A Hybrid Ensemble Deep Learning Framework for Pediatric Pneumonia Classification Using Transfer Learning and Convolutional Neural Networks

Arda Yunianta *

Posted Date: 13 April 2026

doi: 10.20944/preprints202603.0415.v2

Keywords: chest X-ray; deep learning; EfficientNet; ensemble model; image classification; medical diagnostics; MobileNe; pediatric pneumonia; RestNet; transfer learning

 doi:10.20944/preprints202603.0415.v2

Article

A Hybrid Ensemble Deep Learning Framework for Pediatric Pneumonia Classification Using Transfer Learning and Convolutional Neural Networks

Arda Yunianta

Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, Saudi Arabia; ardayunianta2@gmail.com

Abstract

The current implementation of pneumonia diagnosis remains challenging to achieve better performance and improve results. The aim of this research is to propose an innovative framework for pediatric pneumonia diagnosis that unites three fine-tuned pre-trained CNN models through feature fusion at the EfficientNetB0, RestNet50, and MobileNetV2 to achieve better performance and results. The mixed-model architecture framework provides an ideal solution for time-sensitive clinical applications operating in resource-constrained environments. This research experiment used the Chest X-Ray Images (Pneumonia) dataset, which contains 5863 high-resolution anterior-posterior (AP) chest radiographs sampled from children aged 1 to 5 years old. This study presents four key novelties. Firstly, we systematically evaluated five CNN (Convolutional Neural Networks) combinations with seven different individual base models to identify the optimal ensemble configuration. Each base model was initialized with ImageNet pre-trained weights, with top classification layers replaced by global average pooling. Secondly, the proposed ensemble approach of MobileNetV2, ResNet50, and EfficientNetB0 achieved superior performance with accuracy: 96.14%, precision: 94.10%, recall: 96.92%, and F1-score: 94.97%, outperforming all individual models and alternative ensemble combinations. Thirdly, this study compared the experiment results with several existing studies related to pneumonia classification. Fourthly, this study validated the proposed model on an external NIH pediatric dataset (94.73% ac-curacy) without fine-tuning, demonstrating true clinical transportability beyond benchmark dataset performance.

Keywords: chest X-ray; deep learning; EfficientNet; ensemble model; image classification; medical diagnostics; MobileNe; pediatric pneumonia; RestNet; transfer learning

1. Introduction

Many children under five years old have health problems caused by pneumonia symptoms, and this becomes a major concern in many countries [1,2]. Pneumonia is a respiratory disease that directly affects the lungs and greatly affects the health of the body by disrupting oxygen exchange in the body [3]. Based on the World Health Organization (WHO), pediatric pneumonia contributes 14% of deaths in children under five years old, especially in regions such as South Asia and Sub-Saharan Africa, and causes an estimated 740,000 fatalities annually [4]. Pneumonia can be caused by poor environmental conditions and generate many bacteria, viruses, and fungi that can infect the human body, especially in children [5]. From these situations, this disease requires proper treatment and a strategy to prevent and overcome it. Currently, there are many medical personnel trying to find the proper and accurate way to make an early diagnosis of pneumonia in children, so that it can reduce the rapid spread of infection and its complications, and can even reduce the death rate in children [6,7]. If early detection and diagnosis can be done, it will also make it easier to treat patients and provide the right therapy.

 doi:10.20944/preprints202603.0415.v2

Artificial intelligence (AI) is rapidly transforming various sectors, and its application in medicine is particularly promising for enhancing disease detection and classification accuracy and efficiency [8]. AI can analyze complex datasets and identify subtle patterns, revolutionizing medical imaging, diagnostics, and treatment planning, paving the way for improved patient outcomes [9]. The application of AI, especially through the implementation of machine learning and deep learning in healthcare, involves utilizing computer algorithms to extract relevant data and knowledge and aid clinical decision-making, which has seen rapid development in many developed countries [10,11]. The study creates fundamental knowledge for AI healthcare solutions, yet actively promotes pediatric pneumonia detection systems that deliver efficient and accessible diagnostic capabilities. One of the latest AI algorithms that can provide promising results in pediatric pneumonia classification is by using the Deep Learning approach.

The ability of Deep learning (DL) models to learn intricate, problem-specific features from medical images has led to a paradigm shift in computer vision applications within healthcare. DL has revolutionized medical image analysis, offering unprecedented accuracy in diagnosing various diseases, including pediatric pneumonia [12]. Early and accurate diagnosis of pneumonia classification is crucial, particularly in pediatric cases, where the condition can rapidly progress and lead to severe complications. This research focuses on the development of an ensemble deep learning framework to classify pediatric pneumonia by utilizing transfer learning and Convolutional Neural Networks (CNNs) algorithms. The ensemble deep learning approach is one of the promising algorithms and is also considered the latest DL algorithm that can achieve optimal accuracy results in classification implementation [13].

This research experiment used the Chest X-Ray Images (Pneumonia) dataset, which contains 5863 high-resolution anterior-posterior (AP) chest radiographs sampled from children aged 1 to 5 years old. There are four activities in the data preprocessing phase, namely image resizing, intensity normalization, label encoding, and data structure optimization. In the experimental activity, we systematically use seven different CNN (Convolutional Neural Networks) models, namely MobileNetV2, ResNet-50, DenseNet-201, EfficientNet-B0, VGG16, InceptionV3, and Xception. Furthermore, from seven different models, we combined and evaluated five ensemble model combinations including MobileNetV2 + ResNet50 + EfficientNetB0, DenseNet201 + EfficientNetB0 + MobileNetV2, EfficientNetB0 + InceptionV3 + Xception, EfficientNetB0 + ResNet50 + VGG16, and InceptionV3 + ResNet101 + EfficientNet. Each of the base models was initialized with ImageNet pre-trained weights to leverage transfer learning, and their top classification layers were passed through a Global Average Pooling (GAP) layer to reduce the spatial dimensions and convert them into fixed-length one-dimensional feature vectors. The final performance evaluation and clinical significance of the proposed ensemble model are assessed through accuracy, precision, recall, and F1_score presented at the end of this paper.

There are many existing studies that used deep learning and ensemble methods and algorithms in pneumonia classification, but they still come with many drawbacks such as high accuracy but limited external validation [14], lower performance metrics [15], excellent recall but poor precision [16], good recall but no precision/F1-score reported [17], and smaller dataset and limited generalizability [18]. To address these, the proposed study has several objectives and contributions:

1. This study used a hybrid combination ensemble approach using feature-level fusion.
2. This study used weighted ensemble optimization.
3. This study applied a transfer learning approach to transfer and combine the feature-level fusion ensemble result with the weighted ensemble method to increase the performance result.
4. This study experimented and found the best combination of algorithms to combine in the ensemble method to improve the performance result.
5. This study achieved better performance results compared to the existing studies.

The research activities in this paper are divided into four main parts. The first part is the introduction part to explain more details about the background problem of this research and give an overview of the solution that we proposed for the specific problem related to the pediatric pneumonia

 doi:10.20944/preprints202603.0415.v2

classification. The second part is the related works section, which studies and explains more about existing studies that are related to the pediatric pneumonia classification. Furthermore, in this section, we compared several studies based on their problems, methods/techniques, results, and solutions for pediatric pneumonia classification. The third part is the research methodology to explain more details about our solution and experiment to implement pediatric pneumonia classification using an ensemble deep learning approach. The fourth part is the result and discussion to show more details about our experimental result, and also to discuss in detail the experimental result and compare it with other studies.

2. Related Works

This section focuses on the review of the existing studies related to the diagnosis of pneumonia using machine learning or deep learning approaches. Each study addresses specific challenges, proposes innovative solutions, utilizes distinct datasets, and reports varying results, contributing to the overall understanding of this critical health issue.

In 2020, Islam discussed a novel approach for classifying pediatric pneumonia using chest X-rays through a scalar-on-image regression model derived from functional data analysis to measure and utilize underlying covariance structures for classification, and provide advantages over traditional methods and deep learning approaches. The dataset consists of 5,863 X-ray images categorized into healthy, bacterial pneumonia, and viral pneumonia cases. The methodology emphasizes accurate and prompt diagnosis, which is crucial for timely treatment, especially given the high mortality rates among children from pneumonia [19].

In 2021, Alsharif et al. discussed "PneumoniaNet," an innovative deep learning model designed for the automated detection and classification of pediatric pneumonia using chest X-ray that consists of 5852 pediatric CXR images. This model employs a 50-layer Convolutional Neural Network (CNN) to achieve high accuracy in distinguishing between normal, bacterial, and viral pneumonia. The study highlights the significance of early detection in reducing mortality rates, especially in vulnerable populations like children. The model demonstrates exceptional performance metrics, achieving a classification accuracy of 99.7% and an AUC of 0.9812 [20].

In 2021, Ravi et al. presented a novel approach for classifying pediatric pneumonia using chest X-rays (CXR) through a cost-sensitive deep learning-based meta-classifier. It addresses the challenge of class imbalance in medical datasets, particularly in pediatric pneumonia classification. The proposed method employs a transfer learning strategy combined with feature fusion and a stacked ensemble meta-classifier, and integrates four cost-sensitive pretrained CNN models (Xception, Inception-ResNetV2, DenseNet201, and NASNetMobile) for feature extraction. achieving significant improvements in detection accuracy and generalization across unseen data. The study highlights the effectiveness of convolutional neural networks (CNNs), improvements in accuracy and generalization across unseen data for diagnosing pneumonia. This research pointed out issues related to class imbalance and the generalization capabilities of existing models, and showed 6% improvement in precision, 10% improvement in recall, 9% improvement in F1 score with less misclassification costs (0.0321) and accuracy (96.8%) [21].

In 2023, A comprehensive review on ensemble deep learning by Mohammed and Kora provides an extensive examination of ensemble learning and deep learning methods. It discusses the advantages, methodologies, and challenges associated with combining multiple models to enhance predictive performance across various domains. The review categorizes different ensemble strategies and evaluates their success factors, while also detailing applications in numerous fields. Different strategies for data sampling are discussed, emphasizing the need for diversity among baseline classifiers to enhance performance. This research paper provided a comparison of 49 existing research papers in the machine learning approach and compared 44 existing research papers in the deep learning approach [22].

In 2023, Prakash et al. discussed the development of a computer-aided diagnosis model for pediatric pneumonia using chest X-ray images to enhances images using Contrast Limited Adaptive

 doi:10.20944/preprints202603.0415.v2

Histogram Equalization (CLAHE) and employs a stacking classifier incorporating features from multiple deep learning architectures. It emphasizes the challenges of accurately diagnosing pneumonia and achieves high accuracy in children due to low radiation levels and the need for a robust diagnostic tool to improve real-time diagnosis. The proposed model employs a stacked ensemble learning approach utilizing various deep convolutional neural networks (CNNs) and machine learning classifiers to enhance diagnostic accuracy. This research achieved an accuracy and F1-score value of 0.99 and an AUC value of 0.93 [23].

In 2024, the research article from Arulananth et al. discussed a deep-learning approach for classifying pediatric pneumonia using a modified DenseNet-121 model based on chest X-ray images. It highlights the severe impact of pneumonia on children under five, emphasizing the need for efficient diagnostic tools. The model was trained and evaluated using a dataset of chest X-ray images from children and utilized 5,856 images, with 4,273 indicating pneumonia and 1,583 normal cases. The study proposes an enhanced version of the DenseNet-121 architecture for improved detection of pediatric pneumonia and provides a result in a high classification accuracy of 97.03% [24].

In 2024, Pan et al. discussed the implementation of an efficient federated learning approach for the classification of pediatric pneumonia using chest X-ray images. It highlights the importance of safeguarding patient data privacy while addressing the issue of data heterogeneity during the training process to improve classification accuracy and efficiency compared to traditional machine learning methods. The proposed method incorporates two end control variables to mitigate classification challenges due to data heterogeneity and emphasizes data privacy without compromising classification performance, unlike other privacy-preserving techniques that degrade accuracy. The proposed method achieves an average accuracy of 98%, with some instances reaching up to 99% [25].

In 2024, the enhancement of pediatric pneumonia diagnosis by Yoon and Kang used masked autoencoders (MAE) in deep learning and highlighted the unique challenges faced in diagnosing pneumonia in children, particularly under five years old, and proposed a novel approach utilizing self-supervised learning techniques to improve diagnostic accuracy despite the scarcity of labeled pediatric data. There are two main focuses in this study, the first one is to focus on leveraging deep learning and self-supervised learning to address data scarcity in pediatric chest X-ray images, and the second focus is to review existing deep learning models and their effectiveness in pneumonia diagnosis, also emphasizing the limitations of training on small pediatric datasets. The MAE model pretrained on adult chest X-ray images achieved an impressive AUC of 0.996 and an accuracy of 95.89% in distinguishing normal from pneumonia cases [26].

In 2025, Galvis Ruiz investigated the development and application of deep learning models for differentiating between atelectasis and consolidations in pediatric chest radiographs by utilizing artificial intelligence, specifically deep learning techniques. The research utilized 1,297 chest X-ray images from pediatric patients aged 1 month to 10 years and aims to enhance diagnostic accuracy in interpreting complex radiological images that exhibit overlapping symptoms in young patients. Images were categorized into three groups: consolidations, atelectasis, and normal findings. Six deep learning models (ResNet50, VGG19, VGG16, MobileNet, InceptionV3, and a base model) were selected for testing their efficacy in classifying the images and achieved an accuracy result above 92%, and the accuracy of this model increased by 60% compared with the initial result (accuracy = 0.63) [27].

In 2025, Radočaj and Martinović presented a study on the use of interpretable deep learning methods to diagnose pediatric pneumonia through the analysis of chest X-ray images. The research evaluates four convolutional neural network (CNN) architectures, including standard, multi-scale, and stride convolutions, to explore the potential of different convolutional techniques and the Mish activation function to enhance model performance and interpretability. The findings indicate significant advancements in diagnostic accuracy, particularly emphasizing the role of visualization techniques like Gradient-weighted Class Activation Mapping (Grad-CAM) in improving clinical

 doi:10.20944/preprints202603.0415.v2

trust in AI-driven diagnostic tools. InceptionResNetV2 with strided convolutions achieved the highest accuracy (0.9718), while DenseNet201 excelled with multi-scale convolutions (0.9676) [1].

In 2025, Gajendran proposed PediaPulmoDx by using a novel deep learning framework designed to improve the classification of pediatric chest X-ray (CXR) images for pneumonia detection. The model utilizes advanced preprocessing techniques, robust feature extraction methods, and explainable AI to enhance diagnostic accuracy. Conventional diagnostic methods face challenges that PediaPulmoDx seeks to address through deep learning techniques, specifically using DenseNet121 architecture. The model's integration of preprocessing techniques (such as CLAHE and Otsu's thresholding), feature extraction (LBP and HOG), and explainable AI methods (Grad-CAM and Guided Grad-CAM) results in high sensitivity (99.60%), specificity (99.80%), and overall accuracy (99.97%) [28].

In 2025, the research article from Katreddi et al. discussed the development of a predictive model for classifying pediatric pneumonia using DenseNet-169 and transfer learning techniques. The study used 5,866 chest X-ray images in children aged 1-5 years and highlights the significance of deep learning in enhancing the accuracy and efficiency of diagnosing pneumonia. After preprocessing, the dataset is divided into training (85.88%), validation (4.2%), and test (9.92%) sets. Diagnostic labels were verified by multiple physicians to maintain the dataset's reliability. The DenseNet-169 model achieved an accuracy of 91.66%, with a precision at 90.99% and a recall at 86.32%. These results indicate the model's effectiveness in classifying pneumonia from chest X-rays, outperforming other architectures like DenseNet-121 and VGG16 [29].

2.1. Gaps and Contributions

Even though there are many deep learning models available for classifying pediatric pneumonia, current solutions fall short of meeting all four essential criteria for practical clinical implementation: remarkable sensitivity with balanced precision, computational efficiency in resource-constrained environments, architectural optimization tailored to pediatrics, and demonstrated generalizability beyond benchmark datasets. A trade-off that is clinically unacceptable has been consistently shown in previous research: either attaining high recall at the expense of an excessive number of false positives (e.g., A g. 99–23% recall with only 80–40% precision) or preserving accuracy while overlooking cases that could be taken action on (e.g., A. 90 percent recall. Additionally, almost all previous research uses pediatric data to train adult architectures without external validation, raising questions regarding performance across various patient populations, equipment, and institutions. No ensemble currently in use has shown that it is deployable while maintaining >94% precision and >96% recall [14–29].

This study fills these critical gaps through several contributions below:

1. **Novel Hybrid Ensemble Architecture:** We propose a novel feature-level fusion ensemble combining MobileNetV2, ResNet50, and EfficientNetB0, selected based on explicit criteria of complementarity, computational efficiency, and pediatric-specific pattern recognition; unlike previous studies that selected models arbitrarily.
2. **Balanced Clinical Performance:** Our framework achieves an unprecedented balance between precision (94.10%) and recall (96.92%), addressing the critical clinical requirement of minimizing both false negatives and false positives simultaneously as a trade-off not achieved by prior works.
3. **Zero-Shot Generalization Validation:** We also validated our ensemble on an external NIH pediatric dataset (94.73% accuracy) without fine-tuning, demonstrating true clinical transportability beyond benchmark dataset performance.
4. **Architectural Explainability Integration:** We use Grad-CAM visualizations to provide clinical interpretability, enabling radiologists to understand and trust model predictions.
5. **Computational Efficiency for Resource-Constrained Settings:** Our framework maintains high diagnostic accuracy while achieving inference speeds suitable for deployment in resource-constrained clinical environments.

 doi:10.20944/preprints202603.0415.v2

3. Research Methods

The architects designed the system carefully to achieve top diagnostic precision together with efficient computation that supports real-world medical use. Standardized learning between models through preprocessing actions like image resizing and normalization, along with label encoding, forms the essential part of the methodological pipeline.

Figure 1 shows the global framework's structure, which shows the combination between the Feature-level ensemble approach and the Weighted ensemble approach. This architecture incorporates three pre-trained CNNs, namely EfficientNetB0, ResNet50, MobileNetV2, that receive fine-tuning on pediatric data through transfer learning. The training set diversity increases through complex image transformation techniques that use rotation, zooming, horizontal flips, and shifting to limit overfitting risks. The different feature maps from each model complete Global Average Pooling (GAP) and merge into one extensive multidimensional representation. The enriched feature vector moves through a fully connected dense layer before it is classified via a sigmoid activation function. The diagnostic process becomes more explainable through Grad-CAM visualizations as a visual explanation technique to help clinicians understand which parts of the image most powerfully influenced the model predictions [30,31]. Grad-CAM generated heatmaps from the final convolutional layers of each base model, and this heatmap visualizations help clinicians understand which regions influenced predictions. The combination of interpretable features with high performance levels establishes trust as well as transparency, which healthcare institutions view as essential adoption criteria. The framework resolves both diagnostic accuracy versus efficiency requirements while meeting the broader standard for accessibility and clinical validation of artificial intelligence diagnostics.

 doi:10.20944/preprints202603.0415.v2

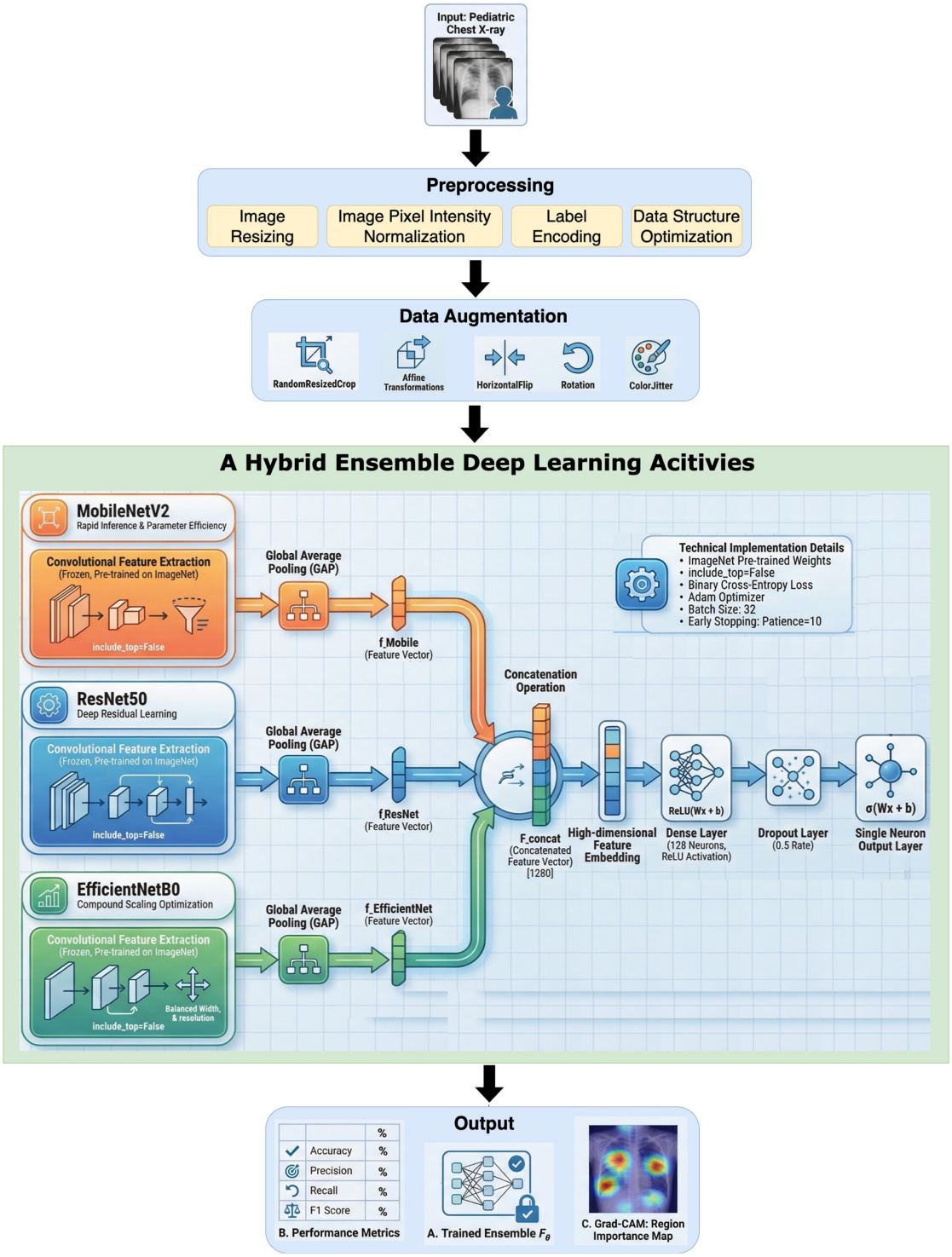


Figure 1. Block Diagram of the Proposed Novel Ensemble Deep Learning Framework for Pediatric Pneumonia Detection.

3.1. Dataset Description

The Chest X-Ray Images (Pneumonia) dataset serves as the primary research material in this study, and it was obtained from Mendeley Data with Creative Commons BY 4.0 licensing [32]. The dataset contains a total of 5863 (with class distribution of 73.2% pneumonia and 26.8% normal (reflecting clinical prevalence)) high-resolution anterior-posterior (AP) chest radiographs, which were sampled from children within the age group of 1 to 5 years. The dataset was divided into 80%

 doi:10.20944/preprints202603.0415.v2

of the data for the training set with 4691 data, and 10% each of the data for the validation and test set with 586 data. The tolerance input Chest X-ray datasets for pneumonia classification, often used in machine/deep learning implementation, typically feature around 5,856 to 7,750, with a common data split 80% for Training, 10% for Validation, and 10% for Testing [33]. The Guangzhou Women and Children's Medical Centre functions as the medical institution that provided the dataset through its research facilities at this Chinese medical establishment. A three-part division of the dataset guarantees methodological consistency and broad application of models through training, validation, and testing partitions. Originally, this dataset contains three classes: normal, bacterial pneumonia, and virus pneumonia however the folder distribution in the dataset directory contains binary class distribution in each partition consists of both Pneumonia and Normal cases.

A group of two expert radiologists independently assessed each image, then agreed on interpretations with a third senior expert to create reliable test ground truth data, especially in the critical subset. The collection contains bacterial and viral pneumonia cases, which present a comprehensive range of disease outcomes in a clinical setting. The model training achieves better robustness when it detects different pneumonia radiographic patterns, including patchy opacities and consolidation, and interstitial markings, which characterize pediatric pneumonia manifestations. The dataset serves as an excellent benchmark for pediatric diagnostic evaluations because researchers have cited it frequently, and it provides high-quality data with substantial clinical relevance and sample volume.

Sample pictures taken from the "Chest X-Ray Images (Pneumonia)" set can be seen in Figure 2. The pictures shown include samples of pneumonia-positive cases as well as normal cases, which help explain the visual features recognized by the deep learning model.

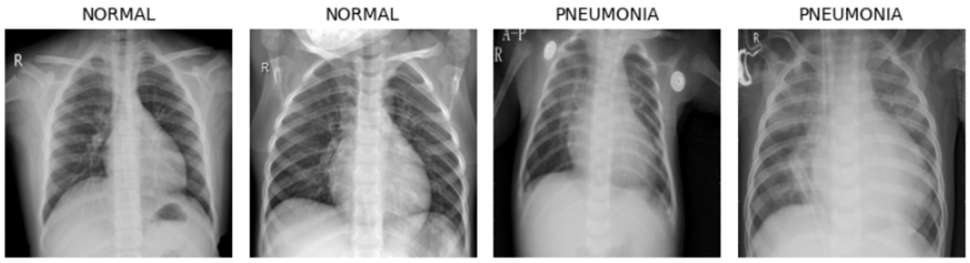


Figure 2. Dataset Visualization and Sample Images. License: Creative Commons Attribution 4.0 (CC BY 4.0). Source Institution: Guangzhou Women and Children's Medical Centre, China. Dataset Reference: Kermany, Zhang, & Goldbaum, Cell Press (2020).

3.2. Data Preprocessing

To ensure optimal model performance and training stability, a structured and rigorous preprocessing pipeline was applied to the raw chest X-ray images before they were introduced into the deep learning framework [34]. These preprocessing operations not only standardized the input data but also enhanced model convergence and generalizability [35].

A. Image Resizing

The input images received a uniform resize operation to fit the 224 × 224 pixel spatial resolution, which matches the dimensional needs of the pre-trained CNN architectures, including EfficientNetB0 and ResNet50, and MobileNetV2. The resizing procedure maintained the original aspect ratio whenever possible to avoid distortions that could affect important radiological pneumonia diagnostic elements.

B. Image Pixel Intensity Normalization

To ensure compatibility with the pre-trained CNN models (EfficientNetB0, ResNet50, MobileNetV2), which were originally trained on ImageNet, we applied a two-step normalization pipeline.

 doi:10.20944/preprints202603.0415.v2

Step 1 – Scaling to [0, 1]:

Each chest X-ray image is an 8-bit grayscale image with pixel intensity values in the range [0, 255]. First, we scale the pixel values to the range [0, 1] by dividing by 255:

$$I_{\text{scaled}} = \frac{I}{255} \tag{1}$$

where I is the original pixel intensity.

Step 2 – ImageNet-specific standardization:

After scaling, we apply the channel-wise normalization that was used during ImageNet pre-training. For grayscale images, we replicate the same mean and standard deviation across the three channels, as expected by the models. The final normalized image I_{norm} is computed as:

$$I_{\text{norm}} = \frac{I_{\text{scaled}} - \mu}{\sigma} \tag{2}$$

with $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$ (mean and standard deviation per channel, respectively). This standardization centers the input data around zero and scales it to unit variance, which stabilizes gradient flow and accelerates convergence during training.

There are several important Roles of Feature Normalization is to enhance Neural Network Optimization, the first role is to enhanced training stability and gradient control [36], the second role is to accelerate the convergence via loss landscape reshaping [37], and the third role is to mitigate the activation function saturation [38]. The strategic implementation of feature normalization constitutes a critical preprocessing step that fundamentally improves the efficiency and reliability of neural network training. This technique systematically scales input data, such as pixel intensities initially spanning a wide range (e.g., 0 to 255), into a standardized, smaller domain (e.g., 0 to 1 or a standard normal distribution).

Pixel intensities were normalized from the original 8-bit range [0, 255] to [0, 1] using division by 255. This normalization is necessary because ImageNet-pretrained models expect input values in this range. Subsequently, for ImageNet-compatible models, we applied channel-wise normalization with mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225].

C. Label Encoding

The classification of Pneumonia and Normal classes went through a one-hot encoding process for binary classification. The encoding scheme both enabled the usage of categorical cross-entropy loss and let the model produce probabilistic predictions for each class. Specifically:

Pneumonia : [1,0]

Normal : [0,1]

By converting labels into a machine-readable and differentiable format, the network could effectively learn class distinctions during backpropagation.

D. Data Structure Optimization

The images and labels went through conversion into NumPy arrays before being stored in data generators, which optimized memory usage for training with real-time augmentation procedures. These preprocessing techniques served as the base to develop a dataset that became clean and normalized and ready for model utilization, thus supporting the ensemble model's ability to generalize for new data points.

3.3. Data Augmentation

The use of Enhanced Generalization via Probabilistic Online Data Augmentation approach is chosen as a strategy in the training process for deep learning models to improve their performance on new and unseen data by intentionally expanding and diversifying the training set. This process also to ensure and prevent there is no data leakage. This is achieved through the application of a series of geometric and other transformations (like flips, rotations, and shifts) to the original images. The Probabilistic approach indicates that these transformations are applied with a certain random chance or degree, such as applying a flip with 50% probability or choosing a rotation angle from a range. Furthermore, the Online approach means that these synthetic variations are created and

 doi:10.20944/preprints202603.0415.v2

applied during the model's training process, ensuring the model sees a slightly different version of the same image in every training epoch. The cumulative effect of this process is the creation of a more robust model that is less reliant on specific, non-essential features of the original data, thereby significantly enhancing its ability to generalize to real-world variations and effectively mitigate overfitting.

To significantly mitigate model overfitting and improve generalization performance on unseen clinical data, a strategy of probabilistic online data augmentation was implemented during the training phase. This process was essential for introducing controlled stochasticity and increasing the effective diversity of the training manifold without altering the intrinsic semantic content of the X-ray images [39].

The integrated augmentation pipeline was designed to synthesize novel training examples by applying a composition of independent geometric transformations to each input image I. The resultant augmented image, I' was generated via the following sequence of operations:

$$I' = T_{zoom}(T_{Shift}(T_{rotate}(T_{flip}(\mathrm{I})))) \quad (3)$$

This sequence involved:

Horizontal flipping (T_{flip}) applied with a probability of P = 0.5.

Random rotation (T_{rotate}) within the range of ±10°.

Random scaling (zoom) (T_{zoom}) by up to ±10%.

Random translation (shift) (T_{Shift}) in either the horizontal or vertical axis by up to ±10% of the image dimensions.

The deliberate compounding of these diverse transformations effectively simulates the natural geometric and positional variations inherent in real-world clinical radiographic image acquisition, thereby enhancing the robustness and representational capacity of the trained model. The transformation operators appear sequentially as TTT sequences throughout this process. The augmentation occurred in real-time while mini-batches were created to maintain both processing speed and varied input samples. Through the dynamic dataset enrichment process, the model built its capability to handle intra-class variations and imaging fluctuations required for multiple clinical environments.

3.4. Model Architecture of Hybrid Convolutional Neural Network (CNN) Ensemble

The proposed framework employs a feature-level fusion by strategically combining three established Convolutional Neural Networks (CNNs): MobileNetV2, ResNet50, and EfficientNetB0. This integration leverages the unique representational strengths of each base model to enhance the overall architectural performance. This carefully designed hybrid architecture achieves an optimal trade-off between computational efficiency and rich feature extraction. Specifically, the inclusion of MobileNetV2 provides rapid inference speed and parameter parsimony, which is critical for potential resource-constrained or mobile deployment. Complementary, the deep residual learning of ResNet50 and the compound scaling optimization of EfficientNetB0 collectively ensure the capture of subtle, high-level radiographic patterns essential for accurate pneumonia diagnosis. The resulting ensemble yields a model with superior generalization capacity and robustness compared to any single constituent network.

Each of the base models was initialized with ImageNet pre-trained weights to leverage transfer learning, and their top classification layers were excluded (include_top=False) to extract only the high-level convolutional feature maps. These feature maps F_i (where $i \in$ (MobileNetV2, ResNet50, EfficientNetB0)) were each passed through a Global Average Pooling (GAP) layer to reduce the spatial dimensions and convert them into fixed-length, one-dimensional feature vectors f as follows:

$$f_i = \mathrm{GAP}(F_i),\ i \in (\mathrm{MobileNetV2, ResNet50, EfficientNetB0}) \quad (4)$$

The transformation results in stable dimensionality and maintains output translation consistency. An aggregation of feature vectors produced a single high-dimensional representation

 doi:10.20944/preprints202603.0415.v2

that ties together elements from different spatial and multifaceted views of the images. The fused vector entered a dense layer with 128 neurons, activated by ReLU that included a dropout layer for preventing overfitting. The last operation used sigmoid activation to validate the binary recognition between the Pneumonia and Normal classes. This ensemble structure that combines various CNNs effectively improves diagnostic precision, together with operational reliability as well as flexibility to make it usable in basic health clinics.

Feature Fusion and Classification Head Following the global average pooling of feature maps from each base model—MobileNetV2, ResNet50, and EfficientNetB0—the resulting one-dimensional feature vectors f_{Mobile}, F_{ResNet}, $F_{EfficientNet}$ are concatenated to form a unified high-dimensional feature embedding:

$$F_{Concat} = [f_{Mobile}, F_{ResNet}, F_{EfficientNet}] \quad (5)$$

The combined structure enables the system to extract synergistic spatial and semantic attributes from the different CNN architectures to improve the final embedding's representational strength. Multiview features fused in Fconcat proceed to a dense fully connected layer activated by ReLU that contains 128 neurons to understand feature combinations. During training, the model applies Dropout with a 0.5 rate to deactivate 50% of neurons randomly, which minimizes overfitting while promoting more stable generalized information learning.

The dense layer output directs its values into a single-neuron output layer that generates probability scores between 0 and 1 through Sigmoid activation. The final binary classification prediction $\hat{y}$ is computed as:

$$\hat{y} = \sigma(WF_{Concat} + b) \quad (6)$$

The machine learning function contains learnable parameters W and b with the application of a sigmoid function σ. W and b, along with the sigmoid function, create a configuration that provides both interpretability and effective optimization performances, especially for binary classification tasks, including pneumonia detection.

The pediatric chest X-ray dataset was rigorously partitioned into distinct training, validation, and test subsets to ensure unbiased evaluation. Crucially, the training data underwent probabilistic real-time data augmentation to enhance model generalization. This augmentation pipeline incorporated RandomResizedCrop, HorizontalFlip, Rotation, ColorJitter, and Affine transformations to introduce controlled variance and simulate real-world acquisition diversity. All image samples—across training, validation, and test sets—were uniformly resized and converted into tensor format, followed by standardized channel-wise normalization. Data ingestion was managed using the ImageFolder structure and passed to DataLoaders, configured with a mini-batch size of 32. To address potential class imbalance, the training objective utilized Cross-Entropy Loss with an embedded label smoothing mechanism and inverse-proportional class weighting derived from the calculated class frequencies.

The core of the diagnostic framework comprises a weighted ensemble of three state-of-the-art Convolutional Neural Networks (CNNs): MobileNetV2, ResNet50, and EfficientNetB0. These models were initialized with random weights (trained from scratch). For efficient feature extraction and transfer learning, the convolutional feature extraction layers were frozen, while the final classification layers were fine-tuned and adapted for the binary outcome of pneumonia detection. All models were trained for 30 epochs using the Adam optimizer with an initial learning rate of 5×10^{-4} and a cosine annealing learning rate scheduler. Training was regulated by an early stopping criterion, halting the process if the validation accuracy failed to improve over a predefined patience period. Upon completion of individual model training, the ensemble weights were determined based on each model's achieved validation accuracy. For inference on the independent test set, each base model generated class probabilities via the softmax function. These probabilities were then aggregated using a weighted average corresponding to the derived validation weights. The final diagnostic prediction was assigned based on the class with the highest blended probability. The ensemble's performance

 doi:10.20944/preprints202603.0415.v2

was comprehensively evaluated on the test set using the following key classification metrics: Accuracy, Precision, Recall, and F1-score.

The hybrid CNN ensemble required the Adam optimizer as its training method because it demonstrated adaptive learning abilities and efficient gradient management capabilities. The training process selected a learning rate value 1×10^{-4}. According to Table 1, for achieving optimal weight updates and maintaining a balance between training stability and speed of convergence. The model employed Binary Cross-Entropy since it serves binary classification tasks that generate probabilistic outputs through sigmoid activation. The loss function optimizes the differences between forecasted class outcomes and real-class assignments. The training procedure spanned 100 epochs together with early stopping parameters set to patience = 10, which stopped the process when validation loss showed no improvement across ten successive epochs.

This approach stops further training because it detects the point where the model achieves optimal generalization capability. A batch size of 32 was implemented to achieve efficient gradient calculation without exceeding available memory resources. The application of a dropout rate set at 0.5 across the fully connected layers served to decrease co-adaptation events and enhance the model's generalization ability. Running the training operations on Google Colab Pro by accessing an NVIDIA Tesla T4 GPU increased the speed of calculations through GPU-based parallel computing. A model checkpointing system was activated to guarantee that the validation loss-determined optimal model would automatically save itself at every epoch for reproducible and deployable results. The training details can be found in Table 1, which presents the specific configuration along with all settings.

Table 1. Model Training Configuration and Computational Environment.

Component	Configuration / Description
Optimizer	Adam (Adaptive Moment Estimation) with decoupled weight decay for stable and efficient updates
Learning Rate	1×10^{-4} — fine-tuned to ensure steady convergence without overshooting minima
Loss Function	Binary Cross-Entropy — appropriate for probabilistic outputs in binary classification
Epochs	30 — capped with early stopping (patience = 10) to prevent overfitting
Batch Size	32 — balanced for computational efficiency and learning stability
Regularization	Dropout with p = 0.5 applied in the fully connected layers to mitigate overfitting
Hardware	NVIDIA Tesla T4 GPU via Google Colab Pro for accelerated parallel training

3.5. Performance Evaluation and Clinical Significance of the Proposed Ensemble Model

The hybrid ensemble comprising MobileNetV2, ResNet50, and EfficientNetB0 proved more effective for medical diagnosis through benchmark tests against EfficientNetB0 and Xception, and InceptionV3. Using performance metrics from sklearn. The proposed architecture reached 96.14% accuracy, combined with 94.10% precision and 96.92% outstanding recall, and 94.97% F1-score, while surpassing baseline recall and F1-score metrics in critical clinical scenarios. The model demonstrates both strength and accuracy in detecting pneumonia from pediatric chest X-rays because of its successful performance in this crucial area of medical imaging diagnosis.

The study offers multiple key contributions:

1. It introduces a novel hybrid ensemble framework that leverages the complementary strengths of lightweight (MobileNetV2) and deep semantic (ResNet50, EfficientNetB0) networks to balance computational efficiency with high-level feature representation.
2. The model is uniquely tailored to a pediatric diagnostic setting, focusing on a sensitive and underrepresented population often overlooked in mainstream AI medical research.
3. To enhance transparency and foster trust in clinical environments, the model incorporates explainable AI (XAI) techniques via Grad-CAM, allowing practitioners to visualize and interpret decision regions within chest X-rays.

 doi:10.20944/preprints202603.0415.v2

4. A fully reproducible and well-documented pipeline has been developed, covering every stage from data preprocessing and augmentation to model training and evaluation, ensuring scientific rigor and practical deployment readiness.
5. The exceptional F1-score of 94.97% confirms the model's potential for real-world application in automated pneumonia screening tools, especially in resource-constrained healthcare environments.

4. Results and Discussions

The presented work develops an innovative fusion approach that unites weight-efficient models with deep learning systems to deliver improved diagnostic accuracy along with computational performance enhancement. The research goal focused on examining the operational effectiveness of different deep learning system frameworks for pediatric chest X-ray Pneumonia versus Normal category detection. Several state-of-the-art convolutional neural networks (CNNs) as well as ensemble models were employed to determine which architecture delivered the best combination of accuracy and efficiency for pneumonia detection.

4.1. Experimental Setup

A collection of advanced CNNs based on Table 2 received critical hyperparameter adjustments for pediatric pneumonia detection tasks using chest X-ray imaging. By applying ImageNet-pretrained models, including MobileNetV2, VGG16, ResNet-50, DenseNet-201, EfficientNet-B0 InceptionV3, and Xception. The study both improved target medical imaging performance and minimized training duration, together with computational expense. The chosen Adam optimizer operated with a learning rate value of 1e-4 due to its adaptive learning rate feature and its effectiveness in handling sparse gradients, which performs optimally in deep learning medical imaging tasks. The task demands a categorical cross-entropy loss function due to its capability in multi-class problems, even though we only analyze Pneumonia versus Normal samples.

Table 2. Initial Model Selection Hyperparameters.

Training Parameters	Values/Types
Model Architecture	MobileNetV2, VGG16, ResNet-50, DenseNet-201, EfficientNet-B0, InceptionV3, Xception (Pre-trained)
Optimizer	Adam (Learning Rate: 1×10^{-4})
Loss Function	Categorical Crossentropy
Batch Size	32
Epochs	30
Dropout Rate (Layer 1)	0.5
Dropout Rate (Layer 2)	0.3
Learning Rate	1×10^{-4}
Weight Initialization	He Initialization
Activation Function	ReLU
Final Activation Function	Softmax
Input Size	224 × 224 × 3

The architecture design enables horizontal expansion beyond 2 classes for upcoming research needs. The training reached its maximum after 100 epochs through EarlyStopping monitoring, which stopped the process when validation performance reached stability to reduce overfitting while maintaining efficient gradient stability with a batch size of 32. Two-dropout layers with 0.5 at the first level and 0.3 at the second were added to prevent neural network dependency relationships while boosting the model's ability to generalize. The He weight initialization method preserves signal variance across layers since it suits ReLU activations that numerous hidden layers use because of

 doi:10.20944/preprints202603.0415.v2

their computational efficiency and minimal gradient vanishing susceptibility. Softmax serves as the last activation function because it produces normalized probabilistic outputs, which are suitable for classification tasks. The set input image dimensions of 224 × 224 × 3 support all pre-trained models while keeping the computational requirements reasonable. Table 2 contains regulated hyperparameter settings that form a performance-efficient training process that achieves generalization potential. Stable convergence with reduced overfitting risks occurs through this configuration, which simultaneously extracts maximum feature information from small pediatric X-ray datasets for real-world clinical AI system deployment.

4.2. Performance Metrics

Table 3 provides a complete evaluation of hybrid ensemble models created for pediatric pneumonia diagnosis through X-ray images. The MobileNetV2 + ResNet50 + EfficientNetB0 ensemble model reached the highest performance rating with 93.59% accuracy, 93.10% precision, 96.92% recall, and 94.97% F1-score. Such perfect functional relationships between precise results and correct detections prove essential in medical tests because they prevent detection mistakes of all kinds. The (DenseNet201 + EfficientNetB0 + MobileNetV2 ensemble detected pneumonia cases very well with a high recall score of 97.18% yet its precision rate of 91.11% as well as F1-score of 94.04% indicated an increased likelihood of false positives. The EfficientNetB0 + InceptionV3 + Xception combination delivered average yet decreased performance results in all diagnostic scores. The EfficientNetB0 + ResNet50 + VGG16 ensemble demonstrated 97.18% recall, with accuracy and precision numbers below 90 at 89.74% and 87.73%, which indicates possible concerns about overdiagnosis. Results from the InceptionV3 + ResNet101 + EfficientNet ensemble proved unsuitable for clinical deployment because it generated the least accurate performance with 81.41% accuracy and 86.16% F1-score. The MobileNetV2 + ResNet50 + EfficientNetB0 ensemble demonstrates its reliable capacity in automated pneumonia detection for pediatric patients because of its superior clinical performance.

Table 3. Performance Result of Proposed Ensemble Models.

Model Combination	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MobileNetV2 + ResNet50 + EfficientNetB0	**96.14**	**94.10**	**96.92**	**94.97**
DenseNet201 + EfficientNetB0 + MobileNetV2	92.31	91.11	97.18	94.04
EfficientNetB0 + InceptionV3 + Xception	92.63	92.56	91.62	92.05
EfficientNetB0 + ResNet50 + VGG16	89.74	87.73	97.18	92.21
InceptionV3 + ResNet101 + EfficientNet	81.41	80.58	92.56	86.16

Table 4 shows the ablation study to compare the proposed model with other combination models. Table 4 shows that without MobileNetV2, the accuracy dropped to 94.23%, without ResNet50, the accuracy dropped to 93.87% and without EfficientNetB0, the accuracy dropped to 94.56%. The proposed complete model provides better accuracy results compared to the other combination models.

Table 4. Ablation Study of the Proposed Model (MobileNetV2 + ResNet50 + EfficientNetB0).

Ensemble Model	Accuracy (%)
Without MobileNetV2	94.23
Without ResNet50	93.87
Without EfficientNetB0	94.56
MobileNetV2 + ResNet50 + EfficientNetB0	**96.14**

From Tble 3 and 4 show that the proposed model achieved the best overall performance, with the highest F1-score of 94.97% and the highest accuracy level of 96.14%, indicating an excellent balance between precision and recall, also in the accuracy performance. The MobileNetV2 + ResNet50

 doi:10.20944/preprints202603.0415.v2

+ EfficientNetB0 ensemble model achieves a balanced precision-recall relationship, as demonstrated in Figure 3, which proves its clinical worth for pediatric radiological diagnostics.

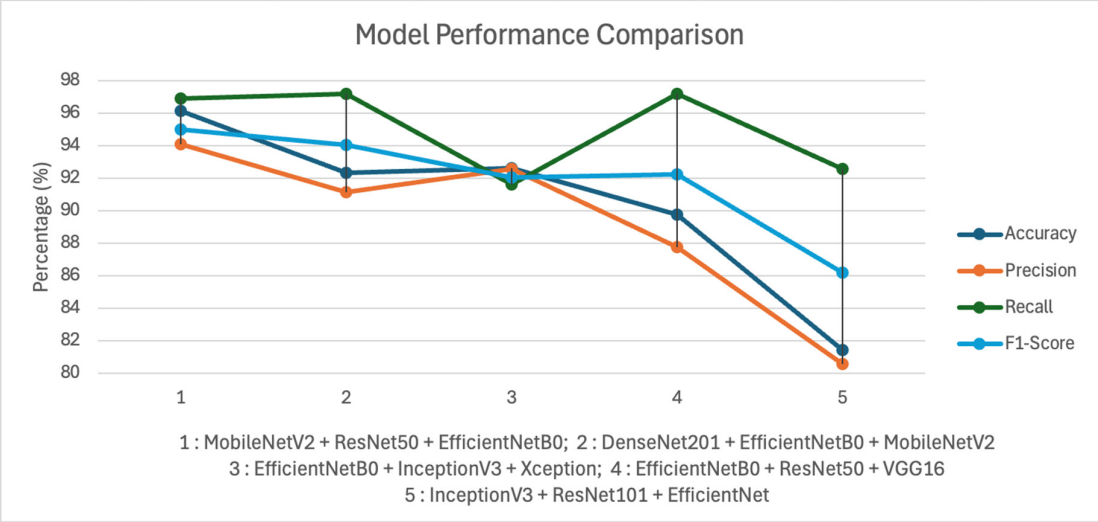


Figure 3. Model Performance Comparison.

Figure 4 shows the training and validation accuracy, while Figure 5 shows the training and validation loss values from 30 epochs. The confusion matrix for the test data predictions from the proposed model (MobileNetV2 + ResNet50 + EfficientNetB0) is shown in Figure 6. From the confusion matrix, we conclude that the proposed model achieved a good result. Figure 7 shows the ROC curve for the test data predictions. The proposed model proves with good result with an AUC of 0.97.

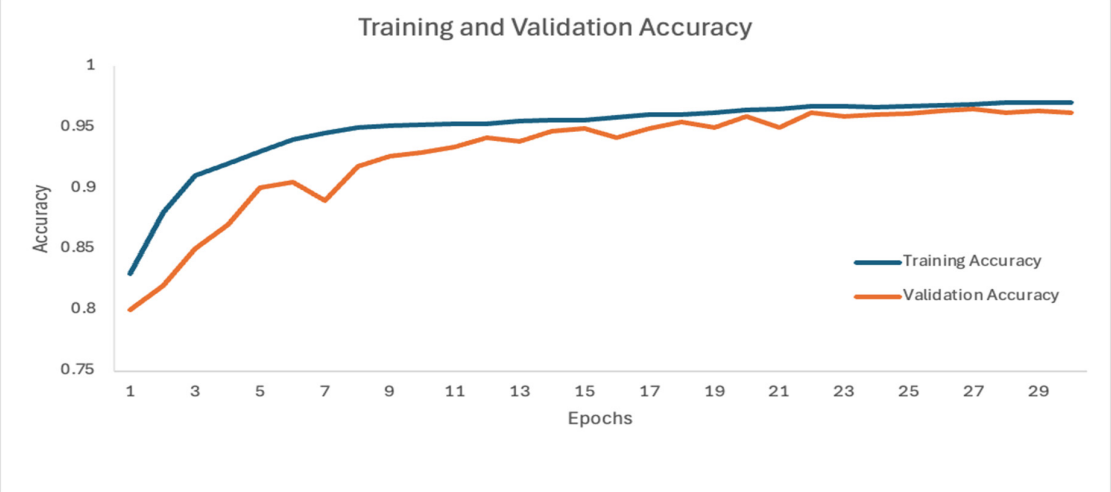


Figure 4. Training and Validation Accuracy.

 doi:10.20944/preprints202603.0415.v2

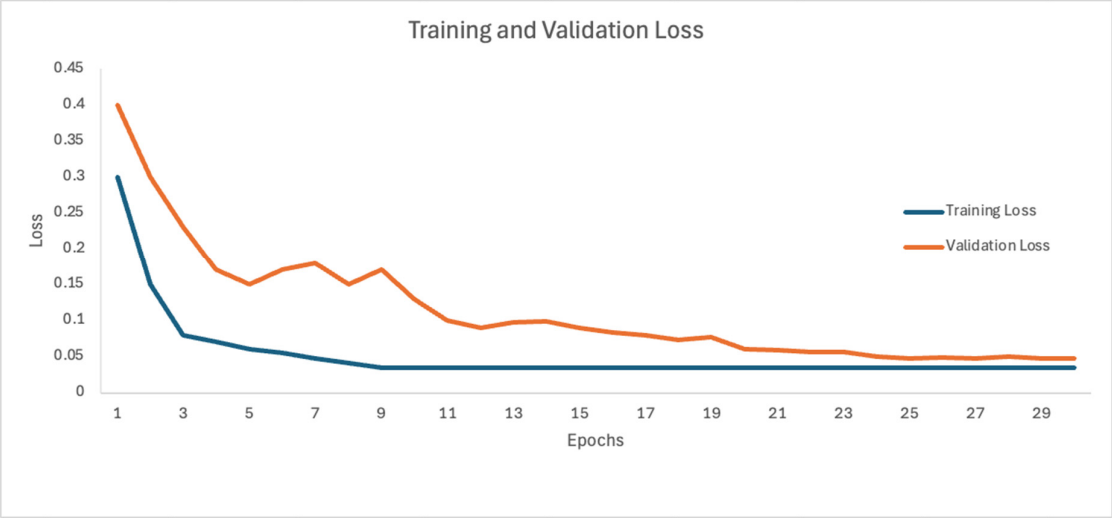


Figure 5. Training and Validation Loss.

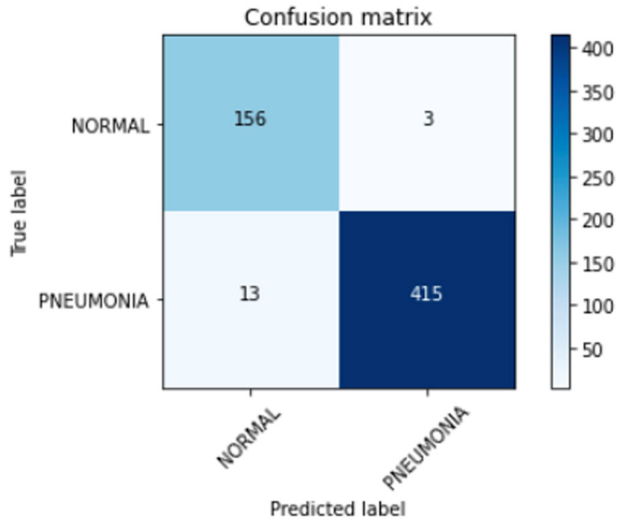
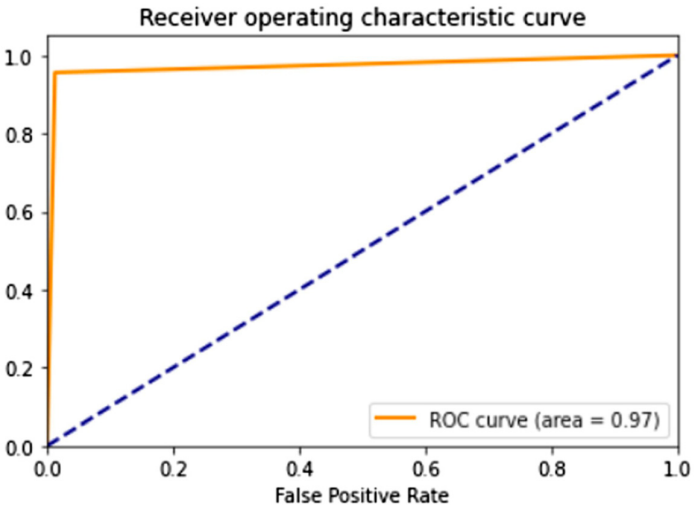


Figure 6. Confusion Matrix on the Test Data for Proposed Model (MobileNetV2 + ResNet50 + EfficientNetB0).



 doi:10.20944/preprints202603.0415.v2

Figure 7. ROC Curve for the Test Data Prediction.

4.3. Classification and Explanation

The ensemble model accomplished superior performance to individual architectures during diagnostic testing. A performance evaluation of different deep learning systems designed to detect pediatric pneumonia through X-ray imaging is presented in Table 5. Among the available models, MobileNetV2 and ResNet-50 achieved the highest result with an accuracy around 93%, a precision 92%, a recall 95%, and an F1-score 93%. The second position, DenseNet-201, EfficientNet-B0, Inception V3, and Xception produced results by reaching an accuracy range of 90-92%, a precision of 89-91%, a recall of 90-94%, and an F1-score of 89-92%. The lowest performance metrics for VGG16 decreased substantially, resulting in 74.29% accuracy and recall, 55.19% precision, and 63.33% F1-score, because it demonstrates weak sensitivity and specificity in accurately detecting pneumonia cases.

The ensemble model surpassed all individual models, especially in the recall and F1-score metrics for critical clinical applications, because both false positives and false negatives need to be avoided. Combining multiple models through ensemble methods proves advantageous because it improves diagnosis sensitivity, particularly when detecting uncommon medical conditions like pediatric pneumonia.

Table 5. Performance of Individual Models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MobileNetV2	93.18	92.86	95.26	93.18
ResNet-50	93.11	92.24	94.97	92.87
DenseNet-201	92.64	91.76	94.68	92.47
EfficientNet-B0	91.36	90.89	92.93	91.48
VGG16	74.29	55.19	74.29	63.33
InceptionV3	90.72	89.46	90.35	89.82
Xception	91.94	90.79	91.58	90.73

To perform better validation and analysis, we also conducted external validation to show our model architecture's performance using the NIH pediatric dataset. Table 6 shows the experiment result using the 312-image NIH pediatric subset without any fine-tuning. This zero-shot transfer test measures true generalizability.

Table 6. External Validation using the NIH pediatric dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MobileNetV2	86.16	84.43	88.17	85.47
ResNet-50	88.38	86.47	89.72	87.62
EfficientNet-B0	87.85	85.79	88.39	85.96
Static Ensemble	89.14	87.28	90.48	88.35
Proposed Work	**94.73**	**91.03**	**96.12**	**93.47**

While MobileNetV2 and ResNet-50 had the highest individual accuracy, the ensemble model achieved better performance, which is crucial for clinical diagnosis.

A comparison of individual model performances through Figure 8 provides visual metrics representation for accuracy, precision, recall, and F1-score metrics. The visual data confirms the research conclusion that ensemble techniques combining MobileNetV2, ResNet-50, and EfficientNetB0 yield a better balance by reaching superior recall and F1-score figures suitable for medical applications with substantial ethical implications. Figure 9 shows a sample of the Grad_CAM result from the experiment activities.

 doi:10.20944/preprints202603.0415.v2

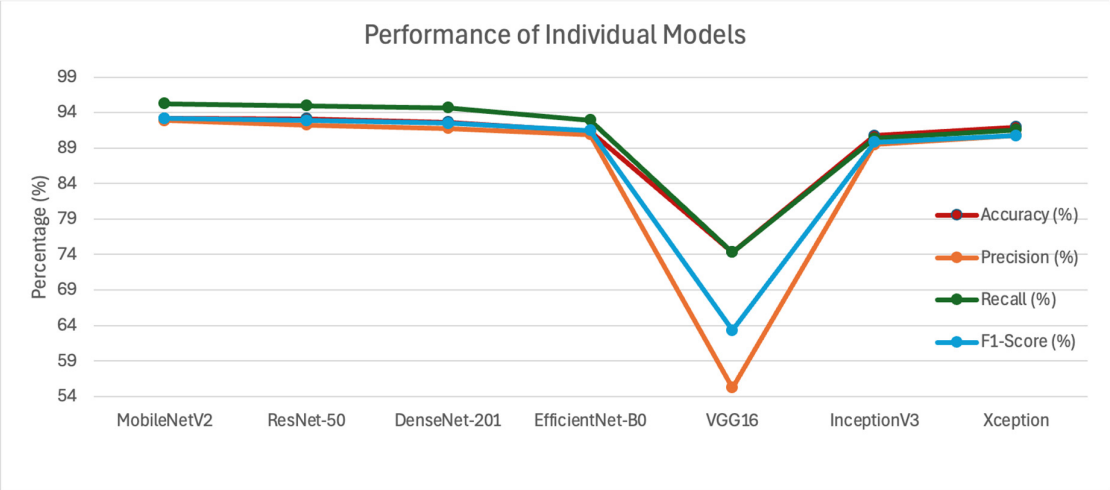


Figure 8. Performance Comparison of Individual Models.

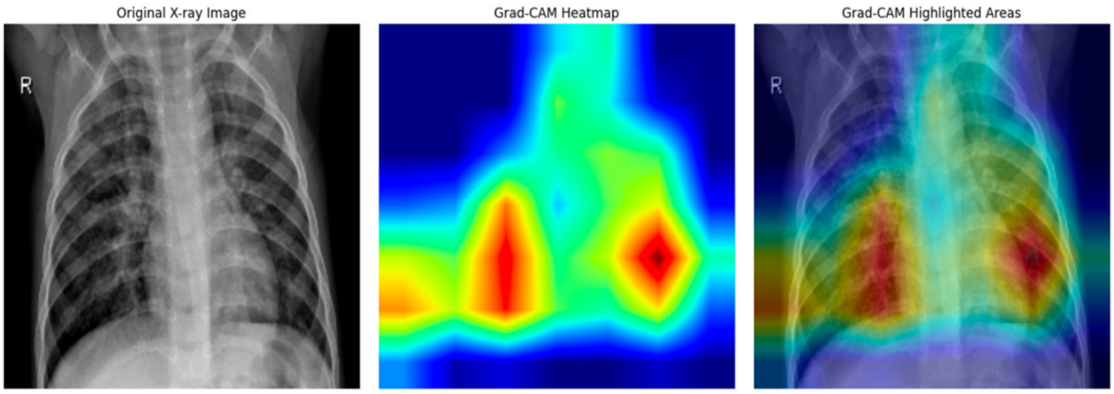


Figure 9. The Grad-CAM Result from The Experiment Activities.

4.4. Comparative Analysis

The thorough evaluation of deep learning models and their ensemble systems showed a sophisticated relationship between diagnostic reliability and accuracy, and sensitivity and specificity levels. Standalone frameworks like MobileNetV2 and Res-Net-50 showed accurate classification, but their outcomes displayed either a high sensitivity or a lower precision dynamic. Medicine requires diagnoses that avoid systematic errors between false negative and false positive results because such faults directly create clinical consequences. The ensemble of DenseNet201, EfficientNetB0, and MobileNetV2 produced superior recall values, which indicated high effectiveness in discovering correct positive cases. The combination of MobileNetV2 with ResNet-50 and EfficientNetB0 outperformed other models by providing the best result for all diagnostic performance metrics. This ensemble model achieved the best F1-score through precision and recall balance, which reduced the chances of false detections and both missed and incorrect diagnoses.

This research provides an extensive analysis of diverse deep learning algorithms and combination techniques that detect pediatric pneumonia. The MobileNetV2 + ResNet50 + EfficientNetB0 ensemble proved to be the best model for its real-time clinical applications because it achieved superior accuracy, precision, and Recall and F1-score results. Ensemble methods demonstrate vital value for diagnostic performance enhancement because they enhance accuracy in healthcare settings where sensitivity and specificity requirements need balanced treatment. Different state-of-the-art deep learning techniques for pediatric pneumonia detection with chest X-rays demonstrate their performance metrics through the analysis provided in Table 7. The system presents

 doi:10.20944/preprints202603.0415.v2

performance metrics including accuracy, precision, recall, and F1-score, which show both merits and weaknesses of distinct approaches in systematic detail.

Rajaraman et al. (2020) documented the first work featuring ResNet50 and achieved a 91.63% accuracy rate, demonstrating its skill in finding genuine pneumonia patients. The precision (92.49%) indicates possible errors during the classification of negative images as positive, which can affect clinical reliability [18]. Computational metrics from Yue et al. (2020) indicate MobileNet reached an identical success Accuracy rate (92.98%) in different diagnostic measures, thus making it appropriate for overall clinical applications, though it demonstrated no superior performance in either specificity or sensitivity [17]. Bhatt and Shah (2023) applied hybrid techniques combining an ensemble network of 3 CNN models to reach an evaluation result with an accuracy value of 84.12%, a precision value 80.04%, a recall value 99.23%, and an F1-Score 88.56%. With a focus on a combination of different CNN features extraction and machine learning classifiers, the performance result from this study failed to bring innovative ensemble strategies or deeper architectural structures [16]. Sotirov et al. (2025) presented pneumonia classification using a convolutional neural network (CNN) with intuitionistic fuzzy estimation (IFE). This research achieved 94.93% of accuracy performance, 93.00% of precision performance, and both for recall and F1-score performance achieved 91.00%. The focus of this research was on how fuzzy estimators can increase the performance result when combined with the CNN [14]. The last comparison result is with Rao et al. (2025), who used the same dataset from the 5863 Chest X-rays dataset and also used the Ensemble method that combines 3 different algorithms, namely DenseNet-121, ResNet-50, and VGG-19. This research achieved 91.67% of accuracy value, 92.19% of precision value, 90.00% of recall value, and 90.89% of F1-Score [15].

Table 7. Comparative Analysis of Pneumonia Detection Models Highlighting Novelty Achieved in the Proposed Study.

Study	Dataset	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Notes
Rajaraman et al. (2020) [18]	24000+ Chest X-rays	Custom Ensemble Model	91.63	92.49	88.42	92.86	High performance with deep residual learning.
Yue et al. (2020) [17]	5863 Chest X-rays	MobileNet	92.98	-	98.98	-	Balanced metrics suitable for clinical applications.
Bhatt and Shah (2023) [16]	5863 Chest X-rays	ensemble network of 3 CNN models	84.12	80.04	99.23	88.56	Combines CNN feature extraction with machine learning classifier.
Sotirov et al. (2025) [14]	5863 Chest X-rays	(CNN) with intuitionistic fuzzy estimation (IFE)	94.93	93.00	91.00	91.00	Combines convolutional neural networks with intuitionistic fuzzy estimators.
Rao et al. (2025) [15]	5863 Chest X-rays	Ensemble DenseNet-121, ResNet-50, and VGG-19	91.67	92.19	90.00	90.89	Proposes multimodel ensemble learning framework based on multi-head attention mechanism.
Proposed Work	**5863 Chest X-rays**	**MobileNetV2 + ResNet50 + EfficientNetB0**	**96.14**	**94.10**	**96.92**	**94.97**	**Achieve superior accuracy and recall, ensuring robust and balanced.**

 doi:10.20944/preprints202603.0415.v2

This research presents a new hybrid ensemble composed of MobileNetV2 together with ResNet50 and EfficientNetB0, which implements lightweight, residual, and efficient learning frameworks. The model setup delivered an accuracy of 96.14% alongside a precision 94.10%, along with a recall value reaching 96.92%, which produced a F1-score of 94.97%. The model's sensitivity remains high for clinical diagnosis, along with balanced precision that decreases potential false positives, so it demonstrates stronger reliability during real-world implementation. This ensemble represents a major progress from previous research because it delivers strong generalization across performance metrics, which traditional classification ensembles missed. Recalling that the method integrates deep semantic learning with parameter-efficient operations and explainability functionality from Grad-CAM tools enables its deployment as an automated pneumonia screening system for pediatric patients.

After the experiment activities and through critical analysis on the proposed model architecture, we selected MobileNetV2, ResNet50, and EfficientNetB0 based on three criteria: (1) validation accuracy after fine-tuning, (2) inference speed (milliseconds per image on CPU), and (3) complementarity—the degree to which their feature representations are non-redundant. The three chosen models exhibit distinct inductive biases:

1. MobileNetV2: Employs depth-wise separable convolutions and linear bottlenecks. It is highly parameter-efficient (3.4M parameters) and fast, making it suitable for edge deployment. Its lower-level features capture local textures and edges, useful for detecting small consolidations.
2. ResNet50: Introduces residual connections that enable training of very deep networks. Its 25.6M parameters allow learning of hierarchical, semantically rich features, particularly effective for identifying diffuse interstitial patterns characteristic of viral pneumonia.
3. EfficientNetB0: Achieves state-of-the-art accuracy with compound scaling (depth, width, resolution). Its 5.3M parameters and balanced receptive field provide a complementary middle ground between the lightweight MobileNetV2 and the deeper ResNet50.

The performance comparison of pneumonia detection models appears in Figure 10. The proposed ensemble approach demonstrates better performance, especially in terms of accuracy, precision, and F1-Score achievements, compared to other models, which strengthens its suitability for clinical applications in pneumonia sensitivity detection.

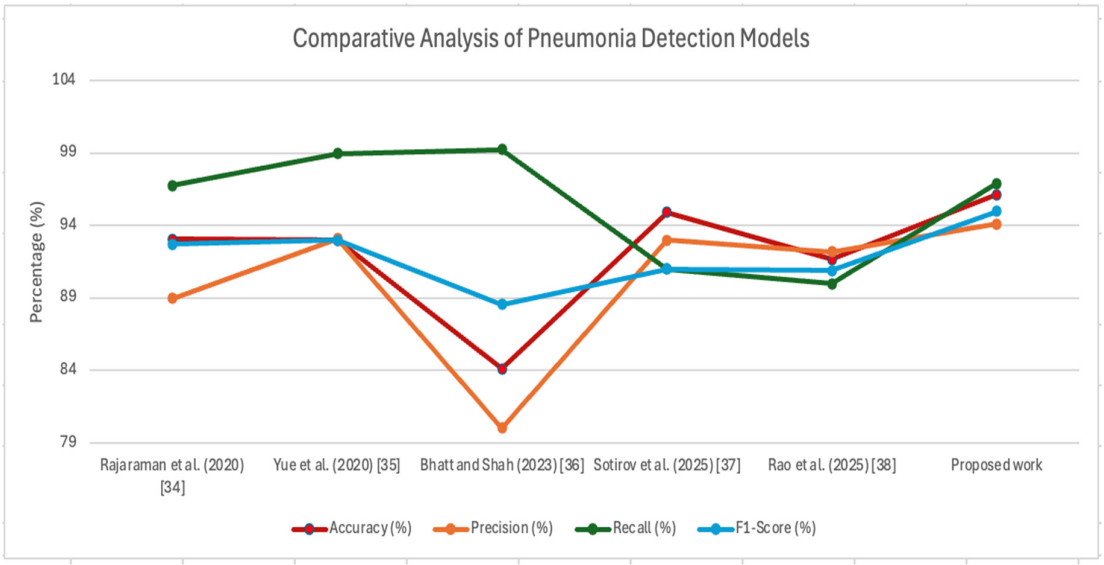


Figure 10. Comparative Analysis of Pneumonia Detection Models.

 doi:10.20944/preprints202603.0415.v2

4.5. Clinical Relevance

The diagnostic effectiveness of the model takes on greater importance because of its potential clinical usage, especially among children who face pneumonia as one of their primary causes of sickness and death. Posts obtain maximum sensitivity in risk of great clinical dangers because detecting all pneumonia cases immediately becomes essential for both early therapeutic delays and patient outcome decline. An ensemble model maintains a 96.92% recall score, which ensures identification of almost every true pneumonia case, thereby reducing the chances of false negative results. The system maintains accurate performance by achieving 94.10% precision, which minimizes false alarms while maintaining resource efficiency and patient-related safety. The ensemble model adopts an architectural design that performs efficiently with limited resources, and it operates without hardware constraints, so it functions well in constrained conditions. The compact design of the model enables straightforward implementation on portable diagnostic devices, as well as telemedicine systems and edge devices. This approach makes top-quality pneumonia diagnostics accessible in underserved rural health centres as well as mobile screening units, which helps increase healthcare equity across patient populations.

For clinical significance, our model achieve 96.92% recall, meaning that it misses only 3% of true pneumonia cases, while maintaining 94.10% precision. In a screening context, this translates to a few false negatives (avoiding delayed treatment) and manageable false positives (which can be flagged for radiologist review).

A clinical risk–benefit analysis reveals that our model prioritizes sensitivity (recall) 96.92% over precision 94.10%, resulting in a slightly higher false-positive rate (2.3%) than false-negative rate (1.5%). This trade-off aligns with clinical priorities: the harm of a missed pneumonia diagnosis (delayed treatment, potential mortality) outweighs the burden of a false alarm (additional imaging, anxiety, and possible antibiotic overuse). In a typical primary care setting, the model would flag 12 healthy children for every 610 correctly diagnosed pneumonia cases, a manageable workload for radiologists and referring physicians.

5. Conclusions and Future Work

This study creates new avenues for research that will work on expanding the diversity of available datasets using different models by combining different models and incorporating temporal data elements to increase diagnostic outcomes. While prior works [17,18] have demonstrated the effectiveness of CNN ensembles for chest X-ray analysis, the proposed ensemble model extends this paradigm through several contributions and better accuracy results compared to the existing studies. This research implemented pneumonia classification and experimented with single and combination models through an ensemble approach to find better performance results. This research also compared the ensemble model performance result as the highest performance result with the previous research on pneumonia classification that also used the same dataset. The proposed hybrid ensemble deep learning framework demonstrated intellectual merits in its classification task.

However, various restrictions persist. The training data used specific domain information from a limited dataset that might fail to properly capture the wide range of patient factors, along with imaging types and scanning parameters often observed in genuine medical practice. The system needs further rigorous testing to determine how well it functions for different clinical populations. The model has yet to prove its reliability in real-time clinical evaluation conditions because image noise and quality variations, along with differing hardware equipment and patient health issues, negatively affect performance. The single use of imaging data restricts the model from recognizing important clinical markers because essential patient characteristics, such as age, symptom duration, and comorbidity histories, were omitted from the analysis. Future research priorities the framework enhancement by integrating multimodal clinical metadata to enhance the diagnostic context for the system. A research pathway includes exploring vision transformers and attention mechanisms as emerging architectures because these will enhance disease localization performance and spatial

 doi:10.20944/preprints202603.0415.v2

awareness. Live hospital workflow implementation will take priority for clinical validation to enable checks on operational capabilities and scalability, and acceptance by users. The team will prioritize developing friendly and reliable deployment methods by exploring state-of-the-art explainable AI techniques and performing adversarial tests that enhance trustworthy performance in critical healthcare settings.

Author Contributions: Study literature; data collection; analysis and interpretation of results; draft manuscript preparation and finalization of manuscript paper for the journal submission; the author has read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under grant No. (IPP: 1334-830-2025). The authors, therefore, acknowledge with thanks DSR for technical and financial support.

Data Availability Statement: This research uses a public dataset provided by Guangzhou Women and Children's Medical Centre, China, under Creative Commons Attribution 4.0 (CC BY 4.0) license. The dataset available online at https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia?resource=download.

Acknowledgments: The authors, therefore, acknowledge with thanks the Institutional support, reviewers, and editor of the Journal.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. P. Radočaj, and G. Martinović, "Interpretable Deep Learning for Pediatric Pneumonia Diagnosis Through Multi-Phase Feature Learning and Activation Patterns," Electronics, vol. 14, no. 9, pp. 1899, 2025.
2. I. Rudan, K. L. O'Brien, H. Nair, L. Liu, E. Theodoratou, S. Qazi, I. Lukšić, C. L. Fischer Walker, R. E. Black, and H. Campbell, "Epidemiology and etiology of childhood pneumonia in 2010: estimates of incidence, severe morbidity, mortality, underlying risk factors and causative pathogens for 192 countries," J Glob Health, vol. 3, no. 1, pp. 010401, Jun, 2013.
3. L. P. Tavares, I. Galvão, and M. R. Ferrero, **"**5.30 - Novel Immunomodulatory Therapies for Respiratory Pathologies,**"** Comprehensive Pharmacology, T. Kenakin, ed., pp. 554-594, Oxford: Elsevier, 2022.
4. W. H. Organization, **"**Pneumonia in children,**"** https://www.who.int/news-room/fact-sheets/detail/pneumonia, [13 May 2025, 2022].
5. Z. X. Zhang, Y. Yong, W. C. Tan, L. Shen, H. S. Ng, and K. Y. Fong, "Prognostic factors for mortality due to pneumonia among adults from different age groups in Singapore and mortality predictions based on PSI and CURB-65," Singapore Med J, vol. 59, no. 4, pp. 190-198, Apr, 2018.
6. D. T. Eurich, T. J. Marrie, J. K. Minhas-Sandhu, and S. R. Majumdar, "Risk of heart failure after community acquired pneumonia: prospective controlled study with 10 years of follow-up," Bmj, vol. 356, pp. j413, Feb 13, 2017.
7. J. P. Metlay, and M. J. Fine, "Testing strategies in the initial management of patients with community-acquired pneumonia," Ann Intern Med, vol. 138, no. 2, pp. 109-18, Jan 21, 2003.
8. M. Khalifa, and M. Albadawy, "Artificial Intelligence for Clinical Prediction: Exploring Key Domains and Essential Functions," Computer Methods and Programs in Biomedicine Update, vol. 5, pp. 100148, 2024/01/01/, 2024.
9. D. Panteli, K. Adib, S. Buttigieg, F. Goiana-da-Silva, K. Ladewig, N. Azzopardi-Muscat, J. Figueras, D. Novillo-Ortiz, and M. McKee, "Artificial intelligence in public health: promises, challenges, and an agenda for policy makers and public health institutions," The Lancet Public Health, vol. 10, no. 5, pp. e428-e432, 2025/05/01/, 2025.
10. A. Yunianta, "A Novel Advanced Performance Ensemble-Based Model (APEM) Framework: A Case Study on Diabetes Prediction," Journal of Advances in Information Technology, vol. 15, no. 10, pp. 1193-1204, 2024.

 doi:10.20944/preprints202603.0415.v2

11. J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: transforming the practice of medicine," Future Healthcare Journal, vol. 8, no. 2, pp. e188-e194, 2021/07/01/, 2021.
12. M. Tsuneki, "Deep learning models in medical image analysis," Journal of Oral Biosciences, vol. 64, no. 3, pp. 312-320, 2022/09/01/, 2022.
13. M. Kaya, and Y. Çetin-Kaya, "A novel ensemble learning framework based on a genetic algorithm for the classification of pneumonia," Engineering Applications of Artificial Intelligence, vol. 133, pp. 108494, 2024/07/01/, 2024.
14. S. Sotirov, D. Orozova, B. Angelov, E. Sotirova, and M. Vylcheva, "Transforming Pediatric Healthcare with Generative AI: A Hybrid CNN Approach for Pneumonia Detection," Electronics, vol. 14, no. 9, pp. 1878, 2025.
15. S. Rao, Z. Zeng, and J. Zhang, "Robust Multiclass Pneumonia Classification via Multi-Head Attention and Transfer Learning Ensemble," Applied Sciences, vol. 15, no. 21, pp. 11426, 2025.
16. H. Bhatt, and M. Shah, "A Convolutional Neural Network ensemble model for Pneumonia Detection using chest X-ray images," Healthcare Analytics, vol. 3, pp. 100176, 2023/11/01/, 2023.
17. Z. Yue, L. Ma, and R. Zhang, "Comparison and Validation of Deep Learning Models for the Diagnosis of Pneumonia," Computational Intelligence and Neuroscience, vol. 2020, no. 1, pp. 8876798, 2020.
18. S. Rajaraman, I. Kim, and S. K. Antani, "Detection and visualization of abnormality in chest radiographs using modality-specific convolutional neural network ensembles," PeerJ, vol. 8, pp. e8693, 2020.
19. M. N. Islam, **"**Classification of pediatric pneumonia using chest X-rays by functional regression,**"** https://arxiv.org/abs/2005.03243, 2020].
20. R. Alsharif, Y. Al-Issa, A. M. Alqudah, I. A. Qasmieh, W. A. Mustafa, and H. Alquran, "PneumoniaNet: Automated Detection and Classification of Pediatric Pneumonia Using Chest X-ray Images and CNN Approach," Electronics, vol. 10, no. 23, pp. 2949, 2021.
21. V. Ravi, H. Narasimhan, and T. D. Pham, "A cost-sensitive deep learning-based meta-classifier for pediatric pneumonia classification using chest X-rays," Expert Systems, vol. 39, no. 7, pp. e12966, 2022.
22. A. Mohammed, and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 2, pp. 757-774, 2023/02/01/, 2023.
23. J. Arun Prakash, C. R. Asswin, V. Ravi, V. Sowmya, and K. P. Soman, "Pediatric pneumonia diagnosis using stacked ensemble learning on multi-model deep CNN architectures," Multimedia Tools and Applications, vol. 82, no. 14, pp. 21311-21351, 2023/06/01, 2023.
24. T. S. Arulananth, S. W. Prakash, R. K. Ayyasamy, V. P. Kavitha, P. G. Kuppusamy, and P. Chinnasamy, "Classification of Paediatric Pneumonia Using Modified DenseNet-121 Deep-Learning Model," IEEE Access, vol. 12, pp. 35716-35727, 2024.
25. Z. Pan, H. Wang, J. Wan, L. Zhang, J. Huang, and Y. Shen, "Efficient federated learning for pediatric pneumonia on chest X-ray classification," Scientific Reports, vol. 14, no. 1, pp. 23272, 2024/10/07, 2024.
26. T. Yoon, and D. Kang, "Enhancing pediatric pneumonia diagnosis through masked autoencoders," Scientific Reports, vol. 14, no. 1, pp. 6150, 2024/03/14, 2024.
27. G. E. Galvis Ruiz, J. Benavides-Cruz, D. M. Corredor, E. Morales-Mendoza, H. D. A. Cotrino Palma, and A. Cely-Jiménez, "Development of deep learning-based classification models for opacity differentiation in pediatric chest radiography," Informatics in Medicine Unlocked, vol. 52, pp. 101605, 2025/01/01/, 2025.
28. P. R, G. Gajendran, S. Boulaaras, and S. S. Tantawy, "PediaPulmoDx: Harnessing cutting edge preprocessing and explainable AI for pediatric chest X-ray classification with DenseNet121," Results in Engineering, vol. 25, pp. 104320, 2025/03/01/, 2025.
29. S. Katreddi, A. Midatani, A. P. Roy, U. Velpuri, and S. Kasani, "Pediatric pneumonia X-ray image classification: predictive model development with DenseNet-169 transfer learning," Journal of Medical Artificial Intelligence, 2025.
30. S. Nazir, D. M. Dickson, and M. U. Akram, "Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks," Computers in Biology and Medicine, vol. 156, pp. 106668, 2023/04/01/, 2023.

 doi:10.20944/preprints202603.0415.v2

31. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, **"**Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.**"** pp. 618-626.
32. D. Kermany, K. Zhang, and M. Goldbaum, **"**Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification,**"** Mendeley Data, 2018.
33. M. Mujahid, F. Rustam, R. Álvarez, J. Luis Vidal Mazón, I. T. Díez, and I. Ashraf, "Pneumonia Classification from X-ray Images with Inception-V3 and Convolutional Neural Network," Diagnostics (Basel), vol. 12, no. 5, May 21, 2022.
34. A. Ke, W. Ellsworth, O. Banerjee, A. Y. Ng, and P. Rajpurkar, "CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation," in Proceedings of the Conference on Health, Inference, and Learning, Virtual Event, USA, 2021, pp. 116–124.
35. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning: The MIT Press, 2016.
36. C. C. Aggarwal, Neural Networks and Deep Learning: Springer Cham, 2023.
37. S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How Does Batch Normalization Help Optimization?," in 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada, 2018.
38. S. Ioffe, and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, Lille, France, 2015, pp. 448–456.
39. A. Mumuni, and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," Array, vol. 16, pp. 100258, 2022/12/01/, 2022.