

# Detection of Pneumonia from Chest X-Ray Images using Deep Learning and Ensemble Approaches with Grad-CAM Explainability

Feyza Toktaş

Department of Computer Engineering  
Ankara University  
Ankara, Türkiye  
[ftoktas@ankara.edu.tr](mailto:ftoktas@ankara.edu.tr)

Melih Osman Gayretli

Department of Computer Engineering  
Ankara University  
Ankara, Türkiye  
[gayretli.melih@hotmail.com](mailto:gayretli.melih@hotmail.com)

Kaan Alp Özdemir

Department of Computer Engineering  
Ankara University  
Ankara, Türkiye  
[okaanalp@gmail.com](mailto:okaanalp@gmail.com)

Makbule Yılmaz

Department of Computer Engineering  
Ankara University  
Ankara, Türkiye  
[makbuleyilmaz5103@gmail.com](mailto:makbuleyilmaz5103@gmail.com)

**Abstract**— This study introduces RDE-PneuNet(ResNet, DenseNet, EfficientNet Pneumonia Network), a robust, balanced, and explainable deep learning framework for detecting pneumonia from chest X-ray images. The proposed architecture integrates three structurally diverse backbone models—ResNet50, DenseNet121, and EfficientNetB0—utilizing an optimized weighted soft-voting ensemble strategy to maximize clinical reliability. To mitigate inherent class imbalances and prevent overfitting on the Chest X-ray dataset, transfer learning, fine-tuning, data augmentation, and strategic class weighting were implemented. Experimental results demonstrate that the ensemble configuration outperforms standalone models, achieving a highly competitive 93% accuracy alongside balanced precision, recall, and F1-score metrics. Crucially, the system overcomes the common clinical trade-off of prioritizing sensitivity at the severe expense of specificity, yielding a well-balanced 92% specificity and an AUC of 0.975. To ensure qualitative transparency and overcome "black-box" limitations, Grad-CAM visual explanations were integrated, confirming that the network consistently attends to anatomically meaningful pulmonary regions rather than background artifacts. Ultimately, RDE-PneuNet provides a reproducible, computationally efficient, and highly transparent decision-support tool suitable for real-world clinical workflows.

**Keywords**— *Pneumonia detection, deep learning, ensemble learning, ResNet50, DenseNet121, EfficientNet, Grad-CAM, RDE-PneuNet, Computer Vision*

## I. INTRODUCTION

Pneumonia is a common respiratory disease that causes infection in the lung tissue and poses a serious risk, especially for children, elderly individuals, and patients with weak immune systems. This disease, which has high morbidity and mortality rates worldwide, can lead to fatal consequences if not diagnosed early. Clinically, pneumonia is divided into different subtypes, primarily bacterial and viral, and accurate diagnosis plays a critical role in determining the treatment process. Chest X-ray images are one of the most used methods for diagnosing pneumonia and hold an important place in clinical applications due to their low cost and rapid result generation. However, the accurate interpretation of these images largely depends on expert radiologists, making this process vulnerable to human error. Therefore, the

development of computer-aided systems is of great importance so that pneumonia can be detected automatically, quickly, and reliably. In recent years, deep learning, particularly Convolutional Neural Network (CNN) based approaches, has achieved significant success in medical image analysis [13]. Models developed using transfer learning methods, with pre-trained architectures such as VGG, ResNet, and DenseNet, have been reported to achieve high accuracy rates. However, training CNN-based models with different optimization strategies directly affects model performance, enabling more balanced and stable results [2, 5]. Furthermore, comparative studies have shown that CNN architecture offers an effective foundational structure when used alongside classical machine learning methods [3, 4] or when evaluating different CNN-based approaches [1, 6, 7].

On the other hand, merely achieving high accuracy in the medical imaging domain is insufficient; the reliability of the model and the explainability of its decision mechanism have become critical requirements. In this context, Explainable AI (XAI) methods like Grad-CAM (Gradient-weighted Class Activation Mapping) increase clinical reliability by visualizing the regions the model relies on to make decisions, thereby contributing to the development of systems that experts can easily interpret [12, 17].

The field of pneumonia detection using deep learning is rich with diverse methodologies. Tripathy et al. proposed a Transfer Learning (TL) framework utilizing deep learning models such as Xception, VGG16, and ResNet152V2 for pneumonia classification, concluding that the ResNet152V2 model [1]. They evaluated their methodology using a highly imbalanced dataset of 5,856 chest X-ray images of pediatric patients obtained from the Kaggle repository. The researchers concluded that the ResNet152V2 model performed exceptionally well, achieving an accuracy of 83.17%, and showcased its potential to assist physicians with rapid and accurate diagnoses.

Manickam et al. utilized deep transfer learning models, including ResNet50, InceptionV3, and InceptionResNetV2, alongside a U-Net architecture for lung segmentation [2]. The models were trained and evaluated on a publicly available database comprising 5,232 chest X-ray images, analyzing the

effects of different optimizers like Adam and SGD. They concluded that the ResNet50 model using the Stochastic Gradient Descent (SGD) optimizer yielded a superior accuracy of 93.06%, significantly simplifying and improving the automated pneumonia detection process.

Trindade et al. compared the performance of multiple Machine Learning classifiers, primarily employing a Convolutional Neural Network (DenseNet121 via transfer learning), Random Forest, and Support Vector Machine models combined with data augmentation [3]. Their analysis utilized the Chest X-Ray Images (Pneumonia) dataset from Kaggle, containing 5,856 X-ray images categorized into normal, viral, and bacterial classes. The study concluded that the CNN-based DenseNet121 model attained the best overall performance with an 81% accuracy and an 84.2% Area Under the Curve (AUC), confirming that accessible AI techniques can effectively support clinical imaging diagnosis.

Rai et al. conducted a comprehensive review surveying various machine learning algorithms, comparing non-complex CNNs, transfer learning architectures, Support Vector Machines (SVM), and XGBoost. Their survey evaluated these methodologies across multiple benchmark datasets found in the literature, including the RSNA dataset and Kaggle collections of pediatric chest X-rays [4]. The authors concluded that using pre-trained models like GoogLeNet or DenseNet201, as well as combining CNN-extracted features with SVM classification, produces highly precise and optimal results for pneumonia recognition.

Chowdhury et al. investigated the impact of different optimizers (RMSProp, Adam, and SGD) on custom CNN models formulated with varying numbers of hidden convolution layers (from one to five) without relying on transfer learning [5]. They tested these architectural combinations using a Kaggle dataset comprising 5,856 chest X-ray images split into training, validation, and testing subsets. The authors concluded that the CNN architecture featuring four hidden layers coupled with the SGD optimizer, as well as a five-layer model with RMSProp, provided the most accurate and reliable results, reaching a testing accuracy of 91%.

Jain et al. experimented with several customized CNNs of varying depths alongside popular transfer learning architectures like VGG16, VGG19, ResNet50, and Inception-v3 [6]. The evaluation was carried out using 5,840 chest X-ray images (5,216 for training and 624 for testing) sourced from the Kaggle repository. They concluded that their customized 3-layer CNN (Model 2) and the VGG19 network outperformed the others, achieving validation accuracies of 92.31% and 88.46%, respectively, making them highly consistent and efficient for diagnostic purposes.

Aledhari et al. proposed an optimized diagnosis system by fine-tuning prominent pre-trained CNN models, specifically comparing VGG16, ResNet-50, and Inception v3, tailored with modified final dense layers [7]. The researchers utilized a massive open dataset from the National Institute of Health (NIH), containing over 112,000 chest X-ray images from more than 30,000 patients. They concluded that their fine-tuned VGG16 model achieved a promising 68% average accuracy and 69% specificity, significantly improving upon earlier state-of-the-art baseline models on this complex dataset.

Sandler et al. introduced MobileNetV2, an architecture utilizing an inverted residual structure with linear bottlenecks and lightweight depth wise convolutions specifically tailored for mobile environments [8]. They validated their proposed methodology using standard benchmark datasets, including ImageNet for classification, COCO for object detection, and PASCAL VOC 2012 for semantic segmentation. The authors concluded that MobileNetV2 drastically reduces the required memory footprint and computational operations while maintaining or beating state-of-the-art accuracy across these varied computer vision tasks.

Slimi et al. designed a highly robust, attention-guided deep learning framework that integrates spatial feature extraction via CNNs, temporal dependencies using Gated Recurrent Units (GRUs), and biologically inspired Spiking Neural Networks (SNNs) [9]. Their proposed framework was rigorously tested on a publicly available dataset of 5,863 pediatric chest X-ray images originally sourced from Guangzhou Women and Children's Medical Center. They concluded that the hybrid SNN-GRU network achieved an exceptional 99.35% accuracy and demonstrated remarkable resilience to various image distortions, making it an ideal, energy-efficient tool for deployment in low-resource clinical settings.

Singh et al. developed a Quaternion Channel-Spatial Attention Network (QCSA) that directly incorporates spatial and channel attention mechanisms into a Quaternion Convolutional Neural Network (QCNN) [10]. They evaluated this advanced architecture using a Kaggle dataset composed of 5,856 chest X-ray images from infants aged one to five years. The study concluded that the QCSA model achieved an accuracy of 94.53%, proving that integrating attention mechanisms within quaternion networks substantially enhances feature extraction and the overall detection of pneumonia.

Özdemir and Keleş proposed a deep learning-based binary classification framework for detecting COVID-19 from chest X-ray images by utilizing transfer learning on pre-trained Convolutional Neural Network (CNN) models, specifically DenseNet121, VGG19, and EfficientNetB5 [11]. To enhance the feature extraction process, their methodology incorporated advanced image preprocessing techniques, including Contrast Limited Adaptive Histogram Equalization (CHALE) and Bi-Histogram Equalization based on Adaptive Sigmoid Function (BEASF), which were fed into the models as multi-channel inputs. Their models were trained and evaluated using a balanced dataset consisting of 747 COVID-19 chest X-ray images sourced from the COVID-CXNet dataset and 754 normal chest X-ray images obtained from the NIH Chest X-ray dataset. The authors concluded that applying these histogram-based preprocessing filters significantly improved the classification success rate across all architectures, with their proposed EfficientNetB5 model achieving the highest performance, recording a 99% accuracy and an F1 score of 99%.

TABLE I. COMPARISON TABLE SUMMARIZING THE METHODOLOGIES, DATASETS, AND KEY VALUES OR RESULTS FOUND IN THE PROVIDED LITERATURE AND RDE-PNEUNET

| Ref. | Methodology / Models     | Dataset                                               | Key Results / Values                 |
|------|--------------------------|-------------------------------------------------------|--------------------------------------|
| [1]  | • Transfer Learning (TL) | • Kaggle repository<br>• 5,856 pediatric chest X-rays | • Best: ResNet152V2<br>• Acc: 83.17% |

|      |                                                                                                                                                                                                                              |                                                                                                                                                        |                                                                                                                                                                               |
|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|      | <ul style="list-style-type: none"> <li>Models: Xception, VGG16, ResNet152V2</li> </ul>                                                                                                                                       |                                                                                                                                                        | <ul style="list-style-type: none"> <li>Precision: 79.87%</li> <li>Recall: 97.69%</li> <li>F1: 80%</li> </ul>                                                                  |
| [2]  | <ul style="list-style-type: none"> <li>U-Net for lung segmentation</li> <li>TL (ResNet50, InceptionV3, InceptionResNetV2)</li> <li>Optimizers: Adam, SGD</li> </ul>                                                          | <ul style="list-style-type: none"> <li>Guangzhou / Kaggle</li> <li>5,232 chest X-rays</li> </ul>                                                       | <ul style="list-style-type: none"> <li>Best: ResNet50 + SGD</li> <li>Acc: 93.06%</li> <li>Precision: 88.97%</li> <li>Recall: 96.78%</li> <li>F1: 92.71%</li> </ul>            |
| [3]  | <ul style="list-style-type: none"> <li>CNN (DenseNet121 via TL)</li> <li>Random Forest</li> <li>Support Vector Machine (SVM)</li> </ul>                                                                                      | <ul style="list-style-type: none"> <li>Kaggle Chest X-Ray</li> <li>5,856 images</li> </ul>                                                             | <ul style="list-style-type: none"> <li>Best: DenseNet121 CNN</li> <li>Multiclass Acc: 81%</li> <li>Sens: 79.53%</li> <li>Spec: 89.31%</li> <li>AUC: 84.2% - 91.85%</li> </ul> |
| [4]  | <ul style="list-style-type: none"> <li>Survey of ML/DL algorithms</li> <li>Pre-trained CNNs vs. from scratch</li> <li>CNN features + SVM/XGBoost</li> </ul>                                                                  | <ul style="list-style-type: none"> <li>RSNA, Kaggle, Guangzhou</li> <li>Multiple benchmark datasets</li> </ul>                                         | <ul style="list-style-type: none"> <li>VGG-19 Acc: 97.3%</li> <li>ResNet-50 Acc: 96.2%</li> <li>SVM + CNN: Highest precision</li> </ul>                                       |
| [5]  | <ul style="list-style-type: none"> <li>Custom CNNs (1 to 5 hidden layers)</li> <li>Optimizers: RMSProp, Adam, SGD</li> </ul>                                                                                                 | <ul style="list-style-type: none"> <li>Kaggle repository</li> <li>5,856 chest X-rays</li> </ul>                                                        | <ul style="list-style-type: none"> <li>Best: 4-layer CNN + SGD &amp; 5-layer CNN + RMSProp</li> <li>Test Acc: 91%</li> </ul>                                                  |
| [6]  | <ul style="list-style-type: none"> <li>Custom CNNs (2 &amp; 3 layers)</li> <li>TL: VGG16, VGG19, ResNet50, Inception-v3</li> </ul>                                                                                           | <ul style="list-style-type: none"> <li>Kaggle repository</li> <li>5,840 images (5,216 train, 624 test)</li> </ul>                                      | <ul style="list-style-type: none"> <li>Best: Custom 3-layer CNN</li> <li>Acc: 92.31%, Recall: 98%</li> <li>VGG19 Acc: 88.46%</li> </ul>                                       |
| [7]  | <ul style="list-style-type: none"> <li>Fine-tuned pre-trained CNNs</li> <li>Models: VGG16, ResNet-50, Inception v3</li> </ul>                                                                                                | <ul style="list-style-type: none"> <li>NIH open dataset</li> <li>112,000 images, 30,000+ patients</li> </ul>                                           | <ul style="list-style-type: none"> <li>Best: Fine-tuned VGG16</li> <li>Test Acc: 75%</li> <li>Avg Acc: 68%</li> <li>Avg Spec: 69%</li> </ul>                                  |
| [8]  | <ul style="list-style-type: none"> <li>MobileNetV2 architecture</li> <li>Inverted residuals</li> <li>Lightweight depth wise convolutions</li> </ul>                                                                          | <ul style="list-style-type: none"> <li>ImageNet, COCO, PASCAL VOC 2012</li> </ul>                                                                      | <ul style="list-style-type: none"> <li>ImageNet Top-1 Acc: 72%</li> <li>Parameters: 3.4 million</li> <li>Multiply-Adds: 300M</li> </ul>                                       |
| [9]  | <ul style="list-style-type: none"> <li>Hybrid Spiking Neural Network (SNN)</li> <li>Gated Recurrent Units (GRU)</li> <li>CNN spatial extraction + Attention</li> </ul>                                                       | <ul style="list-style-type: none"> <li>Guangzhou Medical Center</li> <li>5,863 pediatric chest X-rays</li> </ul>                                       | <ul style="list-style-type: none"> <li>Acc: 99.35%</li> <li>Precision: 99.1%</li> <li>Recall: 99.5%</li> <li>AUC-ROC: 99%</li> </ul>                                          |
| [10] | <ul style="list-style-type: none"> <li>Quaternion CNN (QCNN)</li> <li>Quaternion Channel-Spatial Attention (QCSA)</li> </ul>                                                                                                 | <ul style="list-style-type: none"> <li>Kaggle dataset</li> <li>5,856 chest X-rays</li> </ul>                                                           | <ul style="list-style-type: none"> <li>Acc: 94.53%</li> <li>Precision: 93.56%</li> <li>F1-score: 96.14%</li> <li>AUC: 89%</li> </ul>                                          |
| [11] | <ul style="list-style-type: none"> <li>TL (DenseNet121, VGG19, EfficientNetB5)</li> <li>Preprocessing: CHALE and BEASF</li> </ul>                                                                                            | <ul style="list-style-type: none"> <li>COVID-CXNet and NIH Chest X-ray</li> <li>1,501 images (balanced)</li> </ul>                                     | <ul style="list-style-type: none"> <li>Best: EfficientNetB5 + CHALE/BEASF</li> <li>Acc: 99%</li> <li>F1-score: 99%</li> </ul>                                                 |
| [12] | <ul style="list-style-type: none"> <li>Transfer Learning (Xception, InceptionV3, VGG16)</li> <li>XAI Methods (LIME, SHAP, Saliency Map, Grad-CAM, Grad-CAM++)</li> <li>Proposed: Ensemble XAI (SHAP + Grad-CAM++)</li> </ul> | <ul style="list-style-type: none"> <li>2,235 clinical chest X-rays</li> <li>1,475 prospective chest X-rays</li> <li>NIH Chest X-ray dataset</li> </ul> | <ul style="list-style-type: none"> <li>Acc: 89.9%, AUC: 89%</li> <li>F1-score: 50%</li> <li>Mean Set IOU: 36%</li> <li>Radiologists' Trust: 70.2%</li> </ul>                  |

|                |                                                                                                                                                          |                                                                                                                                                     |                                                                                                                                                           |
|----------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
| RDE - Pneu Net | <ul style="list-style-type: none"> <li>Soft-voting Ensemble</li> <li>ResNet50, DenseNet121, EfficientNetB0</li> <li>Explainable AI (Grad-CAM)</li> </ul> | <ul style="list-style-type: none"> <li>Kaggle Chest X-ray Pneumonia</li> <li>5,863 chest X-rays</li> <li>5,863 images (stratified split)</li> </ul> | <ul style="list-style-type: none"> <li>Best: Weighted Ensemble</li> <li>Acc: 93%</li> <li>F1-score: 93%</li> <li>Spec: 92%</li> <li>AUC: 97.5%</li> </ul> |
|----------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|

Comparison of methodologies, datasets, and key values or results found in the provided literature and our proposed RDE-PneuNet are given in Table I. In this study, a comprehensive comparative analysis was conducted using ResNet50, DenseNet121, and EfficientNet architectures for pneumonia detection, and the outputs of these models were combined using an ensemble learning approach. The results obtained on the utilized dataset are evaluated not only in terms of accuracy but also for generalizability and stability. Additionally, a Grad-CAM-based explainability method was applied to analyze the model's decision-making mechanism. The proposed approach aims to offer a more balanced, applicable, and reliable solution compared to existing studies in the literature.

The main novelty of this study lies in the integration of a weighted ensemble strategy with threshold optimization and class imbalance handling, aiming to achieve a clinically balanced performance rather than maximizing accuracy alone. Unlike existing studies that focus primarily on sensitivity, the proposed approach introduces a decision mechanism that effectively reduces false positives while maintaining high detection capability. Additionally, the use of Grad-CAM provides interpretability, ensuring that the model decisions are clinically meaningful and trustworthy.

## II. MATERIAL AND METHODS

In this study, a deep learning-based classification system was developed to detect pneumonia from chest X-ray images. The proposed method consists of data preprocessing and augmentation techniques, transfer learning-based modeling, mitigating class imbalance, fine-tuning applications, combining multiple model outputs (ensemble learning), and ensuring the explainability of model decisions.

### A. Dataset and Preprocessing

The Chest X-Ray Pneumonia dataset, shared by Paul Mooney on the Kaggle platform, was used for pneumonia detection [19]. The dataset contains a total of 5,863 chest X-ray images. Because the original validation set was highly limited only 16 images, the dataset was restructured using the 5,216 images from the training set to ensure a more reliable evaluation. By applying stratified splitting, the training set was determined as 4,172 images, and the validation set as 1,044 images. An independent test set of 624 images was used for the final testing phase

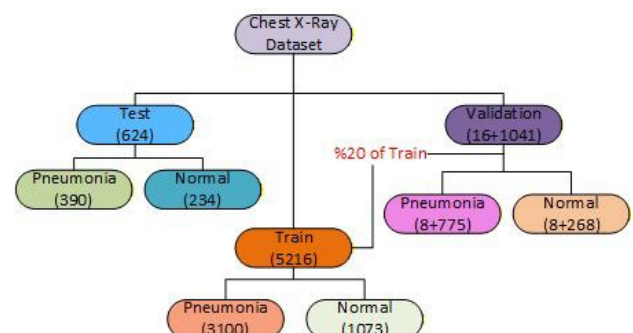


Fig. 1. Chest X-ray dataset split into training, validation, and test sets.

Fig.1 is used to visually outline the hierarchical dataset directory structure, showing the precise stratified splitting strategy applied to the training, validation, and test subsets.

TABLE II. DATA PROCESSING AND AUGMENTATION TECHNIQUES

| Technique        | Description                                           |
|------------------|-------------------------------------------------------|
| Resizing         | All images were resized to $224 \times 224$ pixels    |
| Normalization    | Pixel values were scaled using standard normalization |
| Rotation         | Random rotation up to 10 degrees                      |
| Zoom             | Random zoom range of 10%                              |
| Width Shift      | Horizontal shift of 10%                               |
| Height Shift     | Vertical shift of 10%                                 |
| Horizontal Flip  | Random horizontal flipping applied                    |
| Validation Split | 20% of training data used for validation              |

Table II is used to list the specific preprocessing steps and augmentation parameters, such as resizing, random rotation, zooming, and shifting, applied to enhance the diversity of the training images.

### B. Model Architecture and Training Process

Three distinct deep learning architectures featuring diverse structural properties were evaluated: ResNet50 [18], DenseNet121 [15], and EfficientNetB0 [16]. These models were initialized with pre-trained weights on the ImageNet dataset via a transfer learning approach. Initially, the general feature extraction layers of the models were frozen, and then fine-tuning was applied by making the final layers trainable specifically for this dataset.

#### Algorithm 1: Proposed Weighted Ensemble Method

Input: Test dataset  $T$ , trained models  $M_1, M_2, M_3$   
 1:  $P_1 \leftarrow M_1(T), P_2 \leftarrow M_2(T), P_3 \leftarrow M_3(T)$   
 2:  $[w_1, w_2, w_3] \leftarrow \text{Normalize}(acc_1, acc_2, acc_3)$   
 3:  $P_{ens} \leftarrow w_1 \cdot P_1 + w_2 \cdot P_2 + w_3 \cdot P_3$   
 4: Apply threshold to control decision boundary:  $y_{pred} \leftarrow (P_{ens} > \theta)$   
 5: Select optimal configuration maximizing accuracy and F1-score  
 6: Return  $y_{pred}$

The proposed ensemble method which is shown as Algorithm 1, combines the outputs of three deep learning models using a weighted soft voting strategy and  $M_1, M_2$ , and  $M_3$  denote ResNet50, DenseNet121, and EfficientNetB0 models, respectively,  $P_1, P_2$ , and  $P_3$  represent their predicted probabilities, and  $w_1, w_2$ , and  $w_3$  are the corresponding weighting coefficients. The prediction probabilities obtained from each model are aggregated to produce a final ensemble probability [20]. A threshold mechanism is applied to control the decision boundary and improve classification reliability. Rather than employing the conventional fixed threshold (50%), an optimized threshold value was introduced to modulate the sensitivity-specificity trade-off, thereby significantly enhancing clinical reliability. The weights are assigned based on the individual model accuracies to prioritize more reliable models within the ensemble. This strategy enhances overall performance while reducing the influence of weaker models. Normalization ensures a balanced and stable contribution from all models. In equation (1);  $P_i$ , namely,  $P_1, P_2, P_3$  denote the predicted probabilities of ResNet50, DenseNet121, and EfficientNetB0, respectively, and  $w_1, w_2$ , and  $w_3$  represent their corresponding weighting coefficients, ensuring a reproducible and balanced ensemble formulation.

$$P_{ensemble} = \sum_{i=1}^3 w_i P_i \quad (1)$$

The weighting coefficients used for ResNet50, DenseNet121, and EfficientNetB0 were explicitly defined as  $w_1=36\%$ ,  $w_2=29\%$ , and  $w_3=35\%$ , respectively. These values were obtained through a systematic search process to maximize classification performance. The explicit definition of these coefficients ensures reproducibility and allows fair comparison with future studies. Also, experiments were performed on a high-performance workstation with an AMD Ryzen 7 7840HS processor, 32 GB of RAM, and an NVIDIA GeForce RTX 4060 GPU with 8 GB VRAM. The use of GPU acceleration significantly reduced training time and ensured a fair comparison of computational cost among the evaluated models.

TABLE III. HYPERPARAMETER CONFIGURATION OF RESNET50, DENSENET121, AND EFFICIENTNETB0 MODELS

| Parameter     | ResNet50             | DenseNet121          | EfficientNetB0       |
|---------------|----------------------|----------------------|----------------------|
| Batch Size    | 32                   | 32                   | 32                   |
| Epoch         | 10                   | 20                   | 20                   |
| Learning Rate | $1 \times 10^{-5}$   | $2 \times 10^{-5}$   | $1 \times 10^{-5}$   |
| Optimizer     | Adam                 | Adam                 | Adam                 |
| Loss Function | Binary Cross entropy | Binary Cross entropy | Binary Cross entropy |

Table III illustrates the exact training hyperparameters, including batch size, epochs, learning rate, and optimizer settings configured for the ResNet50, DenseNet121, and EfficientNetB0 models [14]. To benefit from the complementary features of different models, the model outputs were combined using an ensemble learning approach of soft voting. In this approach, the probability values generated by each model were averaged to obtain the final decision. The schematic workflow of the proposed ensemble model architecture is delineated in Fig. 2.

#### Proposed Ensemble Model Architecture

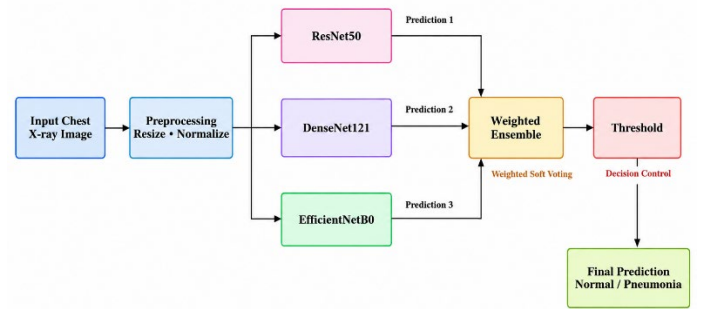


Fig. 2. RDE-PneuNet Ensemble Model Architecture

All images were resized to  $224 \times 224$  pixels and normalized. To improve the model's generalization ability and prevent overfitting, data augmentation operations were applied to the training dataset.

### III. EXPERIMENTAL RESULTS

The performances of the utilized ResNet50, DenseNet121, and EfficientNetB0 models were evaluated using accuracy, precision, recall, F1-score, and specificity metrics.

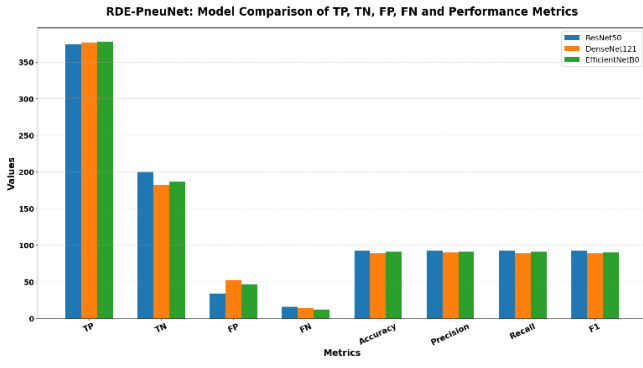


Fig. 3. RDE-PneuNet: Model Comparison of TP, TN, FP, FN and Performance Metrics

As illustrated in Fig. 3, while all architectures demonstrate substantial classification efficacy, EfficientNetB0 provides superior diagnostic reliability compared to ResNet50 and DenseNet121 due to its lower false negative rate and well-balanced metric profile.

Some metrics used to evaluate model performance indicate that in (2), precision represents the proportion of correctly predicted positive instances; in (3), recall represents the correctly identified actual positive cases; in (4), accuracy reflects the overall correctness, and in (5), F1-score expresses the balanced mean of precision and recall as follows [10]:

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \times 100\% \quad (4)$$

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (5)$$

In equations (2), (3), (4) and (5), the terms TP, TN, FP, and FN correspond to True Positives, True Negatives, False Positives, and False Negatives, respectively. The upcoming sections elaborate on the pixel intensity histograms, feature map visualizations, empirical evaluations of the three individual models, and a comparative analysis against prior methodologies established in the literature. The diagnostic efficacy of the proposed architecture was assessed using fundamental performance metrics, specifically precision, recall, accuracy, specificity, and the F1-score.

Precision acts as an indicator of diagnostic exactness, measuring the model's proficiency in avoiding false positives. A high precision score guarantees that healthy chest X-rays are rarely misdiagnosed as having viral or bacterial pneumonia. Recall, or sensitivity, evaluates the completeness of the diagnosis; a high recall signifies a superior disease detection rate, ensuring that true pneumonia infections are not overlooked. Accuracy represents the overall proportion of correct classifications made by the system. To provide a comprehensive and rigorous measure of the model's reliability, the F1-score calculates the harmonic mean of both precision and recall.

Table IV illustrates the comparison of the quantitative individual performance metrics of the standalone models on the internal test set before they were combined into an ensemble.

TABLE IV. MODEL COMPARISON AND FINAL RDE-PNEUNET SELECTION AND PERFORMANCE METRICS.

| Model          | Accuracy | Precision | Recall | F1  | Specify |
|----------------|----------|-----------|--------|-----|---------|
| ResNet50       | 92%      | 92%       | 92%    | 92% | 85%     |
| DenseNet121    | 89%      | 90%       | 89%    | 89% | 78%     |
| EfficientNetB0 | 91%      | 91%       | 91%    | 90% | 80%     |

As seen in the results, ResNet50 achieved a strong performance with 92% accuracy, precision, recall, and F1-score. EfficientNetB0 showed similar performance with 91% accuracy and balanced metric values. DenseNet121 exhibited slightly lower performance with 89% accuracy. In terms of specificity, ResNet50 reached the highest value at 85%.

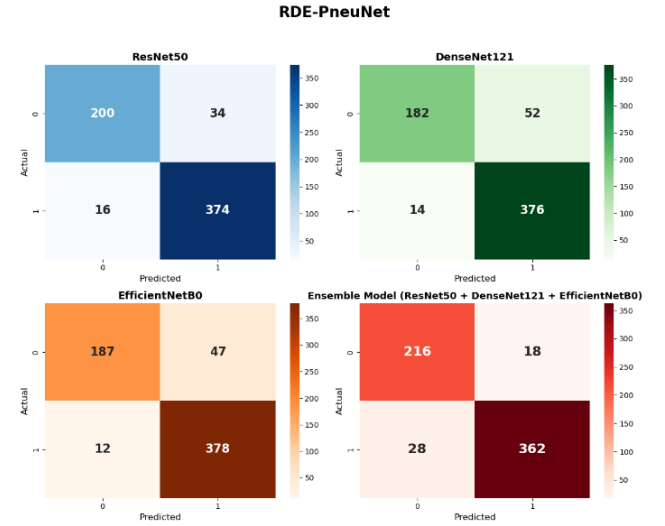


Fig. 4. Confusion matrices of ResNet50, DenseNet121, EfficientNetB0, and the proposed ensemble model

Fig.4 displays the confusion matrices for ResNet50, DenseNet121, EfficientNetB0, and the proposed ensemble model, providing a clear visual breakdown of correct and incorrect predictions per class. The confusion matrices reveal that ResNet50 correctly classified 200 normal images and 374 Pneumonia images, while the ensemble model correctly classified 216 normal and 362 Pneumonia images, demonstrating a lower error rate specifically in normal class.

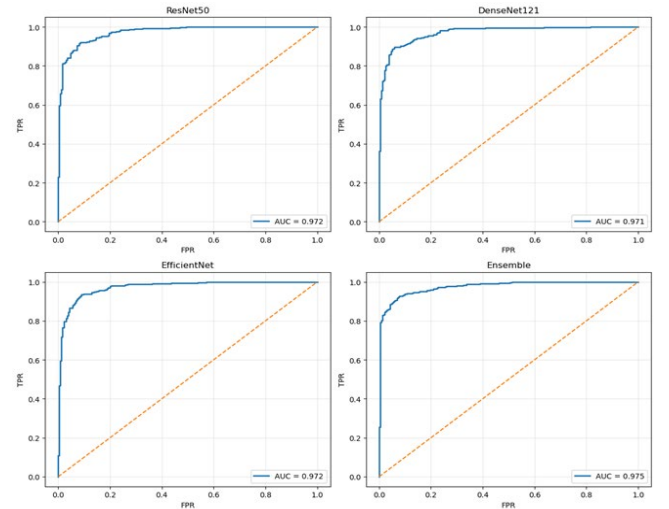


Fig. 5. ROC curves of ResNet50, DenseNet121, EfficientNet, and the proposed ensemble model.

The inter-class discrimination performances of the models were evaluated using Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values. Fig.5 presents the ROC curves for the individual models and the ensemble model, graphically illustrating the trade-off between the true positive rate and false positive rate

Fig.6 plots the training and validation accuracy and loss curves across epochs, proving that the models learned successfully and converged without significant overfitting. Quantitatively, the precise AUC scores are derived from the ROC curves, confirming that the proposed ensemble model achieves the peak performance with a score of 97.5%. The training processes of the models were analyzed by observing the epoch-based changes in accuracy and loss values. Different ensemble configurations were evaluated, and their results were comparatively analyzed. Performance of the final weighted ensemble model of the proposed model achieves a balanced and robust performance with 93% accuracy, precision, recall, and F1-score, along with a specificity of 92%.

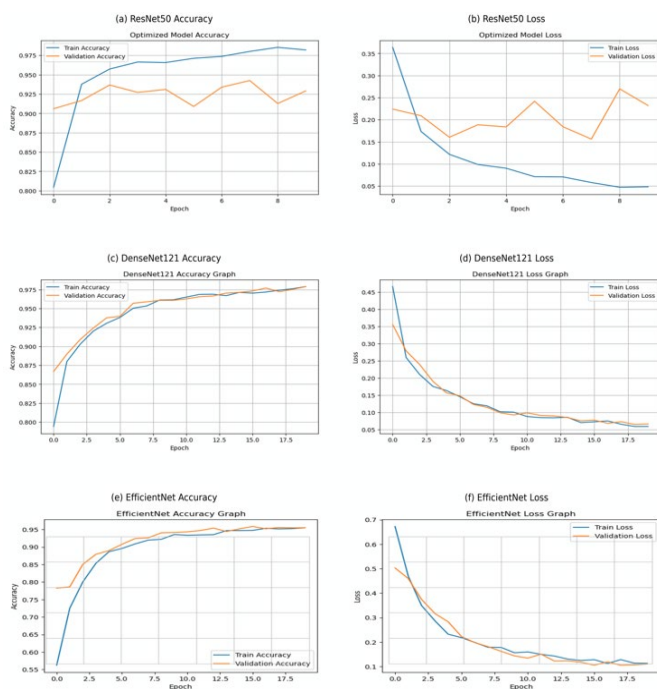


Fig. 6. Training and validation accuracy and loss curves across epochs.

These results indicate that the model maintains a strong equilibrium between sensitivity and specificity, making it suitable for reliable clinical decision support. The consistency across all evaluation metrics demonstrates the effectiveness of the ensemble strategy in capturing discriminative features from chest X-ray images.

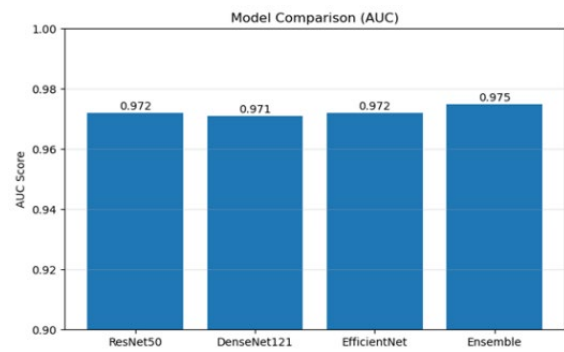


Fig. 7. Overall comparison of AUC

Fig.7 illustrates the overall AUC and accuracy scores of the standalone models against the final ensemble model using

#### IV. EXPLAINABILITY AND GRAD-CAM

A major limitation of deep learning-based models is the inability to directly interpret their decision-making processes. Therefore, the Grad-CAM (Gradient-weighted Class Activation Mapping) method was used by adopting an Explainable AI approach [17].

Fig.8 provides Grad-CAM-based visual evidence of how the proposed RDE-PneuNet model makes its predictions.

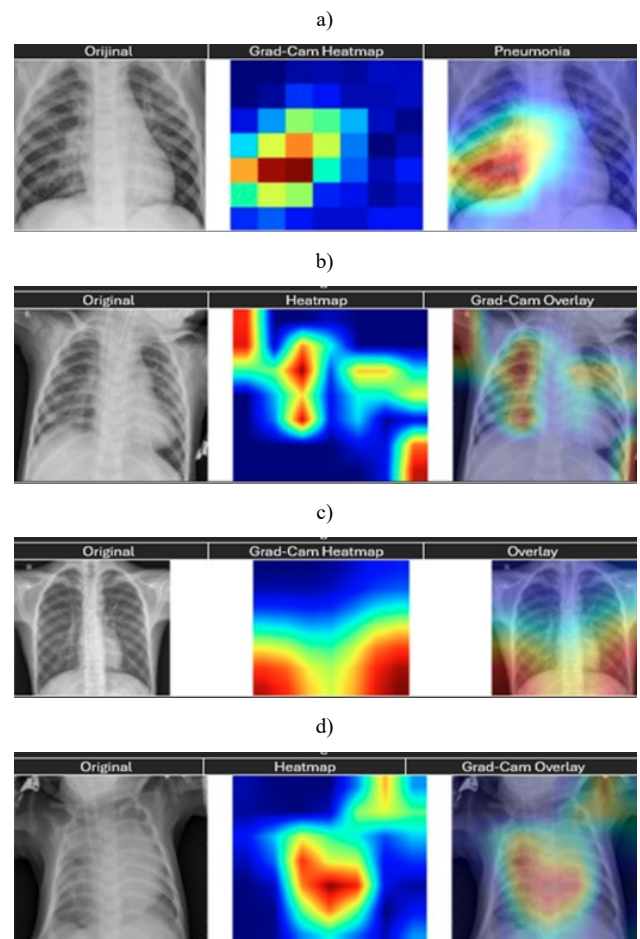


Fig. 8. Grad-CAM visualizations of the proposed RDE-PneuNet model. a) Correct localization on infected lung regions in a pneumonia case. b) Partial attention outside lung boundaries despite correct classification. c) Low activation in lung regions for a normal case. d) Strong activation highlighting severe pneumonia patterns.

As seen in Fig.8.a, the model accurately localizes the infected lung region in a pneumonia case. In Fig.8.b, the model still produces the correct prediction, although the attention map partially extends outside the lung boundaries, revealing a slight localization deviation. In contrast, Fig.8.c shows a normal chest X-ray with relatively low activation over the lung region, indicating that the model does not focus strongly on healthy tissue. Moreover, Fig.8.d demonstrates a severe pneumonia case, where the activation is concentrated over the pathological area, further supporting the clinical relevance and interpretability of the proposed model.

With Grad-CAM, class activation maps were generated by combining feature maps from the final convolutional layer with class-specific gradient information. By overlaying these maps onto the input images, it was visually confirmed that the model primarily focuses on lung regions and anatomically meaningful features. The analysis proved that the model has an interpretable and reliable structure, strengthening its potential for clinical use. Fig.8 is used to visualize Grad-CAM results for samples a, b, c, and d. The highlighted regions indicate the areas that contribute most to the model's decisions. The model consistently attends to relevant pulmonary regions, thereby validating its interpretability and strict alignment with clinical diagnostic expectations. Grad-CAM analysis reveals that the proposed RDE-PneuNet model consistently attends to clinically meaningful pulmonary regions in pneumonia cases while suppressing irrelevant activations in normal samples, thereby demonstrating strong discriminative capability and clinically interpretable decision-making, despite minor localization deviations.

## V. DISCUSSION

Experimental evaluation confirms that the RDE-PneuNet ensemble establishes superior clinical equilibrium over individual baseline networks. While standalone models tend to over-detect pneumonia—yielding elevated false-positive rates—the proposed weighted soft-voting strategy significantly enhances healthy subject identification, securing a 92% specificity. This diagnostic balance minimizes patient anxiety and overtreatment while maintaining a high sensitivity of 93% to ensure critical pathologies are not overlooked. Unlike alternative frameworks that sacrifice specificity for raw accuracy, RDE-PneuNet delivers a clinically safer solution via an optimally stratified dataset without relying on overly complex hybrid topologies. Furthermore, the integration of Grad-CAM heatmaps resolves the inherent "black-box" limitations of standard CNNs by verifying that predictions are driven by anatomically relevant features, such as pulmonary opacities, rather than background artifacts.

Future research will focus on validating this framework using real-world clinical data from Kocaeli City Hospital under institutional ethics committee approval. To preserve patient confidentiality while assessing generalizability, cross-dataset validation and federated learning paradigms will be explored. Additionally, future iterations will investigate Vision Transformers (ViTs), lightweight deep learning architectures, and meta-heuristic optimization algorithms to refine the ensemble selection strategy. Ultimately, the objective is to optimize the computational footprint of RDE-PneuNet for embedded edge hardware deployment, transforming it into a portable, real-time clinical decision support system.

## REFERENCES

- [1] B. Tripathy, S. Khan, S. Beborra, A. Kamal, S. S. Tripathy, M. Fazil, and A. M. Albarrak, "TL-PneuNet: A transfer learning-based pneumonia classification framework," *Scientific Reports*, vol. 15, no. 1, pp. 1–12, 2025.
- [2] A. Manickam, J. Jiang, Y. Zhou, A. Sagar, R. Soundrapandiyan, and R. D. J. Samuel, "Automated pneumonia detection on chest X-ray images: A deep learning approach with different optimizers and transfer learning architectures," *Measurement*, vol. 184, p. 109953, 2021.
- [3] L. de F. A. Trindade, L. A. da Silva, M. L. C. S. Serrano, I. G. S. de A. Araújo, F. S. F. Prof., and V. C. Amador, "Comparative analysis of machine learning techniques for pneumonia detection in chest X-ray images," *Cureus*, vol. 17, no. 1, pp. 1–10, 2025.
- [4] D. Rai, H. Verma, M. Arham, and S. Shukla, "A comprehensive survey on comparative analysis of machine learning algorithms in pneumonia detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, pp. 1–10, 2023.
- [5] Y. S. Chowdhury, R. Dasgupta, and S. Nanda, "Analysis of various optimizers on CNN models in the application of pneumonia detection," *Procedia Comput. Sci.*, vol. 218, pp. 123–130, 2022.
- [6] R. Jain, P. Nagraath, G. Kataria, V. S. Kaushik, and D. J. Hemanth, "Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning," *Measurement*, vol. 165, p. 108046, 2021.
- [7] M. Aledhari, S. Joji, M. Hefaida, and F. Saeed, "Optimized CNN-based diagnosis system to detect pneumonia from chest radiographs," *IEEE Access*, vol. 8, pp. 1–10, 2020.
- [8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4510–4520.
- [9] H. Slimi, A. Balti, S. Abid, and M. Sayadi, "Trustworthy pneumonia detection in chest X-ray imaging through attention-guided deep learning," *Scientific Reports*, vol. 15, no. 1, pp. 1–15, 2025.
- [10] S. Singh, M. Kumar, A. Kumar, B. K. Verma, and S. Shitharth, "Pneumonia detection with QCSA network on chest X-ray," *Scientific Reports*, vol. 13, no. 1, pp. 1–12, 2023.
- [11] Z. Özdemir and H. Y. Keleş, "Derin öğrenme ile göğüs röntgeni görüntülerinde COVID-19 algılama," in *Proc. 29th Signal Processing and Communications Applications Conf. (SIU)*, 2021, pp. 1–4.
- [12] L. Zou *et al.*, "Ensemble Image Explainable AI (XAI) Algorithm for Severe Community-Acquired Pneumonia and COVID-19 Respiratory Infections," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 2, pp. 242–254, April 2023, doi: 10.1109/TAI.2022.3153754.
- [13] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.
- [14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.
- [15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700–4708.
- [16] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.
- [17] 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," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
- [19] P. Mooney, "Chest X-Ray Images (Pneumonia)," *Kaggle*, 2018. [Online] Available: <https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia>
- [20] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2025.