Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 2, 2026

# A TRANSFER LEARNING–BASED VGG-16 FRAMEWORK FOR AUTOMATED PNEUMONIA DETECTION FROM CHEST X-RAY IMAGES WITH REAL-TIME WEB DEPLOYMENT

**Malaika Nasir[1], Noreen Khalid[2], Zaeem Nasir[3], Asra Nasir[*4]**

[1,2]*Department of Computer Science, Beaconhouse National University, Lahore, Punjab 54000, Pakistan.*
[3]*Department of Data Science, National University of Computer & Emerging Sciences, Lahore, Punjab 54000, Pakistan.*
[*4]*Department of Computer Science, Kinnaird College for Women University, Lahore, Punjab 54000, Pakistan.*

[1]malaika1nasir@gmail.com, [2]noreen.khalid@bnu.edu.pk, [3] zaeem0077@gmail.com, [*4]asranasir09@gmail.com

DOI: https://doi.org/10.5281/zenodo.18677554

**Keywords**
*Pneumonia detection; Deep learning; Binary classification; Pre-trained model; VGG-16; Convolutional neural network;*

**Article History**
*Received: 17 December 2025*
*Accepted: 01 February 2026*
*Published: 16 February 2026*


**Corresponding Author: ***
**Asra Nasir**

**Abstract**
*As part of our focus on enhancing respiratory diagnostic processes, we examine the use of advanced machine learning techniques to help automate pneumonia detection from chest X-rays. Our goal is to help radiologists manage their workloads by giving them screening tools to aid in the diagnosis of patients in resource-constrained, high patient volume environments. Using Python in Jupyter notebooks, we created a pneumonia classification model utilizing transfer learning combined with techniques of medical image analysis. The model was built using the Kaggle Chest X-ray Pneumonia dataset. This dataset included images of pediatric chest X-rays which were divided into two classes: Normal and Pneumonia. Additionally, we improved the model's robustness with data preprocessing and augmentation techniques. We used a VGG-16 model as a feature extractor, then modified the model to suit the needs of our classification task. We added a binary classification layer to the top of the model. The results of our experiments were very successful with a validation accuracy of 99.04% and a test accuracy of 98.28% along with strong precision, recall, and F1 scores. There was very little misclassification. Additionally, to improve accessibility, we created a simple web-based interface. This allows users to upload X-ray images of their chest and receive predictions along with a confidence score in real time. This research proves that a transfer learning framework based on VGG-16 is not only accurate, but also demonstrates the potential of AI assistance in pneumonia screening.*

## INTRODUCTION

Pneumonia as an infection of the lung parenchyma is one infection that causes inflammation of the alveoli, which may lead to gas exchange and clinical issues. It is a serious and continues to be a big problem in the world especially in children under the age of five and elderly people [1]. As a global pediatric health concern, in the year 2019 alone, there were pediatric deaths amounting to hundreds of thousands [2]. Common symptoms include a cough, shortness of breath, and fever, and low oxygen levels. But, different populations present

it differently, which makes it difficult to diagnose and treat [3]. As a result, most are diagnosed and treated late. In developing regions where there are scarce health care services, environmental factors that put people at risk, timely diagnosis and treatment become a necessity [4].

Chest X-ray (CXR) is the most used diagnostic technology because it is cheap and does not require an invasive X-ray [5]. Nevertheless, it is difficult to interpret the x-ray because of the structures, and the opacity that is subtle, as well as some diseases that may have the same recapture, which may result in a misdiagnosis, especially in developing regions where there is limited expertise in radiology [6]. Deep learning (DL) is an advanced technology that can be used to automate the diagnosis of pneumonia using X-ray [7]. Furthermore, attention-based models and transformer architectures have also performed well due to their ability to understand patterns and global context in medical images [8].

This research develops a pneumonia detection framework using transfer learning and the VGG-16 model, which the authors also fine-tune for binary classification of chest X-ray images into two categories: Normal and Pneumonia. The model is trained and evaluated under the same experimental conditions to measure its accuracy and ability to generalize. A web-based front-end is developed to provide a practical solution for real-time predictions, confidence scores, and possible clinical utilization.

### 1.1. Motivation

Pneumonia diagnosis from chest X-rays still depends largely on manual interpretation, which can be slow and inconsistent due to radiologist workload, image-quality variations, and the visual similarity of pneumonia to other lung conditions. Because of the lack of fast, consistent, and accessible web-based screening tools, especially in low-resource settings, this study was designed to develop a deep learning–based pneumonia detection system with reliable performance. To improve usability, the trained model was also deployed as a web application, allowing users to upload chest X-ray images and instantly receive predictions with confidence scores.

### 1.2. Contribution

In this study, the main contributions are as follows:

- Development of an AI-based pneumonia detection framework using transfer learning with the VGG-16 architecture for binary classification of chest X-ray images into Normal and Pneumonia classes.
- The preprocessing and data augmentation pipeline (noise reduction, 224x224 image resizing and normalization) contributed to the model's robustness and helped mitigate overfitting.
- Fine-tuning strategy for adapting the pre-trained VGG-16 backbone with a customized classification head, resulting in streamlined training and enhanced generalization on medical imaging data.
- Comprehensive experimental evaluation using validation and test sets, with performance documented through accuracy/loss reporting, confusion matrices, classification reports, and thorough analysis of model reliability and misclassification behavior.
- Exceptional diagnostic performance, with validation accuracy of 99.04% and test accuracy of 98.28%, indicating strong generalization on unseen chest X-ray samples.
- The trained model is implemented as a web-based application with a practical deployment that includes a front-end landing page (GitHub Pages) and Gradio-based Hugging Face interface for user accessibility, allowing real-time predictions with confidence scores.

### 1.3. Paper Organization

This study presents a structured approach for automated pneumonia detection using chest X-ray images. Section 1 introduces the problem, motivation, and research objectives. Section 2 reviews recent deep learning–based pneumonia detection methods. Section 3 describes the dataset, preprocessing, augmentation, and the proposed VGG-16 transfer learning model. Section 4 reports experimental results and discussion using accuracy/loss curves, confusion matrices, and classification reports. Finally, Section 5 concludes the work.

## 2. RELATED WORK

Chest X-rays are a familiar and valuable tool in medical diagnosis and treatment. However, without a qualified radiologist, it is challenging to interpret. Therefore, many scientists are working to automate X-ray analysis. In recent years, the use of machine learning models including deep learning and transfer learning in the medical field has been growing, with clear successes in the predictions of various medical conditions. However, most datasets used for training these models are not diverse enough and this leads to disparities in healthcare data across different geographical regions. Several studies have been conducted to address the issue of data disparities in medical image analysis.

Alshanketi et al. (2025) evaluated deep learning models including VGG16, ResNet50, and Vision Transformers for pneumonia detection across multiple imbalanced datasets. The results showed that ViT achieved the highest baseline accuracy, while VGG16 produced superior recall. Class weighting, data augmentation, and semi-supervised learning significantly improved performance on imbalanced data. Transfer learning advocated strategies to achieve an accuracy of ~82% with only 64 BRAX images, despite showing unsatisfactory zero-shot accuracy of 58%. The authors suggest using more sophisticated semi-supervised techniques, GANs for augmentation, and model compression in future work [9] .

Aljawarneh et al. (2025) developed and validated an enriched CNN model for pneumonia detection from chest X-ray images comparing it with VGG-19, ResNet-50, and fine-tuned ResNet-50 models. His enriched CNN model, with the expanded Kaggle dataset 5863 images, 92.46% and outperformed all the transfer learning approaches. ResNet-50 produced the lowest accuracy (82.8%), while fine-tuned ResNet-50 reached 91%. The study shows that proprietary CNN model architectural train on medical images can outperform general-purpose pre-trained models. The authors cite the limited size and computational capability of the datasets as the primary obstacles, and suggest more extensive datasets and enhanced computational resources for future work [10].

Yanar et al, (2025) implemented PELM, an ensemble deep learning model which integrates InceptionV3, VGG-16, ResNet50, and Vision Transformer for pneumonia detection on a large-scale curated dataset of 50,000 CXR images. The model reached 96% accuracy, surpassing all individual architectures and previous works. Feature-level fusion allowed the model to understand both local and global patterns in images, which enhanced generalization on varied datasets. Despite strong performance, the study highlighted some limitations, including the curated and filtered images, absence of real-world diversity, and low interpretability. The authors proposed multi-institutional validation and combining explainable AI as future directions [11].

Mokarram et al. (2025) developed a hybrid lightweight deep learning model for pneumonia classification on CXR images by combining EfficientNet, ResNet50, and MobileNetV2. Their model feature extraction and diagnostic accuracy of 96.04% and above, exceeding single architectures. They applied dataset augmentation on multiple publicly available datasets to improve generalization. Despite the strong performance, limitations were noted such as dataset imbalance, limited real-world validation, and low explainability. The authors suggested the integration of diverse datasets, explainable AI, and testing in a clinical setting [12].

Bhatt et al. (2023) presented a lightweight ensemble CNN model for pneumonia detection by integrating three convolutional networks using different kernel sizes (3x3, 5x5, 7x7). Using the ensemble technique, the model achieved recall of 99.23%, with a small number of false negatives, however the total accuracy was 84.12%. The model was built to be computationally inexpensive and to be implemented in unstimulated environments. Results indicate that simple CNNs can be more successful in recall-sensitive medical applications when compared to deeper transfer-learning models. The authors indicate that the model can be improved in the future by using larger, more heterogeneous

datasets, applying better pre-processing techniques, and expanding the model to classify the sub types of pneumonia [13].

Al Reshan et al, (2023) create a MobileNet-based deep learning model for pneumonia detection using CXR imaging. The model was tested using two datasets (5,856 images and 112,120 images) and for the performance, he outperformed seven other pre-trained CNNs by obtaining the highest accuracy of 94.23% for ADAM optimizer, 16 batch size, 64 epochs. The study MobileNet performs strong and computationally inexpensive. However, results with different optimizers were inconsistent and a few models displayed poor validation. The authors recommend future studies to include larger and more balanced datasets and a greater variety of clinical datasets to improve real-world applicability and to improve the overall robustness of the models [14]

Nettur et al. (2024) developed a lightweight ensemble model approach and weighted average model using Kermany pediatric CXR dataset for pneumonia detection, MobileNetV2 and NASNetMobile. The ensemble approach outperformed single models achieving an impressive 98.63% accuracy. The authors employed fine-tuning, transfer learning, and optimal weight selection to to improve classification and remain computationally efficient. Despite strong results, authors mentioned dataset, single-center, and binary classification limitations and will seek to address these in future work by using larger, more varied datasets, adding multi-class prediction, and explainable AI for clinical trustworthiness [15].

Reddy et al. (2025) proposed a framework for pneumonia detection using explainable deep learning MobileNetV2, transfer learning, and Grad-CAM for enhanced interpretability. The model was trained on the augmented CXR Pneumonia dataset, and obtained 92% accuracy. The model was deployed on Streamlit, which offers single and batch image uploads, and provides confidence scores and visual heatmaps. The authors found the model to be lightweight and interpretable, which makes it applicable for use in resource limited clinical environments. They suggest field testing the model on varied populations, increasing certainty in borderline cases, and multi-modal methods to address limitations of binary classification [16].

Vasilevschi et al, (2025) developed the implementation of a federated deep learning framework with a custom CNN and a VGG16 transfer-learning model for pneumonia detection in smart healthcare environments. The system was trained on the CXR Pneumonia dataset and evaluated in both centralized and federated environments with IID and non-IID client distributions. With more than 91% accuracy, the federated VGG16 model was successful. The planned next steps of the project involve clinical validation and real-world implementation, as well as detection and evaluation of pneumonia disease in multiple classes across various hospital imaging environments [17].

Mustapha et al. (2025) introduced a hybrid deep learning model that integrates CNN layers with modified Swin Transformer blocks to capture both local and global features for pneumonia detection. Using the Guangzhou Women and Children's Medical Center X-ray dataset with CLAHE enhancement and extensive augmentation, the model achieved 98.72% accuracy, outperforming a baseline CNN. Hyperparameter optimization via Optuna significantly strengthened robustness, reflected in high precision, recall, and F1-score 98.7%. Future work includes refining the architecture, adding explainable AI methods, and validating performance across more diverse and real clinical datasets [18].

In [19, 20] the author present vgg16 model for multi-class classification.

**Table 1. A review of recent studies for investigating different techniques of Pneumonia detection and classification.**

| Author/ Publication year | Machine or Deep Learning Model / Technique | Dataset | Accuracy | Limitations |
|---|---|---|---|---|
| Alshanketi et al. (2025) | VGG-16<br>ResNet50<br>ViT | BRAX<br>CheXpert<br>CXR | 92.6%<br>80.7%<br>86.5% | Model generalization across demographics limited. Zero-shot performance poor on small datasets. Dependency on ImageNet pre-trained models. GAN-based augmentation not tested. Semi-supervised methods limited to Mean Teacher only. |
| Aljawarneh et al. (2025) | CNN<br>VGG-19<br>ResNet-50 | CXR | 92.4%<br>87.3%<br>91% | Small and imbalanced dataset; heavy reliance on augmentation. No external validation; small test set. Computational limitations restricted deeper experiments Lack of explainability tools (e.g., Grad-CAM). |
| Yanar et al. (2025) | PELM | PadChes CXR<br>NIH CheXpert | 96% | Dataset artificially balanced (1:1), not reflecting real clinical prevalence.<br>Exclusion of low-quality images (blur, noise, poor contrast) may reduce real-world robustness. Deployment challenges: real-world validation, workflow integration, and prospective clinical testing still needed. |
| Mokarram et al. (2025) | ResNet-50<br>Resnet-101 | Cohen kermany | 96.04%<br>93.75% | Domain shift due to multiple heterogeneous datasets. Lack of explainability. Computationally heavier hybrid model. Risk of overfitting due to curated/augmented images. Only X-ray modality used. |
| Bhatt et al. (2023) | CNN | CXR | 84.12% | Low accuracy and precision despite excellent recall model prone to false positives. No external validation or multi-hospital testing, limiting clinical adoption.Model trained only on frontal X-rays no lateral or multimodal imaging. |
| Al Reshan et al. (2023) | MobileNet | CXR | 94.23% | Severe dataset imbalance; heavy reliance on augmentation. Unstable performance across optimizers (SGD dropped to 35%). |
| Nettur et al. (2024) | MobileNetV2<br>NASNetMobile | CXR | 97.10%<br>96.25% | Model trained on dataset with a small, imbalanced test split, causing limited generalization; no real-world validation or explainability. |
| Reddy et al. (2025) | MobileNetV2 | CXR | 92% | Model trained dataset with moderate confidence scores (60–70%) for some cases; requires external validation across different populations and imaging devices. |

| | | | | |
|---|---|---|---|---|
| Vasilevschi et al. (2025) | VGG-16 | CXR | 91% | Model trained on a public dataset under simulated federated conditions does not model hardware/network variability, and supports only binary classification. |
| Mustapha et al. (2025) | CNN | CXR | 98.72% | Needs clearer images for even higher performance, and hybrid model still requires improved preprocessing and architecture refinement. |

## 3. METHODOLOGY

The proposed pneumonia detection system follows a deep learning–based classification pipeline. Chest X-ray images are first collected and preprocessed to ensure uniformity and noise reduction. The processed images are then fed into a pre-trained VGG-16 model for feature extraction. Finally, fully connected layers perform binary classification to distinguish between normal and pneumonia cases. The overall framework consists of data acquisition, preprocessing, feature extraction using VGG-16, classification, and performance evaluation, as shown in **Figure 1.**

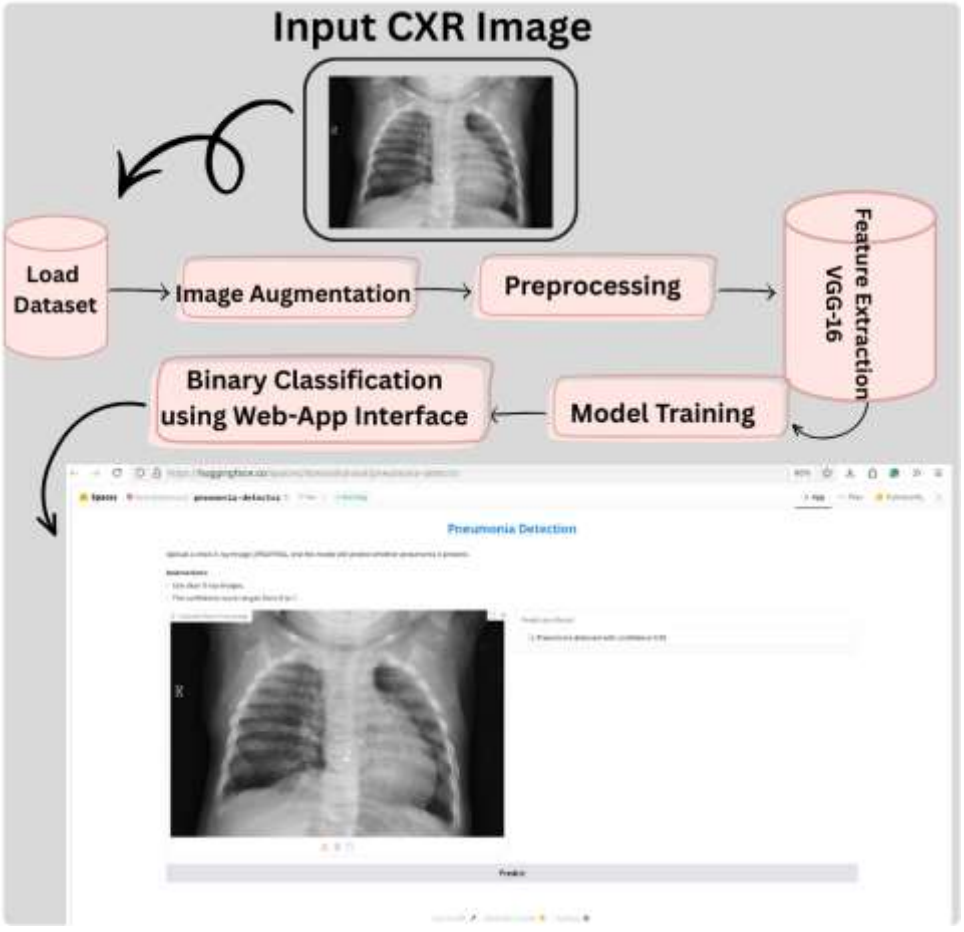


**Figure 1. Proposed framework of Pneumonia detection and classification Model**

### 3.1. Explanation of Dataset

The collected dataset with anterior posterior (AP) CXR images from a retrospective cohort of children aged 1 to 5 years at the Guangzhou Women and Children's Medical Center, where imaging was done as part of routine clinical practice is publicly available on kaggle [21]. Before collection, every radiograph in the dataset underwent a meticulous quality control review and was discarded if found to be unreadable and or blurry. A pair of expert radiologists provided each of their confirmations for the image diagnosis, as well as a third specialist who reviewed the evaluation set for the sake of consistency and in order to reduce the number of grading errors.

The dataset is made up of 5,856 annotated CXR images, making it suitable for binary classification as each image is classified as either Normal or Pneumonia. The images come as JPEG format and have varying resolution, each being over 1000 pixels across both dimensions and having file sizes between 100 and 500 KB. The dataset is split into two classes: 1583 Normal and 4273 Pneumonia. The clear skew in class distributions, where the number of pneumonia cases is far greater than the number of normal cases, will impact which aspects of the model the focus while evaluating performance. **Figure 2** shows a few representative CXR images for the Normal and Pneumonia classes while **Table 2** shows more details on the class-wise distribution of the dataset. Additionally, the class imbalance which motivates the need for augmentation and other techniques to balance the dataset during training. In **Figure 3** as a bar chart outlining the distribution of images across each class.

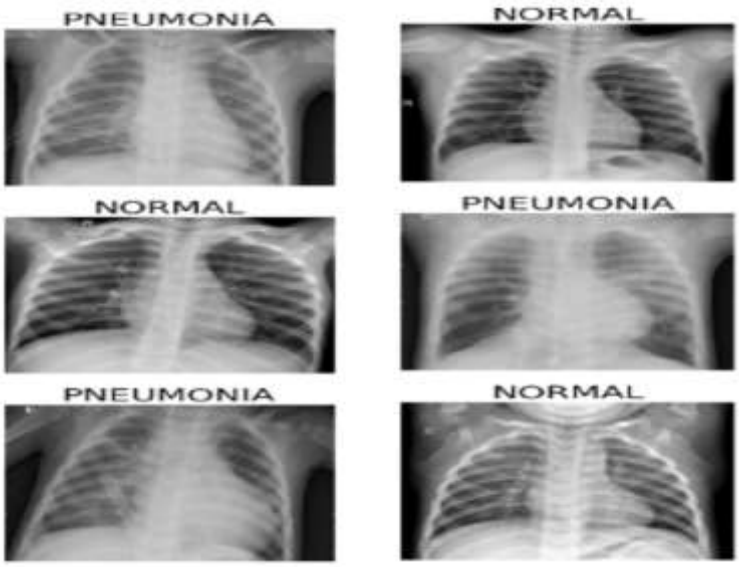


**Figure 2. Pictorial illustration from the chest X-Ray dataset.**

**Table 2. Dataset Description.**

| | |
|---|---|
| **Dimensions** | The dataset contains AP-view chest X-ray images with varying resolutions (e.g., 1782×1434, 1570×1164). |
| **Picture Type** | .JPEG |
| **Chest X-Ray Type** | |
| **Normal** | 1583 |
| **Pneumonia** | 4273 |
| **Total** | 5856 |

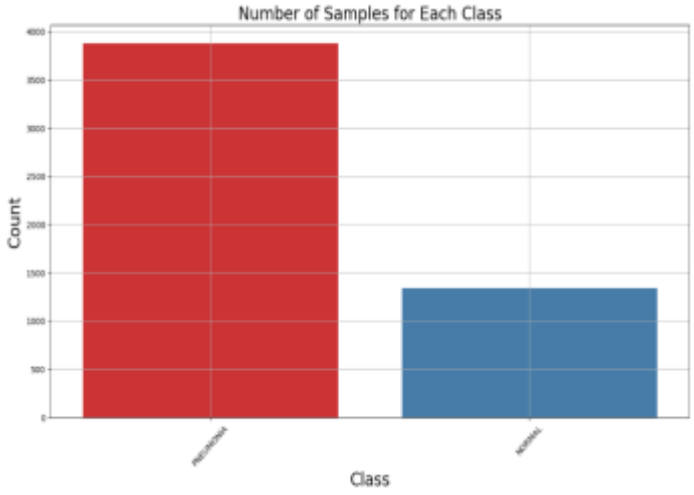


**Figure 3. Illustrate class distribution from the chest X-Ray dataset**

### 3.2. Implementation Environment and Tools

The studies were implemented using the Python 3.10.12 programming language and its corresponding deep learning libraries and data processing, model training and evaluation support tools. The main working environment was on MacBook with the Apple M1 chip, 16 GB of RAM, and the device was used for data preparation, code writing and evaluation of outcomes. For training of the models requiring a lot of computational power, the experiments were carried out in the Kaggle Notebook environment, which allows the use of GPUs, thereby shortening training periods and making deep learning more productive.

The main kaggle's software functionalities are used in this study include:

- The core implementation of the project was done using the Python programming language, version (v3.10.12).
- The deep learning framework utilized for the model development and training was Tensorflow, version (v2.15.0).
- The software used as a high-level API to build and fine-tune the VGG-16 model architecture on Keras, (v2.15.0).
- The process of manipulation, preprocessing and numerical calculations regarding the dataset is implemented on Pandas (v2.1.1) and NumPy (v2.24.4).
- The visualization and plotting the training, validation and testing curves results and other results of the experiment using Matplotib, (v3.7.2).
- To assess and create report of classification results and generate confusion matrices and other evaluative statistics by using Scikit-learn, (v1.2.2).

To avoid running out of memory in the computer while working with a complete dataset of CXR images, the Keras ImageDataGenerator was utilized, which allows images to be read one at a time from specified directories and only loads them into memory during the training or evaluation phase. Furthermore, the ImageDataGenerator library also permits image preprocessing functions to be applied in a way that modifies the original images in memory at the same time, thereby using memory more efficiently. Such a method also helps to enhance the generalization of the model by exposing it to a wider variety of the original data images.

### 3.3. Dataset Augmentation

This study utilized data augmentation techniques to increase the size and complexity of the training dataset to aid the deep learning model in training. Data augmentation is the process of

creating new training examples from existing training images. For this, label-preserving transformations such as rotation, horizontal flip, and zooming were used. The transformations utilized lead to a more diverse dataset and lower the chances of the model overfitting by providing the model with the tools to learn more flexible and robust features rather than memorize specific patterns. The model becomes more tolerant to the changes in the appearance of the CXR, and increases the model's performance on images that were not seen in the training dataset. **Figure 4** shows examples of CXR that were added after augmentation to the dataset in the Normal and Pneumonia classes. In **Table 3**, the class-wise representation and distribution of the dataset after augmentation is presented. **Figure 5** shows a bar-chart representation of the dataset, with a breakdown of images by class after augmentation, illustrating that the dataset is more diverse and increased the training opportunity for the model.

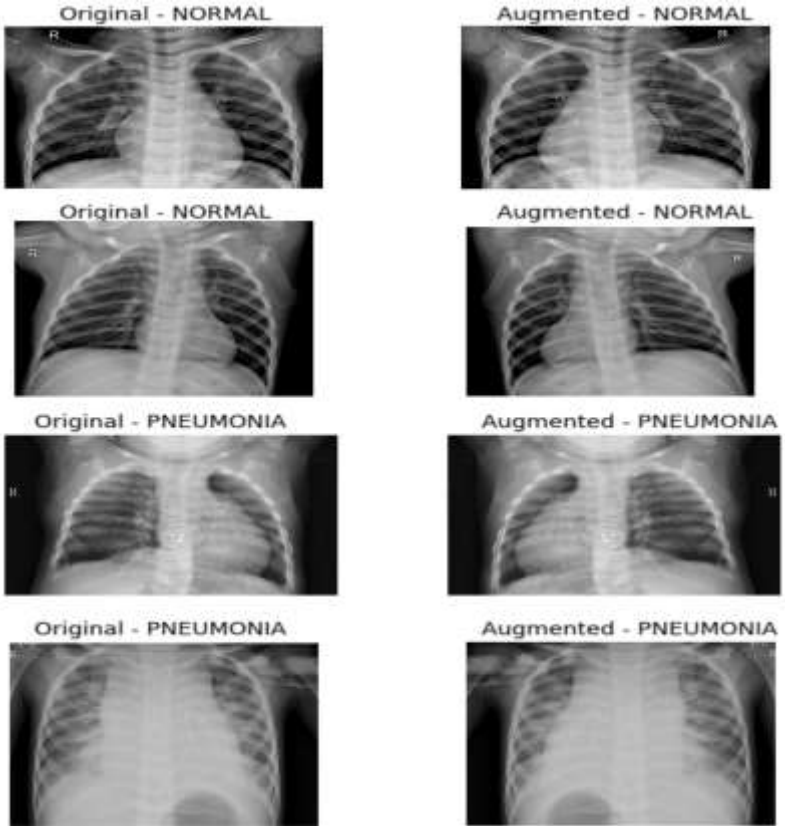


**Figure 4. Illustrates the few pictorial representations of our augmented dataset along with original images.**

**Table 3. Dataset Description after augmentation.**

| Chest X-Ray Type | Total number of Images before augmentation | Total number of Images before augmentation |
|---|---|---|
| **Normal** | 1,583 | 2,682 |
| **Pneumonia** | 4,273 | 7,750 |
| **Total** | 5,856 | 10,432 |

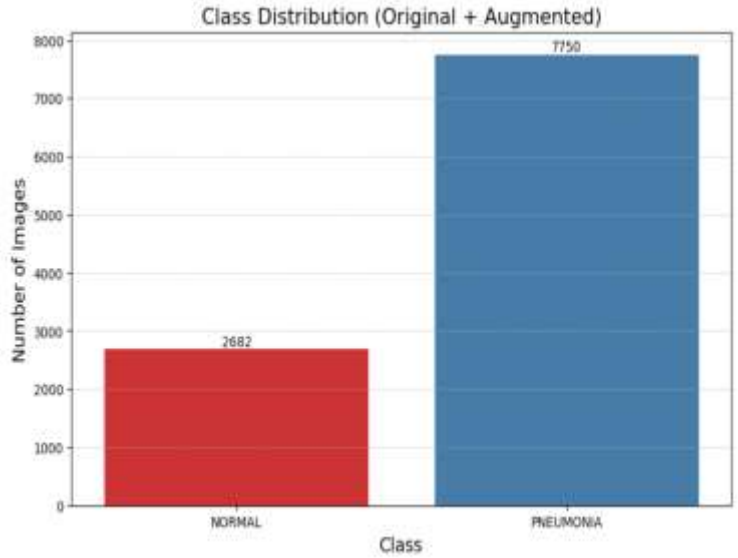


**Figure 5. Illustrate class distribution from the chest X-Ray dataset after augmentation**

### 3.4. Dataset Preprocessing

Several of the preprocessing steps, which were intended to streamline training and improve performance for the model, were applied to the CXR images prior to their use for classification. The steps included noise reduction, image labeling, resizing and normalization.

#### 3.4.1. Noise Reduction

Due to difficulties in image acquisition and the storage formats the CXR images contain noise. Consequently, the first preprocessing step was noise reduction, using filtering methods. With filtering techniques, an image is analyzed, and the pixels are replaced with values determined by specific algorithms, resulting in a "smoothing" of the image. This is particularly useful in removing noise such as salt and pepper noise in conjunction with preserving the edges of structures. More particularly, in this approach, the central pixel within the image is replaced with the median pixel value of the surrounding pixels. Since in many images noise pixels differ from the median in a substantial manner, this helps suppress the noise without blurring critical details in the images.

#### 3.4.2. Image Labeling

Supervised learning requires specific labels for the images within the dataset. Since the dataset was already sorted in folders corresponding to the individual classes, the labels were assigned directly based on the names of the folders. Each training image was assigned a label L, where L=0 is indicative of the Normal class and L=1 the Pneumonia class..

#### 3.4.3. Image Resizing

Due to the varying resolutions of the dataset images, a resizing procedure was therefore conducted to achieve a uniform input size for the deep learning model. All images were adjusted to be 224×224 pixels to satisfy the input specifications of the VGG-16 architecture. This increases the uniformity of feature extraction across the entire dataset while also simplifying the computational load.

#### 3.4.4. Normalizing

For enhancing the stability of the model training, and to speed up the convergence, the pixel intensity values were adjusted to a common scale. In the filtering of extreme values, and the masking of important data, normalizing the data in terms of its illumination levels was necessary. For this analysis, the pixel values were expressed and limited to the range of 0 to 1 by dividing by 255. This increased the efficiency and stability while also improving the performance of the model.

### 3.5. VGG-16 Architecture

Automated pneumonia detection using CXR images requires a deep learning pipeline based on transfer learning in this work. For binary classification into Normal and Pneumonia classes, the proposed technique uses VGG-16, which has been pre-trained on the ImageNet database, as a powerful feature extractor. Using a model 'from scratch' is avoided and instead the VGG-16 convolutional base is used with include_top=False to enable the network to derive meaningful hierarchically organized features from the medical images. This is depicted in **Figure 6** where the pre-trained convolutional layers create feature maps which are used for classification. 224×224 pixel resolution is set for all input images and normalized with pixel rescaling (1/255). **Table 4** illustrates more specifications for each layer.

To enhance generalization and reduce overfitting, on-the-fly data augmentation is applied during training using Keras' ImageDataGenerator, including random rotation, zooming, width/height shifting, and horizontal flipping. A customized classification head is added on top of the VGG-16 base model consisting of a Flatten layer, a Dense layer with 256 neurons (ReLU), and a Dropout layer (0.5), followed by a sigmoid output layer for binary prediction; the complete architecture is presented in **Figure 7**. The model is compiled using the Adam optimizer with a learning rate of $1\times10^{-4}$ and trained using binary cross-entropy loss. Data are loaded through directory-based generators (flow_from_directory) with class_mode='binary' for memory-efficient batch-wise training. The network is trained for 15 epochs with a batch size of 32, followed by an additional 20 epoch fine-tuning phase, and the final model is saved in H5 format for reproducibility and future deployment.

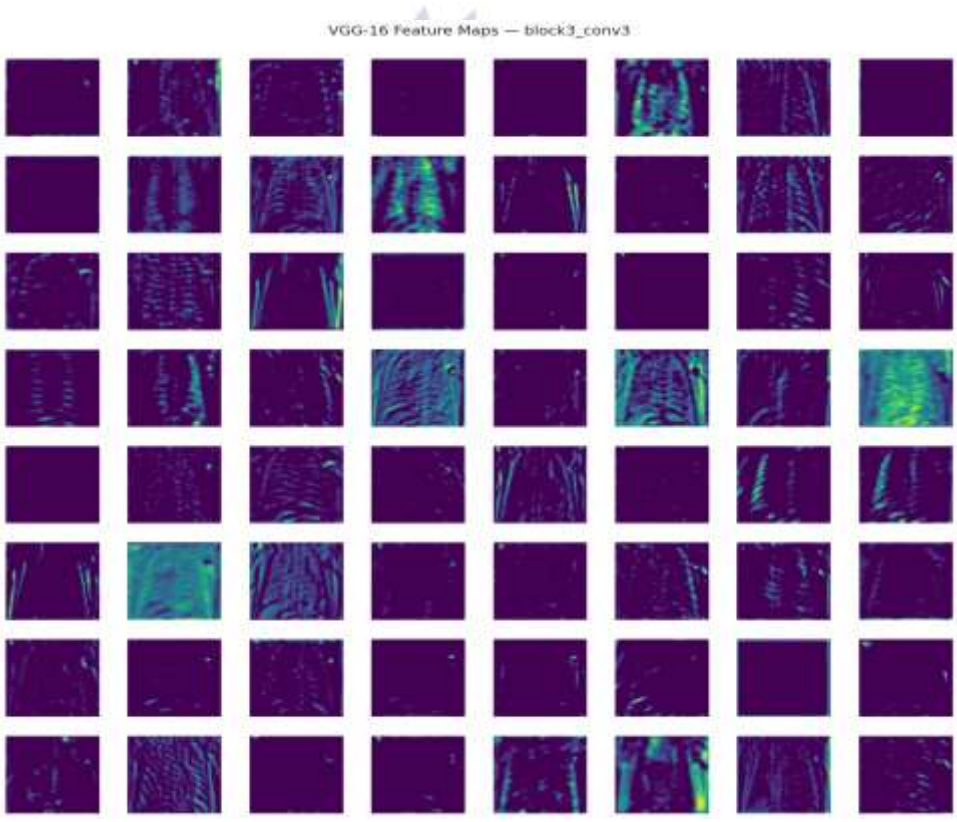


**Figure 6. Illustrates the Feature Extraction Map of VGG-16**

**Table 4. Steps of VGG-16 architecture for Pneumonia detection and classification Model.**

| Step no. | Step Name | Characteristics |
|---|---|---|
| 1 | Base Model | VGG-16 (ImageNet weights), include_top=False |
| 2 | Input Size | 224 × 224 × 3 |
| 3 | Task Type | Binary classification (Normal vs Pneumonia) |
| 4 | Frozen Layers | All VGG-16 base layers frozen initially (layer.trainable = False) |
| 5 | Data Split | Split into training and validation sets. Ratio: 80% training, 20% validation. |
| 6 | Callbacks | ModelCheckpoint: Save best weights. EarlyStopping: Prevent overfitting. |
| 7 | Classification Head | Image data normalized to range [0, 1].Flatten → Dense(256, ReLU) → Dropout(0.5) → Dense(1, Sigmoid) |
| 8 | Loss Function | Binary Cross-Entropy |
| 9 | Optimizer | Adam |
| 10 | Learning Rate | 1e−4 |
| 11 | Batch Size | 32 |
| 12 | Epochs | 10 epochs (initial training) + 10 epochs (fine-tuning) |
| 13 | Saved Model Format | .h5 file: pneumonia_vgg16_model.h5 |
| 14 | Evaluation | Model performance evaluated using accuracy and loss metrics. |

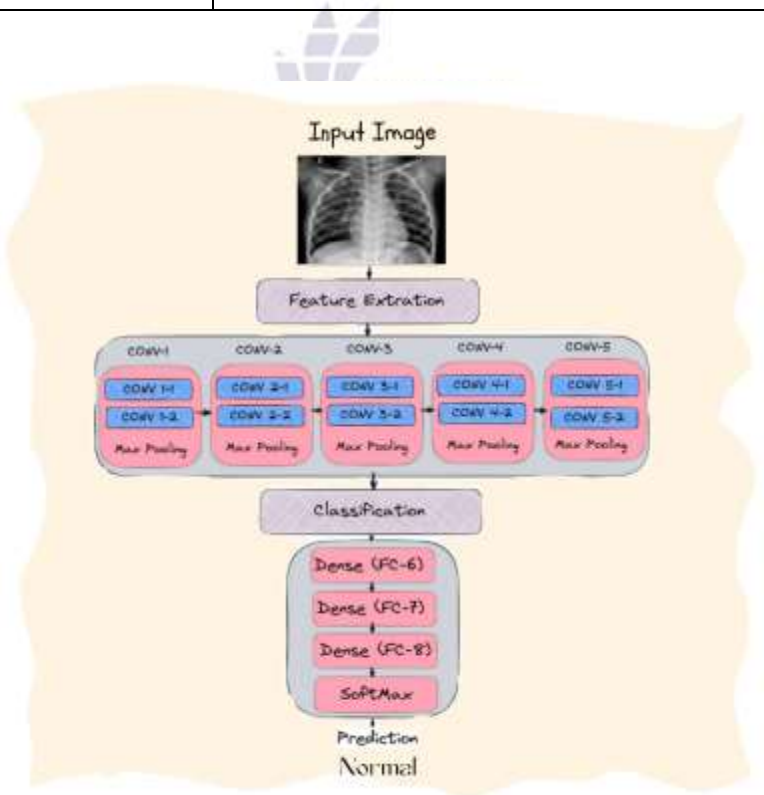


**Figure 7. VGG-16 basic Architecture**

### 3.6. Web App Deployment and User Interface

To facilitate real-time and user-friendly pneumonia screening, a complete web-based interface was developed and integrated with the trained deep learning model. The system includes a front-end landing page hosted on GitHub Pages [22], which acts as the entry point for users and provides direct access to the pneumonia detection platform. In **Figure 8**, the landing page has a prominent option saying "Open Pneumonia Detector" that allows users to access the deployed application. The detection system is integrated with a Gradio interface on Hugging Face Spaces, where users can submit CXR images in JPEG or PNG format, which will be used for predictions. The model classifies the images and provides a result with a confidence score to facilitate the understanding of the outcomes. The result of the model's prediction for a case of Pneumonia is illustrated in **Figure 9**. This model prediction is the most accessible form of the proposed method because it allows users to evaluate the model without needing to configure it locally, which demonstrates its practical usability and possible clinical usability.


**Figure 8. Web-App for the deployed pneumonia detection and classification system**


**Figure 9. Gradio-based Hugging Face deployment interface showing real-time pneumonia prediction with a confidence score.**

## 4. RESULTS AND DISCUSSION

In this section the evaluation of VGG-16 based pneumonia detection model using the CXR Pneumonia dataset is presented. The performance of the model was evaluated using the training, validation, and test splits, and the evaluation was done using model accuracy, precision, recall, F1 score. Confusion matrices and classification reports were also generated to describe model performance and misclassification patterns. The training accuracy and loss plots were generated to analyze the learning process and evaluate the model convergence and generalization.

**Table 5** provides information on the distribution of evaluation samples by class. In our experiments, the validation set had 939 images (698 Pneumonia, 241 Normal), while the test set had 522 images (388 Pneumonia, 134 Normal). To measure the performance of the proposed model, we employed a confusion matrix, which identifies and contrasts predicted and actual labels, indicating true and false identifications.

**Table 5. Detail of Experiment.**

| Class Type | Validation | Test |
|---|---|---|
| **Normal** | 241 | 134 |
| **Pneumonia** | 698 | 388 |
| **Total** | 939 | 522 |

### 4.1. Pneumonia Detection Confusion Matrix Analysis

In our research, **Figure 10** displays the confusion matrix of the predicted classes versus the actual classes as derived from our VGG-16 based pneumonia detection model. We concentrated our attention on the results of the validation set, and relaxed after accomplishing a validation set accuracy of 99.04% in 939 images. The confusion matrix indicates that the model successfully identified and classified 241 normal images and 689 pneumonia images. For the evaluation to be dependable, the dataset was partitioned into the training, validation, and test set, and model performance was monitored using the validation set. The model employed the Adam optimization algorithm set to a learning rate of $1\times10^{-4}$ and was trained for 20 epochs for optimal convergence, thereby ensuring a consistent validation accuracy.

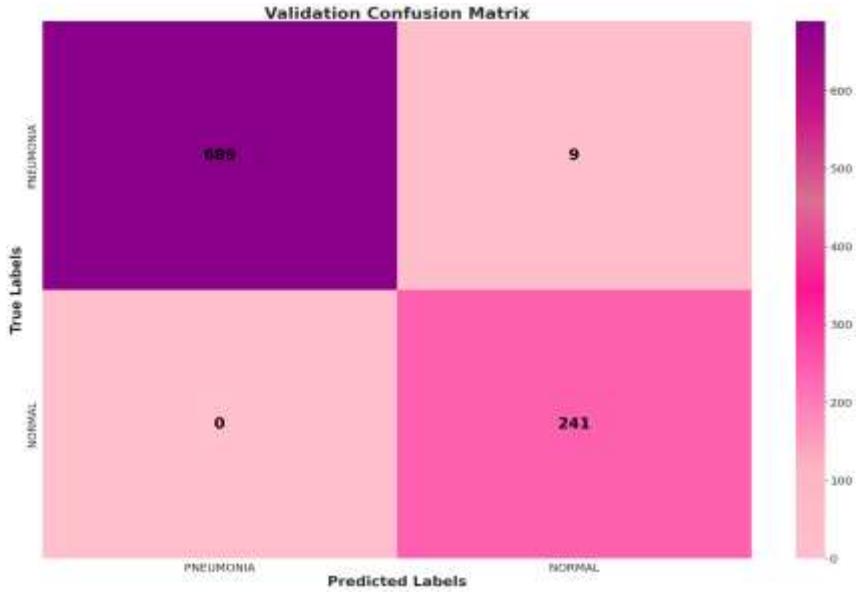


**Figure 10. The confusion matrix illustrated the results for validation set.**

**Figure 11** displays the confusion matrix of the predicted classes versus the actual classes as derived from our VGG-16 based pneumonia detection model. We concentrated our attention on the results of the validation set, and relaxed after accomplishing a validation set accuracy of 98.28% in 522 images. The confusion matrix indicates that the model successfully identified and classified 130 normal images and 383 pneumonia images. For the evaluation to be dependable, the dataset was partitioned into the training, validation, and test set, and model performance was monitored using the test set. The model employed the Adam optimization algorithm set to a learning rate of $1\times10^{-4}$ and was trained for 20 epochs for optimal convergence, thereby ensuring a consistent test accuracy.

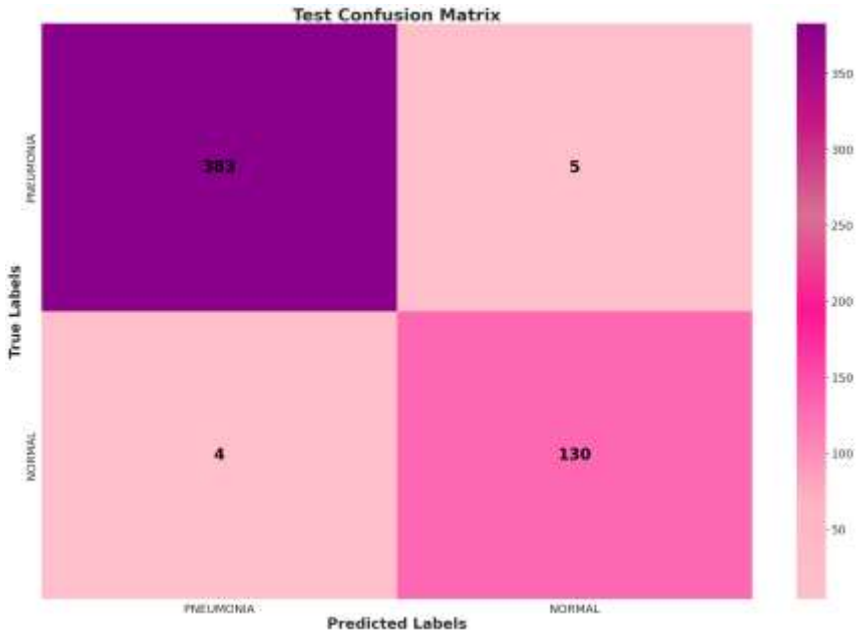


**Figure 11. The confusion matrix illustrated the results for test set.**

#### 4.1.2. Pneumonia Detection Classification Report

In addition to accuracy, we calculated precision, recall, and F1-score for each class to provide a more detailed performance analysis on the validation set. The classification report as show in **Figure 12** generated using the classification report function summarizes these evaluation metrics. Precision represents the proportion of correctly predicted positive cases out of all predicted positive cases, recall indicates the proportion of correctly predicted positive cases out of all actual positive cases, and the F1-score is the harmonic mean of precision and recall. For the validation set, the model achieved an overall accuracy of 99.04% on 939 samples, with the Normal class (label 0) obtaining precision = 1.00, recall = 0.99, and F1-score = 0.99, while the Pneumonia class (label 1) achieved precision = 0.96, recall = 1.00, and F1-score = 0.98. We also computed the overall validation misclassification count as 9, obtained by summing the off-diagonal values in the validation confusion matrix, which reflects the number of incorrect predictions.

```
Validation Set Evaluation:
Accuracy: 0.9904153354632588
F1 Score:
              precision    recall  f1-score   support

           0       1.00      0.99      0.99       698
           1       0.96      1.00      0.98       241

    accuracy                           0.99       939
   macro avg       0.98      0.99      0.99       939
weighted avg       0.99      0.99      0.99       939

Validation Misclassification: 9
```

**Figure 12. Represents the classification report in terms of accuracy, precision, recall, and f1-score of validation set.**

**Figure 13** shows the classification report which gives values of precision, recall, as well as the F1-score for the two classes and adds to the understanding of the model beyond just the accuracy. The model was able to achieve an total accuracy of 98.28% in the test set, having been able to accurately classify 522 samples. The Normal class (label 0) showed strong performance with precision = 0.99, recall = 0.99, and F1-score = 0.99, while the Pneumonia class (label 1) achieved precision = 0.96, recall = 0.97, and F1-score = 0.97. The total number of instances in the test set, which the model misclassified, was 9. This was calculated from the off-diagonal components in the test confusion matrix which shows the model has a high ability to generalize and also shows that its classification errors are minimal.

```
Test Set Evaluation:
Accuracy: 0.9827586206896551
F1 Score:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       388
           1       0.96      0.97      0.97       134

    accuracy                           0.98       522
   macro avg       0.98      0.98      0.98       522
weighted avg       0.98      0.98      0.98       522

Test Misclassification: 9
```

**Figure 13. Represents the classification report in terms of accuracy, precision, recall, and f1-score of test set.**

#### 4.1.3. Pneumonia Detection Visual Analysis Accuracy-Loss Curves

With respect to model evaluation and learning behavior, **Figure 14** shows the training, validation, and test accuracy curves over the multiple epochs. The x-axis represents epochs gone through, and the y-axis represents the model accuracy. The training accuracy blue line increases, and then reaches approximately 1.00. This shows us that the model has effectively learned the training data. The validation accuracy orange line is also high, and remains fairly consistent, effectively demonstrating no overfitting and strong generalization on the

model as it remains at 0.98–0.99. The dashed red line is the test accuracy, which follows validation test accuracy closely, demonstrating consistent performance on previously unseen CXR on image dataset.

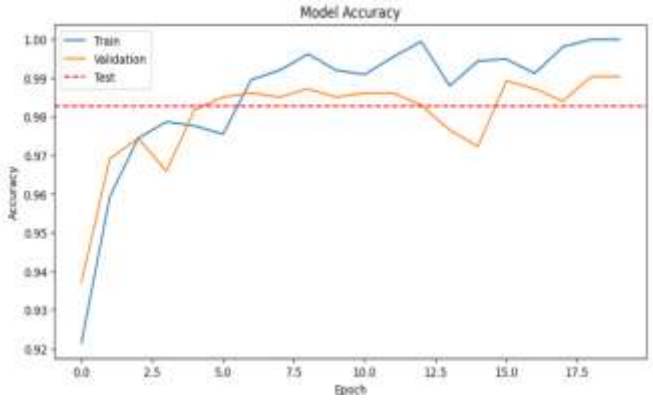


**Figure 14. Training, validation, and testing accuracy curve over 20 epochs.**

Likewise, **Figure 15** shows the training, validation, and test loss for each of the training epochs. Here we now focus on the loss curve, the validation and test loss ledge red line both remain fairly low, and at the endpoints, they capture the overall area. The training loss blue line on the other hand, is the one that shows a very low loss at the beginning, and then quickly drops to near zero, which we define low loss as indicated by increasing confidence of our predictions. In overall support of the model sufficiently demonstrating the ability to generalize and maintain low overfitting, the validation and test losses also remain low and closely aligned.

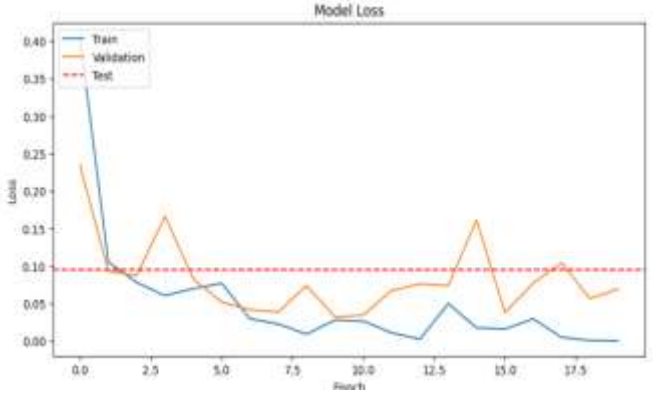


**Figure 15. Training, validation, and testing loss curve over 20 epochs.**

### 4.3.5 Visual Examples of Pneumonia Predictions

In order to showcase the effectiveness of the proposed model even more, there are examples of predicted outputs. The model predictions and corresponding actual labels are shown for some of the randomly selected examples of CXR in **Figure 16**. These displayed examples consist of both pneumonia and normal instances. The model demonstrates that he correctly identifies the class and shows agreement with the labels. Results like these qualitatively verify the model's findings and demonstrates the model's ability to identify patterns of pneumonia in normal lung structures.

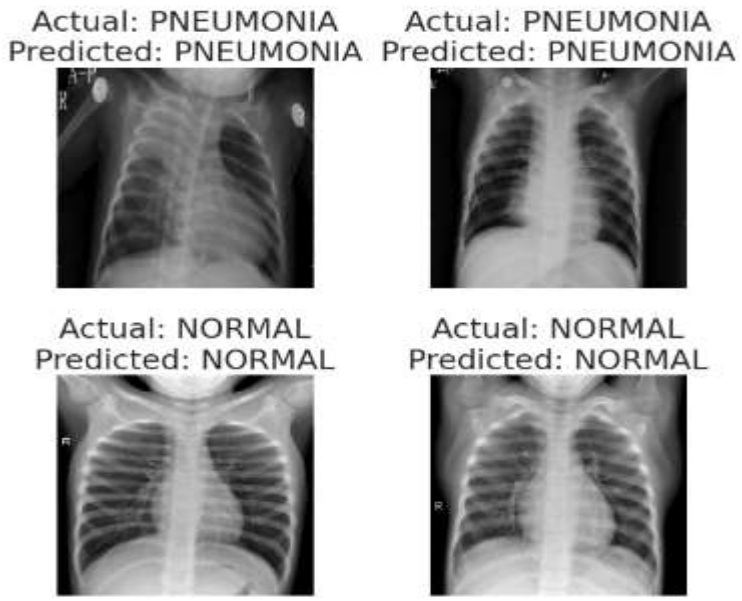


**Figure 16: Represent examples of predictions made by the model on chest X-rays with the actual labels and model predictions are included.**

Also, the model's performance and behavior in real-time was tested after being deployed via a web-based Gradio interface. The deployed application is shown in **Figure 17**, where the user is able to upload a chest X-ray, and the application outputs the predicted class along with a confidence score. The deployed application, where the user is able to upload X-ray images and the application outputs the predicted class with a confidence score, demonstrates that the model can be used in a practical, user-friendly way with no local execution required. These visual outputs confirm the model's pneumonia detection system is both usable and accurate.


**Figure 17: The real-time pneumonia prediction Gradio web application deployed to Hugging Face Spaces. After a chest X-ray is uploaded, a confidence score is provided as well.**

### 4.4. Results Analysis & Discussion

The experimental results show that VGG-16 transfer learning model for binary pneumonia detection from CXR images is optimally functioning. The accuracy curves indicate that learning is stable, with training accuracy reaching 1.00 and validation accuracy remaining consistently up to 0.99. Just the same, the loss curves describe training loss that is decreasing and validation loss that is relatively low, suggesting good convergence and that overfitting is not an issue.

The model validated as having an accuracy of 99.04% with 9 misclassifications out of a total of 939 samples. The confusion matrix describes 689 pneumonia images and 241 normal images that were correctly identified. Predictive misclassifications included only 9 pneumonia images that were identified as normal, and there were no normal images that were misclassified as pneumonia. The classification report also shows strong class-wise performance, with Normal class (precision = 1.00, recall = 0.99, F1 = 0.99) and Pneumonia class (precision = 0.96, recall = 1.00, F1 = 0.98). Lastly, for the final test evaluation, the model validated as having an accuracy of 98.28% with 9 misclassifications out of a total of 522 unseen samples, correctly identifying 383 pneumonia images and 130 normal images. Predictive misclassifications included 5 pneumonia images that were identified as normal and 4 normal images that were identified as pneumonia. This shows that for the sake of clinical safety, efforts to improve the reduction of remains a focus in evaluation summaries. The full breakdown of results is included in **Table 6**.

**Table 6. Performance evaluation.**

| Experiment | Accuracy% | Misclassification% | Precision% | Recall% | F1-Score% |
|---|---|---|---|---|---|
| **Validation** | 99% | 1% | 98% | 99% | 99% |
| **Test** | 98% | 2% | 98% | 98% | 98% |

Along with these evaluations, the model was used on Hugging Face Spaces, where it was built as a Gradio web app to generate real-time predictions and confidence scores. The overall approach is, on the one hand, accurate and pragmatic. On the other hand, future work needs to involve validation against different clinical datasets, explainability, and multiclass pneumonia classification to enhance the clinical trust. Also, recent studies on pneumonia detection are compared in **Table 7**, where the techniques, datasets, and results reported in existing studies are summarized and compared to the proposed method.

**Table 7. Comparison of existing studies on pneumonia detection.**

| Reference | Dataset | Technique | Accuracy% |
|---|---|---|---|
| [23] | CXR | InceptionNet_V2 | 97.23% |
| [24] | CXR | DenseNet121 | 90.06% |
| [25] | CXR | CNN | 91% |
| [26] | CXR | Federated Learning | 94%% |
| [27] | CXR | Res-WG-KNN | 97% |
| [28] | CXR | IFE | 94.93% |
| Proposed Model | CXR | VGG-16 | 99.04% |

## 5. CONCLUSION

An automated pneumonia detection system using transfer learning with the VGG-16 deep CNN architecture and CXR images was developed in this study. Through effective preprocessing and data augmentation, the proposed model attained

a high diagnostic performance with 99.04% validation accuracy and 98.28% test accuracy, supplemented with a great precision, recall, and F1-score, and a low count of misclassified instances. Evidence of the model's ability to generalize to new data was also provided in the confusion matrix and classification report. Apart from model development, a web-based application front-end interface was created and linked to the developed model so that users have an online platform to upload CXR images and receive predictions.

## REFERENCES

[1] WHO. Pneumonia "Pneumonia in Children". 2021. Available online: https://www.who.int/news-room/fact-sheets/detail/pneumonia (accessed on 24 May 2025).

[2] Sajed, S.; Sanati, A.; Garcia, J.E.; Rostami, H.; Keshavarz, A.; Teixeira, A. The effectiveness of deep learning vs. traditional methods for lung disease diagnosis using chest X-ray images: A systematic review. Appl. Soft Comput. 2023, 147, 110817.

[3] Sharma, S.; Guleria, K. A deep learning based model for the detection of pneumonia from chest X-ray images using VGG-16 and neural networks. Procedia Comput. Sci. 2023, 218, 357–366.

[4] Kanwal, K.; Asif, M.; Khalid, S.G.; Liu, H.; Qurashi, A.G.; Abdullah, S. Current diagnostic techniques for pneumonia: A scoping review. Sensors 2024, 24, 4291.

[5] Ahmadova, A.; Huseynov, I.; Ibrahimov, Y. Improving pneumonia diagnosis with RadImageNet: A deep transfer learning approach. Authorea 2023, 8, 25.

[6] Shirwaikar, R. A machine learning application for medical image analysis using deep convolutional neural networks (cnns) and transfer learning models for pneumonia detection. J. Electr. Syst. 2024, 20, 2316–2324

[7] An, Q.; Chen, W.; Shao, W. A deep convolutional neural network for pneumonia detection in X-ray images with attention ensemble. Diagnostics 2024, 14, 390.

[8] Shavkatovich Buriboev, A.; Abduvaitov, A.; Jeon, H.S. Binary Classification of Pneumonia in Chest X-Ray Images Using Modified Contrast-Limited Adaptive Histogram Equalization Algorithm. Sensors 2025, 25, 3976.

[9] Alshanketi, F., Alharbi, A., Kuruvilla, M., Mahzoon, V., Siddiqui, S. T., Rana, N., & Tahir, A. (2025). Pneumonia detection from chest x-ray images using deep learning and transfer learning for imbalanced datasets. Journal of imaging informatics in medicine, 38(4), 2021-2040.

[10] Aljawarneh, S. A., & Al-Quraan, R. (2025). Pneumonia detection using enhanced convolutional neural network model on chest x-ray images. Big Data, 13(1), 16-29.

[11] Yanar, E., Hardalaç, F., & Ayturan, K. (2025). PELM: A Deep Learning Model for Early Detection of Pneumonia in Chest Radiography. Applied Sciences, 15(12), 6487.

[12] Saber, A., Fateh, A., Parhami, P., Siahkarzadeh, A., Fateh, M., & Ferdowsi, S. (2025). Efficient and accurate pneumonia detection using a novel multi-scale transformer approach. Sensors, 25(23), 7233.

[13] Bhatt, H.; Shah, M. A convolutional neural network ensemble model for pneumonia detection using chest X-ray images. Healthc. Anal. 2023, 3, 100176.

[14] Reshan, M. S. A., Gill, K. S., Anand, V., Gupta, S., Alshahrani, H., Sulaiman, A., & Shaikh, A. (2023, May). Detection of pneumonia from chest X-ray images utilizing mobilenet model. In Healthcare (Vol. 11, No. 11, p. 1561). MDPI.

[15] Nettur, S. B., Karpurapu, S., Nettur, U., Gajja, L. S., Myneni, S., Dusi, A., & Posham, L. (2025). Lightweight Weighted Average Ensemble Model for Pneumonia Detection in Chest X-Ray Images. arXiv preprint arXiv:2501.16249.

[16] Reddy, K. T. (2025). Explainable deep learning for automated pneumonia detection from chest X-ray images using MobileNetV2 and Grad-CAM CNN. Authorea Preprints.

[17] Vasilevschi, A. M., Coman, C. A., Ianculescu, M., & Coman, O. A. (2025). Artificial Intelligence for Pneumonia Detection: A Federated Deep Learning Approach in Smart Healthcare. Future Internet, 17(12), 562.

[18] Mustapha, B., Zhou, Y., Shan, C., & Xiao, Z. (2025). Enhanced pneumonia detection in chest x-rays using hybrid convolutional and vision transformer networks. Current Medical Imaging, 21(1), e15734056326685.

[19] A. Nasir, G. Usman, U. Ahmad, A. Afzal, and Z. Nasir, "A Novel Approach for Pakistani Urdu Sign Language Alphabets Recognition and Classification using various Deep Learning Models with Hand Sign Gesture Data Generation," *Data Intelligence*, Sep. 2025, doi: https://doi.org/10.3724/2096-7004.di.2025.0179.

[20] A. Nasir, S. Zafar, and Z. Nasir, "Pakistani Currency Recognition and Classification for Visually Impaired People Using Convolutional Neural Network," pp. 1–6, Oct. 2024, doi: https://doi.org/10.1109/icodt262145.2024.10740188

https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

[22] https://itsmalaikanasir.github.io/AiProject/

[23] Rabbah, J., Ridouani, M., & Hassouni, L. (2025). Improving pneumonia diagnosis with high-accuracy CNN-Based chest X-ray image classification and integrated gradient. Biomedical Signal Processing and Control, 101, 107239.

[24] M. K. Dalei and S. Mahapatra, "Pneumonia detection and classification from chest X-Ray images using Modified DenseNet121 Deep Learning Approach," 2025 International Conference on Ambient Intelligence in Health Care (ICAIHC), Raipur Chattisgarh, India, 2025, pp. 1-5, doi: 10.1109/ICAIHC64101.2025.10956683.

[25] Saarvaan, K. K. K. R. (2025). Deep Learning Approach to Pneumonia Detection and Classification from Chest X-Ray.

[26] Rana, N., & Marwaha, H. (2025). Pneumonia detection from X-ray images using federated learning–an unsupervised learning approach. Measurement: Sensors, 37, 101410.

[27] Shati, A., Hassan, G. M., & Datta, A. (2025). A comprehensive fusion model for improved pneumonia prediction based on KNN-wavelet-GLCM and a residual network. Intelligent Systems with Applications, 26, 200492.

[28] Sotirov, S., Orozova, D., Angelov, B., Sotirova, E., & Vylcheva, M. (2025). Transforming Pediatric Healthcare with Generative AI: A Hybrid CNN Approach for Pneumonia Detection. Electronics, 14(9), 1878.