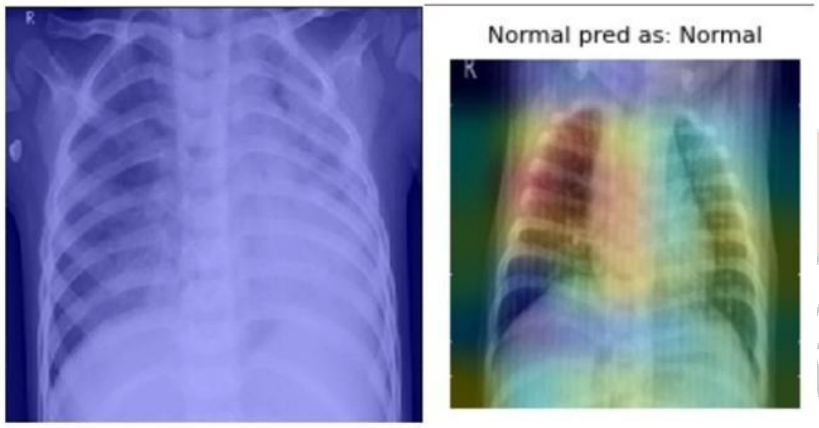

Fig. 5. GRAD-CAM INTERPRETATION

From here, the aforementioned models are ensembled together to build 2 ensembles separately for binary and multiclass classification. Ensemble_Binary shows the following confusion matrix:

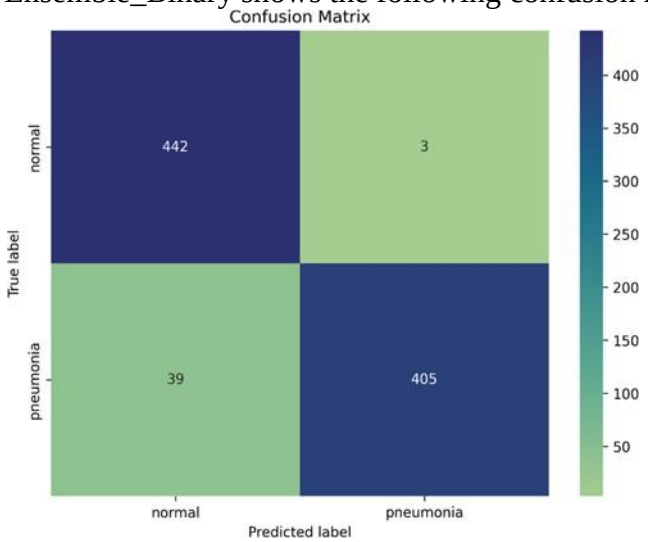

Fig. 6. CONFUSION MATRIX

Fig. 6. provides a clear summary of our model's classification performance. Out of all the test samples, the model correctly identified 442 normal cases and 405 pneumonia cases. It only misclassified 3 normal cases as pneumonia (false positives), which is minimal. However, there were 39 pneumonia cases misclassified as normal (false negatives). While this is still a relatively small number, in medical diagnosis, false negatives are critical as they represent missed cases of illness. Nevertheless, the overall accuracy and balance between the two classes appear strong, indicating a well- performing model.

The training and validation metric plots in Fig. 7. offer deeper insight into the learning dynamics of your model over 18 epochs. The loss curves show a consistent downward trend in both training and validation loss, suggesting that the model is learning effectively without overfitting. Accuracy steadily increases and stabilizes above 94–95% for both training and validation, reinforcing the model's strong performance. The precision graph indicates very high precision on the validation set, consistently close to 98–99%, which means the model rarely misclassifies normal cases as pneumonia. The recall plot shows the model is also highly capable of identifying pneumonia cases, with a validation recall hovering around 93–95%, though with slight variability. This balance between precision and recall is essential, particularly in a healthcare setting.

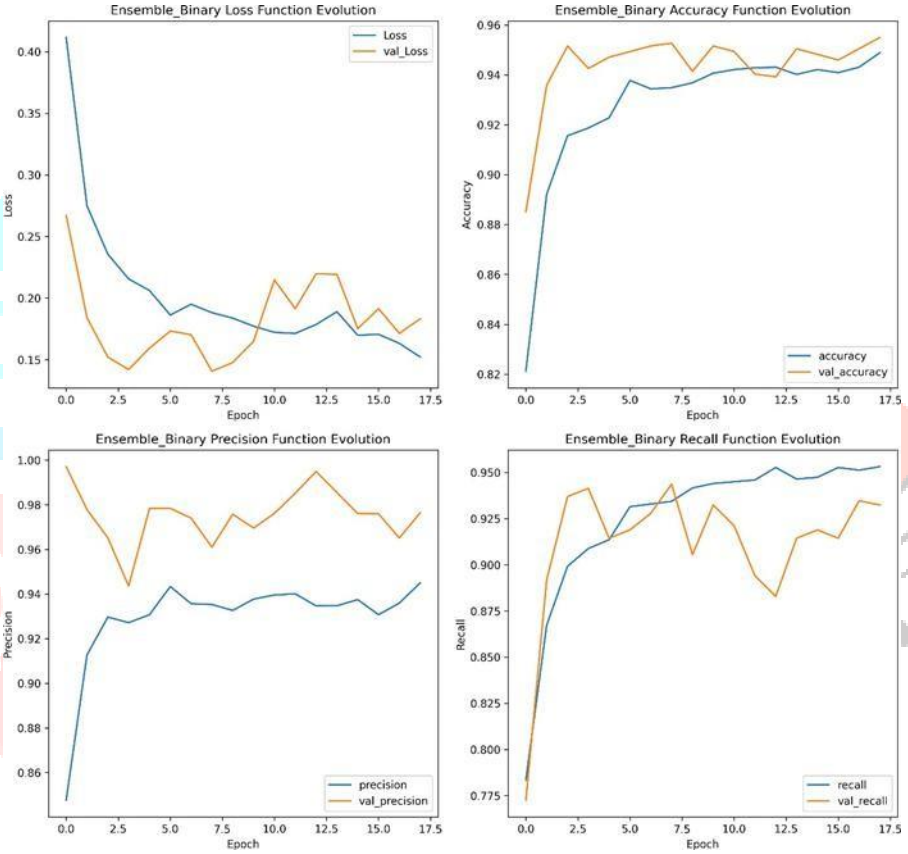

Fig. 7. TRAINING AND VALIDATION PLOT FOR ENSEMBLE BINARY

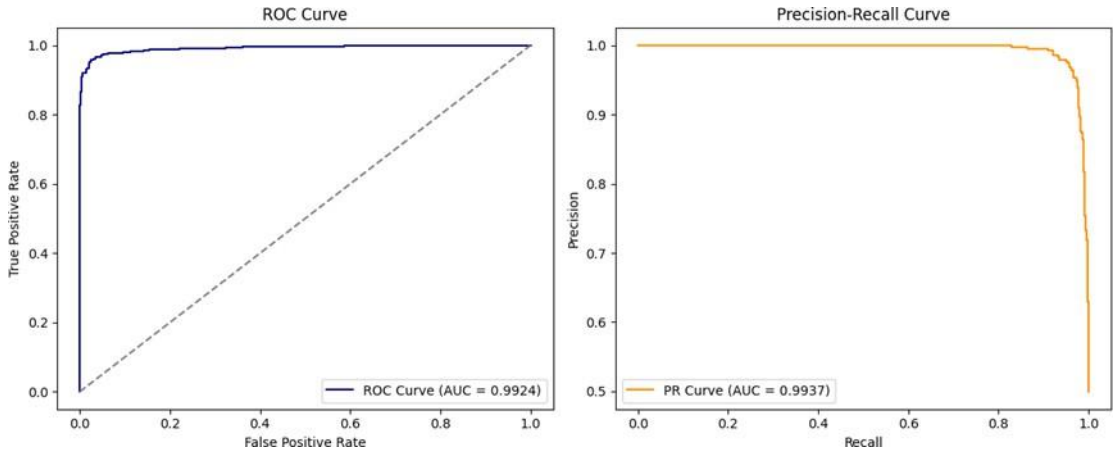

Fig. 8. .ROC AND PR CURVE FOR ENSEMBLE BINARY.

Fig 8. presents the ROC (Receiver Operating Characteristic) and Precision- Recall curves, which further validate the model's excellent classification capability. The ROC curve has an Area Under the Curve (AUC) of 0.9924, indicating near-perfect discrimination between the two classes. Similarly, the Precision-Recall curve yields an AUC of 0.9937, which is particularly valuable when dealing with class imbalances — as is common in medical datasets. These curves confirm that your ensemble model not only classifies well across various thresholds but also maintains high precision and recall consistently.

For Ensemble_multi, Fig 9. shows the training and validation loss (left) and accuracy (right) over 30 epochs. The training and validation loss curves show a consistent downward trend, with the validation loss being lower than the training loss, indicating that the model is generalizing well. The accuracy plots demonstrate a steady increase, with validation accuracy reaching over 82% by the final epoch, outperforming training accuracy—another positive sign of good generalization without overfitting.

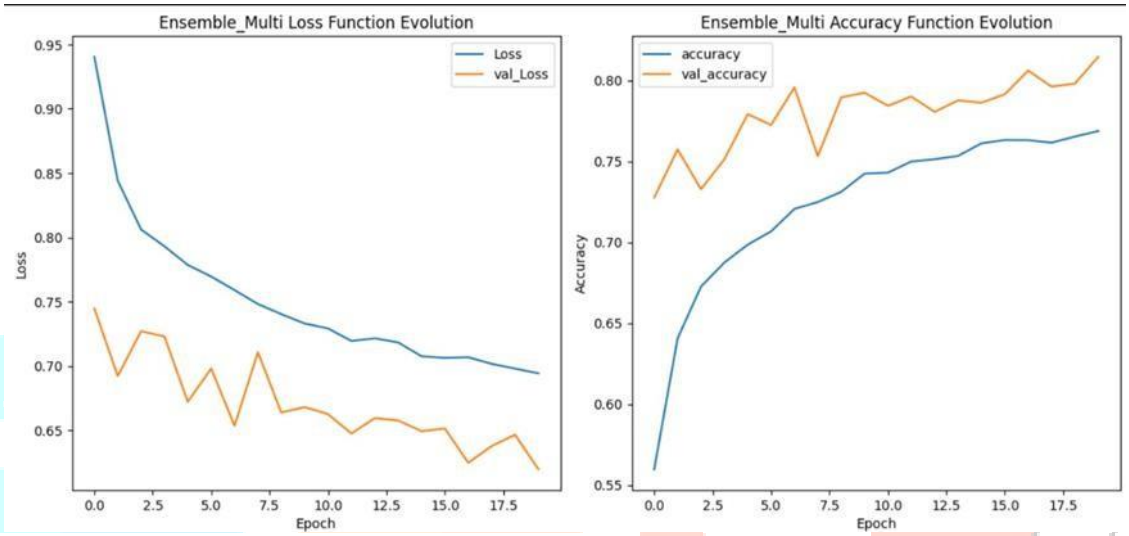

Fig. 9. TRAINING AND VALIDATION PLOT FOR ENSEMBLE MULTI

Fig 10. presents a confusion matrix for the three-class classification (bacteria, normal, virus). The model correctly identifies a large number of 'normal' cases (694) and performs reasonably well on 'bacteria' (523) and 'virus' (487). However, it shows confusion between 'bacteria' and 'virus', as evidenced by the 159 virus predictions for bacteria and 163 bacteria predictions for virus. The misclassifications are more pronounced between these two classes, possibly due to overlapping radiographic features.

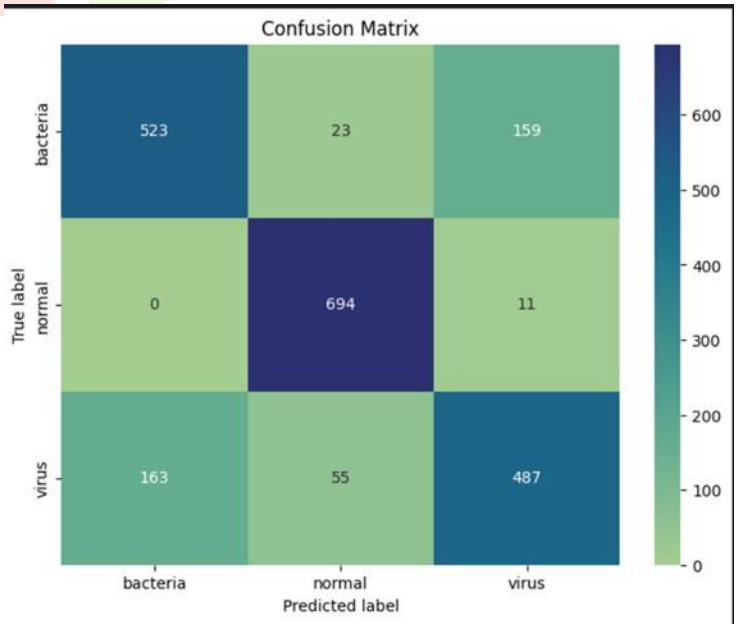

Fig. 10. CONFUSION MATRIX OF ENSEMBLE MULTI

Fig 11. depicts ROC curves for each of the three classes. Class 1 (likely 'normal') exhibits the highest performance with an AUC of 0.99, followed by class 0 ('bacteria') with an AUC of 0.91, and class 2 ('virus') with an AUC of 0.88. These high AUC values reflect the model's strong ability to distinguish between the classes, especially for the 'normal' category.

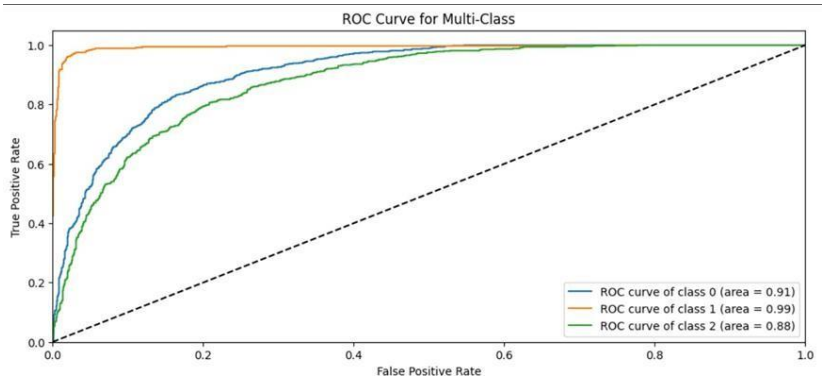

Fig. 11. ROC AND PR CURVE FOR ENSEMBLE MULTI.

Fig 12. shows the binary ROC curve for a pneumonia vs normal classification of the same ensemble model. The pneumonia class achieves a high AUC of 0.96, while the normal class has an AUC of 0.81. The micro-average and macro-average AUCs are both 0.89, suggesting a robust performance across both classes in this binary setup.

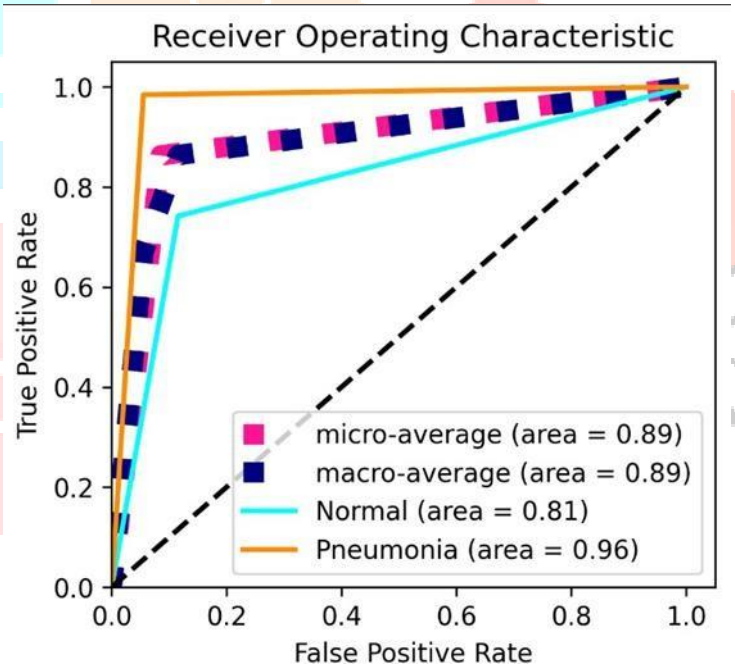

Fig. 12. ROC CURVE FOR ENSEMBLE MULTI WITH MICRO AND MACRO AVERAGE.

The ensemble model for pneumonia detection demonstrates exceptional performance, achieving high accuracy, precision, and recall, with minimal signs of overfitting. The most significant area for potential improvement lies in reducing false negativ es, which is vital for ensuring that pneumonia cases are not overlooked. Overall, these results suggest a robust and clinically valuable model.

Overall, this paper not only validates the effectiveness of deep learning for medical image classification but also emphasizes the importance of model selection and architecture tuning in achieving optimal results. The integration of these models into an interactive Streamlit-based interface bridges the gap between complex AI algorithms and real-world usability, offering a powerful diagnostic aid that can support healthcare professionals in making quicker and more accurate decisions. Future work could focus on optimizing ensemble strategies for multi-class classification, incorporating explainable AI techniques, and expanding the model’s capabilities to detect additional thoracic diseases, further enhancing its clinical value.

V. CONCLUSION

In conclusion, the Pneumonia Detector paper successfully demonstrates the potential of deep learning models in automating and enhancing medical image diagnosis, specifically for detecting pneumonia from chest X-rays. The application not only provides a user-friendly platform for real-time predictions but also leverages some of the most advanced convolutional neural network architectures to ensure high diagnostic accuracy.

While the Pneumonia Detector demonstrates strong performance and usability, there are several opportunities for enhancement to further improve its accuracy, reliability, and real-world impact. One of the key areas for development is the implementation of ensemble learning for multi-class classification. While it has already shown promise in binary classification by slightly boosting performance, integrating a robust ensemble strategy for distinguishing between normal, viral, and bacterial cases could significantly enhance multi-label accuracy and reduce misclassifications.

Another valuable addition would be the integration of explainable AI (XAI) techniques, such as Grad-CAM or LIME, which would allow users—especially healthcare professionals—to visually understand which regions of the X-ray contributed to the model's prediction. This transparency is crucial in clinical settings where trust in AI decisions is essential.

From a usability standpoint, deploying the application to a cloud platform (e.g., AWS, Azure, or GCP) would enable remote access and scalability, making the tool more accessible in real-world healthcare environments, especially in under-resourced or remote areas. Additionally, implementing multi-language support and a mobile- responsive design would expand the tool’s usability across diverse regions and devices.

On the data side, the system could benefit from being trained on an even larger and more diverse dataset, including images from different demographics, hospitals, and equipment sources, to further improve generalization and reduce bias. Incorporating other lung conditions such as COVID-19, tuberculosis, or lung cancer could also make the tool a more comprehensive diagnostic assistant.

Lastly, integrating with electronic health record (EHR) systems and report generation features could position the tool for clinical use, allowing it to contribute directly to the medical decision-making process. These enhancements would move the Pneumonia Detector from a proof-of-concept toward a deployable, trustworthy AI assistant in the medical domain.

VI. REFERENCE

[1]. Alapat, D. J., Menon, M. V., & Ashok, S. (2022). A review on detection of pneumonia in chest X-ray images using neural networks. Journal of Biomedical Physics and Engineering, 12(6), 551–588. https://doi.org/10.31661/jbpe.v0i0.2202-1461

[2]. Asnake, N. W., Salau, A. O., & Ayalew, A. M. (2024). X-ray image-based pneumonia detection and classification using deep learning. Multimedia Tools and Applications, 83, 60789–60807. https://doi.org/10.1007/s11042-023-17965-4

[3]. Beghoura, I., Benssalah, M., & Sbargoud, F. (2023). An improved CovidConvLSTM model for pneumonia-COVID-19 detection and classification. arXiv. https://doi.org/10.48550/arXiv.2408.11507

[4]. Cao, Z., Huang, J., He, X., & Zong, Z. (2022). BND-VGG-19: A deep learning algorithm for COVID-19 identification utilizing X-ray images. Knowledge-Based Systems, 258. https://doi.org/10.1016/j.knosys.2022.110040

[5]. Kundu, R., Das, R., Geem, Z. W., Han, G.-T., & Sarkar, R. (2021). Pneumonia detection in chest X-ray images using an ensemble of deep learning models. PLOS ONE, 16(9). https://doi.org/10.1371/journal.pone.0256630

[6]. Li, D. (2023). Attention-enhanced architecture for improved pneumonia detection in chest X-ray images. BMC Medical Imaging, 24, Article 6. https://doi.org/10.1186/s12880-023-01177-1

[7]. Lu, Z., Whalen, I., Dhebar, Y., Deb, K., Goodman, E., Banzhaf, W., & Boddeti, V. N. (2019). Multi-objective evolutionary design of deep convolutional neural networks for image classification. arXiv. https://doi.org/10.48550/arXiv.1912.01369

[8]. Majumder, T., Das Sarma, U., Choudhury, S., & Debnath, D. (2024). Pneumonia detection using asynchronous split learning method. IEEE Transactions on Consumer Electronics, 70(3). https://doi.org/10.1109/TCE.2024.3413893

[9]. Mehta, S., & Rastegari, M. (2021). MOBILEVIT: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv. https://doi.org/10.48550/arXiv.2110.02178

[10]. Qin, D., Bu, J.-J., Liu, Z., Shen, X., Zhou, S., Gu, J.-J., Wang, Z.-H., Wu, L., & Dai, H.-F. (2021). Efficient medical image segmentation based on knowledge distillation. arXiv. https://doi.org/10.48550/arXiv.2108.09987

[11]. 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. arXiv.

[12]. Raghaw, C. S., Bhore, P. S., Rehman, M. Z. U., & Kumar, N. (2024). An explainable contrastive-based dilated convolutional network with transformer for pediatric pneumonia detection. Applied Soft Computing. https://doi.org/10.1016/j.asoc.2024.112258

[13]. Ravi, V. (2020). Deep fine-tuned efficientNetV2 ensemble deep learning approach for pediatric pneumonia detection using chest radiographs. Journal of Intelligent & Fuzzy Systems. https://doi.org/10.3233/JIFS-219397

[14]. Shakouri, M., Iranmanesh, F., & Eftekhari, M. (2022). DINO-CXR: A self- supervised method based on vision transformer for chest X-ray classification. arXiv.

[15]. Sharma, S., & Guleria, K. (2024). A systematic literature review on deep learning approaches for pneumonia detection
using chest X-ray images. Multimedia Tools and Applications, 83, 24101–24151.
https://doi.org/10.1007/s11042-023- 16419-1

[16]. Ukwuoma, C. C., Qin, Z., Heyat, M. B. B., Akhtar, F., Bamisile, O., Muaad, A. Y., Addo, D., & Al-antari, M. A. (2022). A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. Journal of Advanced Research, 7(48), 191–211. https://doi.org/10.1016/j.jare.2022.08.021

[17]. Ukwuoma, C. C., Qin, Z., Heyat, M. B. B., Akhtar, F., Smahi, A., Jackson, J. K., Qadri, S. F., Muaad, A. Y., Monday,
H. N., & Nneji, G. U. (2022). Automated lung- related pneumonia and COVID-19 detection based on novel feature extraction framework and vision transformer approaches using chest X-ray images.
Bioengineering, 9(11), 709. https://doi.org/10.3390/bioengineering9110709

[18]. Widodo, C. S., Naba, A., Mahasin, M. M., Yueniwati, Y., Putranto, T. A., & Patra, P. I. (2022). UBNet: Deep learning- based approach for automatic X-ray image detection of pneumonia and COVID-19 patients. Journal of X-Ray Science and Technology, 30(1), 57–71. https://doi.org/10.3233/XST-211005