Identifying pneumonia in chest X-ray images

Maciej Lorens, 419763

1. Introduction

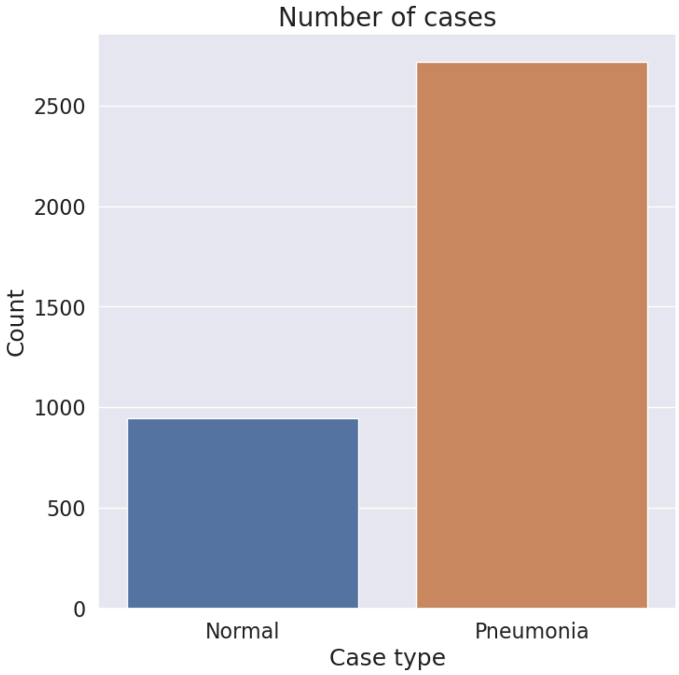
Annually, pneumonia takes the lives of around 2 million children under the age of 5, persistently standing as the foremost cause of childhood mortality according to the World Health Organization (WHO). WHO highlights that nearly all cases (95%) of new-onset childhood clinical pneumonia surface in developing regions, primarily Southeast Asia and Africa, where swift interpretation of radiographic data is often lacking. This project aims to address this challenge by developing an ensemble deep learning model capable of accurately identifying pneumonia in chest X-ray images.

2. Dataset description

Source: https://data.mendeley.com/datasets/rscbjbr9sj/2

The dataset comprises 5,856 JPEG chest X-ray images from children, each accompanied by a binary label (1 for Pneumonia, 0 for Normal). Within the training set, 3,883 images portray pneumonia (2,538 bacterial and 1,345 viral cases), while 1,349 images represent chest X-rays without anomalies. Moving to the test set, it consists of 234 normal images and 390 chest X-ray images indicating pneumonia. There is a clear imbalance in the target value, the majority of which are positive cases. The difference was visualized in Figure 1.

Figure 1. Imbalance in training set classes

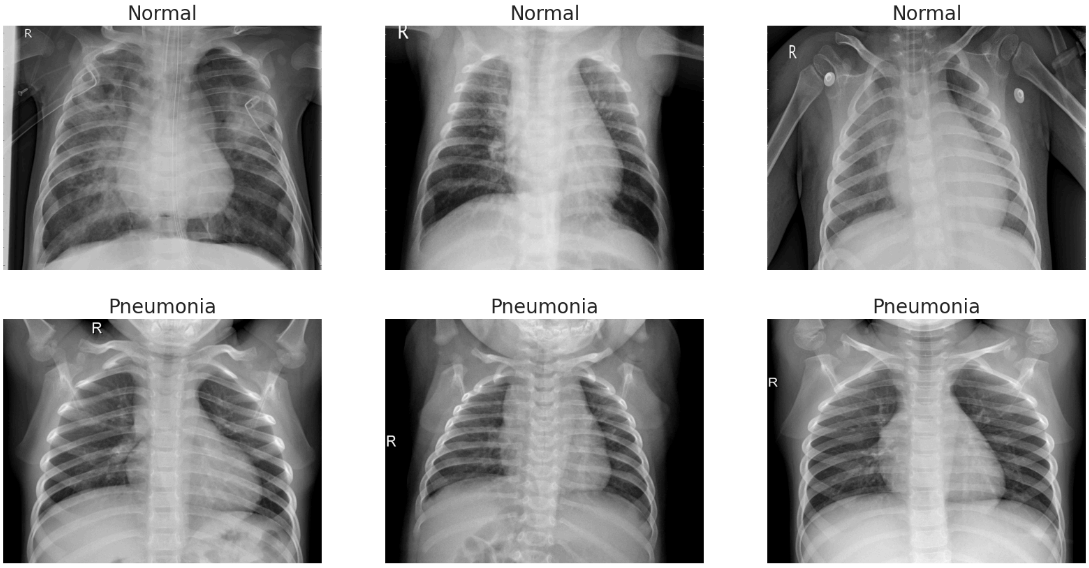


3. Methodology

3.1. Data preprocessing

X-ray images are typically examined for distinctive white spots in the lungs, which can indicate the presence of pneumonia. Examples of the types of images that the models will be trained and tested on are provided in Figure 2.

Figure 2. Examples of X-ray images



Out of the training images, 70% were allocated for training, and 30% for validation, utilizing a stratified train-test split. Due to variations in image sizes, each image was resized to 150x150 and normalized. Additionally, within the training set, images underwent augmentation through horizontal flipping with a probability of 20% and random rotation of 30 degrees maximum.

3.2. Metrics for evaluation of results

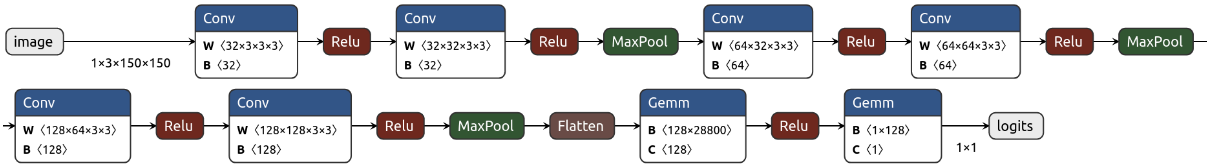
In the context of identifying pneumonia in developing regions, it is crucial to account for limited resources, emphasizing the need to balance Precision and Recall.

- Area Under the Receiver Operating Characteristic (ROC-AUC) - Evaluates the model's ability to distinguish between positive and negative classes across different threshold values, providing a comprehensive measure of its discriminatory power. It was also used in the original paper that introduced the dataset.
- Balanced Accuracy - It takes into account the imbalance between the classes and provides a more balanced view of the model's performance.
- F1 Score - Harmonic mean of precision and recall, a single metric that considers both false positives and false negatives.

3.3. Models

The first evaluated model is a simple benchmark CNN model. Its architecture was presented in Figure 3.

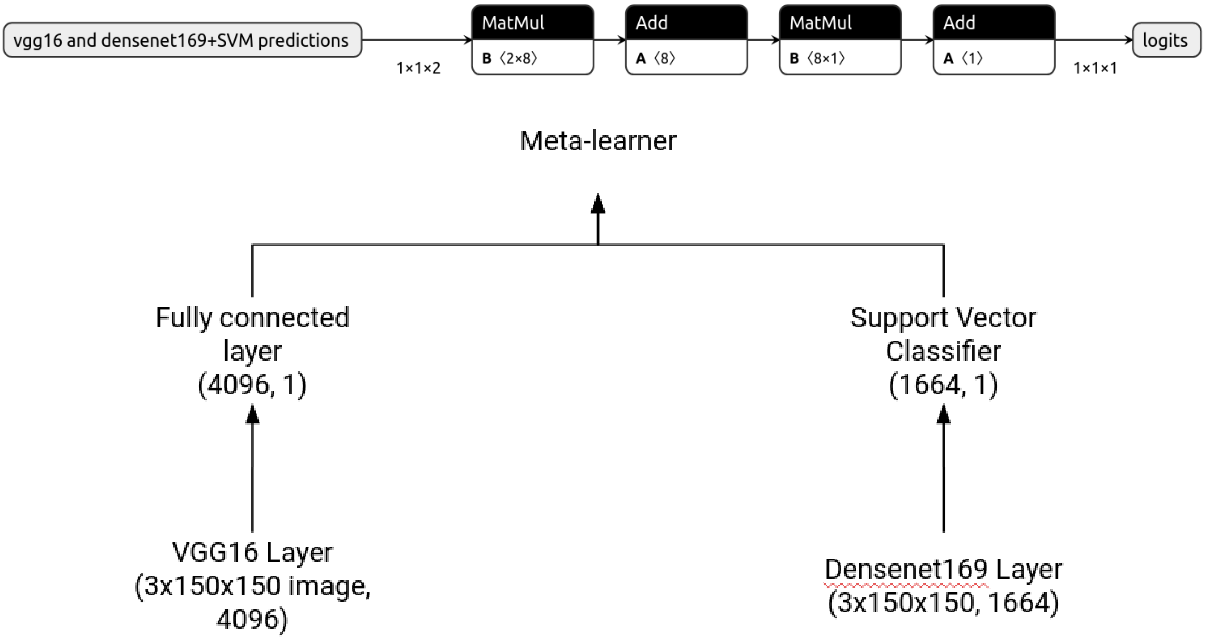
Figure 3. Architecture of the benchmark CNN model



The other two models are transfer learning models that were leveraged as inputs to a stacking ensemble model. In the first model, the architecture of the VGG16 model with frozen weights (not trainable) was used, and its final layer was replaced with a trainable fully-connected layer. In a similar fashion, the second model took an approach using

Densenet169, with the final layer being replaced by a Support Vector Classifier. The last model is a meta learner that takes predictions of transfer learning models, expressed as probabilities and later combines them via a neural network.

Figure 4. Architecture of the stacking model


vgg16 and densenet169+SVM predictions
1×1×2
MatMul
B ⟨2×8⟩
Add
A ⟨8⟩
MatMul
B ⟨8×1⟩
Add
A ⟨1⟩
1×1×1
logits
Meta-learner
Fully connected layer (4096, 1)
Support Vector Classifier (1664, 1)
VGG16 Layer (3x150x150 image, 4096)
Densenet169 Layer (3x150x150, 1664)

3.4. Hyperparameter optimization

The optimal hyperparameters were determined through grid search, except for the regularization parameter C in SVM, which was identified using Bayesian search. The chosen loss function was weighted binary cross-entropy to account for the class imbalance. The search spaces and best hyperparameters were outlined in Tables 1. and 2.

Table 1. Hyperparameter search space

VGG16	**Densenet169 + SVM**	**Meta-learner**
● Batch size = [64, 128, 256] ● Dropout = [0.2, 0.4] ● Learning rate = [0.01, 0.001]	● C = (0.01, 0.1) ● Kernel = [“linear”, “poly”, “rbf”] ● Class weight = [“balanced”, None]	● Batch size = [64, 128, 256] ● Dropout = [0.2, 0.4] ● Learning rate = [0.001, 0.01]

		- Number of neurons = [2, 4, 6, 8] - Activation function = [Identity, Sigmoid, ReLU]

Table 2. Best hyperparameters

VGG16	**Densenet169 + SVM**	**Meta-learner**
- Batch size = 64 - Dropout = 0.2 - Learning rate = 0.001	- C = 0.01 - Kernel = “linear” - Class weight = “balanced”	- Batch size = 64 - Dropout = 0.2 - Learning rate = 0.01 - Number of neurons = 8 - Activation function = Identity

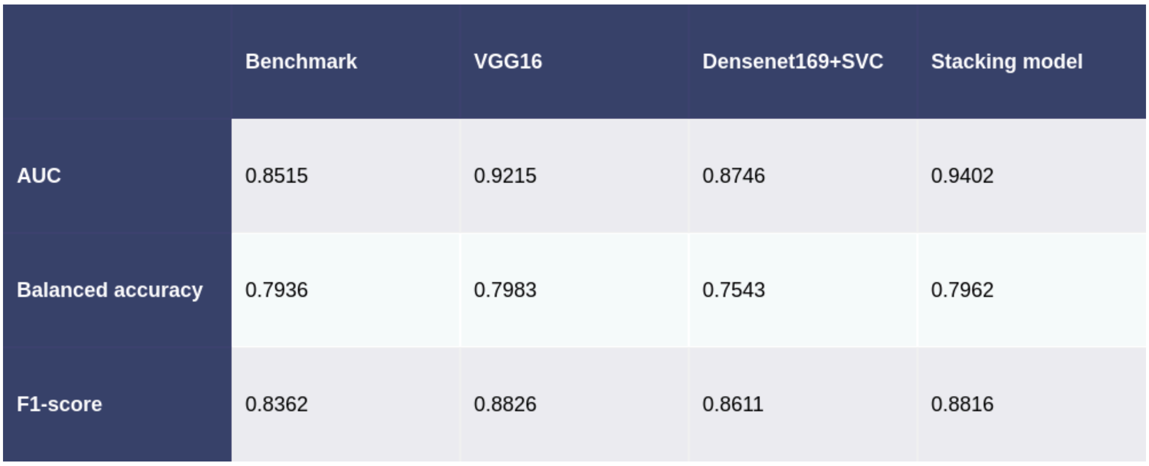
As the models provide probabilities, it is essential to establish suitable thresholds for assessing their performance using balanced accuracy and F1-score. These thresholds were determined through a standard approach of maximizing the G-mean (geometric mean of sensitivity and specificity) on the validation set. The selected thresholds are as follows: 0.4 for the benchmark, 0.5 for VGG16, 0.86 for Densenet169 + SVM, and 0.63 for the stacking model.

4. Model training and results

Each deep learning model (Benchmark, VGG16, Meta-learner) was trained for 20 epochs with early stopping monitoring the validation loss, which was the same weighted binary cross entropy loss function.

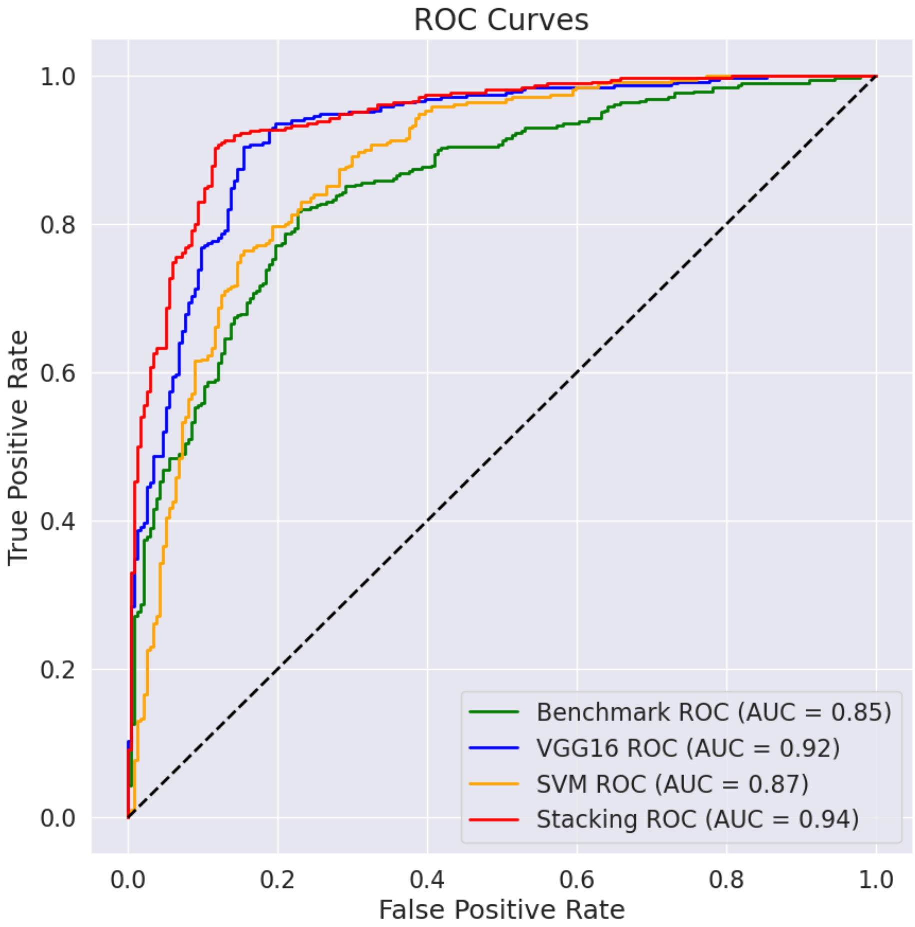
The metrics on the test set are presented in Figure 5. The stacking model exhibited superior performance compared to other models in terms of the AUC score, which was also exhibited on the ROC curve in Figure 6. However, with the chosen threshold for both balanced accuracy and F1-score, the VGG16-based model demonstrated better performance in terms of both balanced accuracy and F1-score.

Figure 5. Metrics on the test set



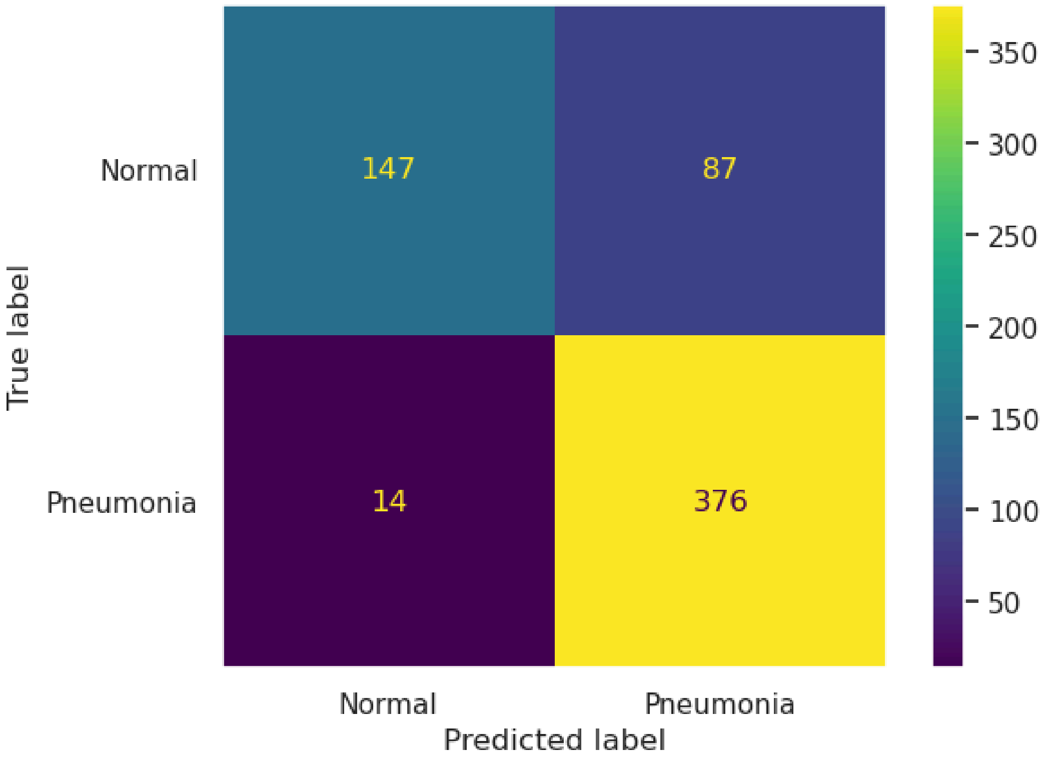
	Benchmark	VGG16	Densenet169+SVC	Stacking model
AUC	0.8515	0.9215	0.8746	0.9402
Balanced accuracy	0.7936	0.7983	0.7543	0.7962
F1-score	0.8362	0.8826	0.8611	0.8816

Figure 6. ROC curves of each model on the test set



Upon examining the confusion matrix in Figure 7., it is evident that there are fewer False Negatives than False Positives. This suggests that the model generally accurately identifies most cases of pneumonia but occasionally errors by assigning a positive label to healthy children. Further refinement is necessary, particularly given the context of identifying pneumonia in environments with limited resources, where both precision and recall are crucial.

Figure 7. Stacking model’s confusion matrix on the test set



5. Explainable Artificial Intelligence

As Neural Networks operate as "black box" models, discerning their decision-making process is not straightforward. To unravel the impact of each feature on predictions, Shapley values are commonly employed. Figure 8. visualizes the Shapley values of the VGG16 model, revealing a predominant focus on the lungs - a critical area for detecting signs of inflammation. Nevertheless, anomalies emerge, with some attention directed towards the patient's shoulders.

Figure 9. sheds light on the contribution of VGG16 and Densenet169 + SVM models to the stacking model's predictions. Notably, lower probabilities predicted by

Densenet169 + SVM exert a more pronounced impact in reducing the overall probability predicted by the stacking model.

Figure 8. Shapley values of VGG16

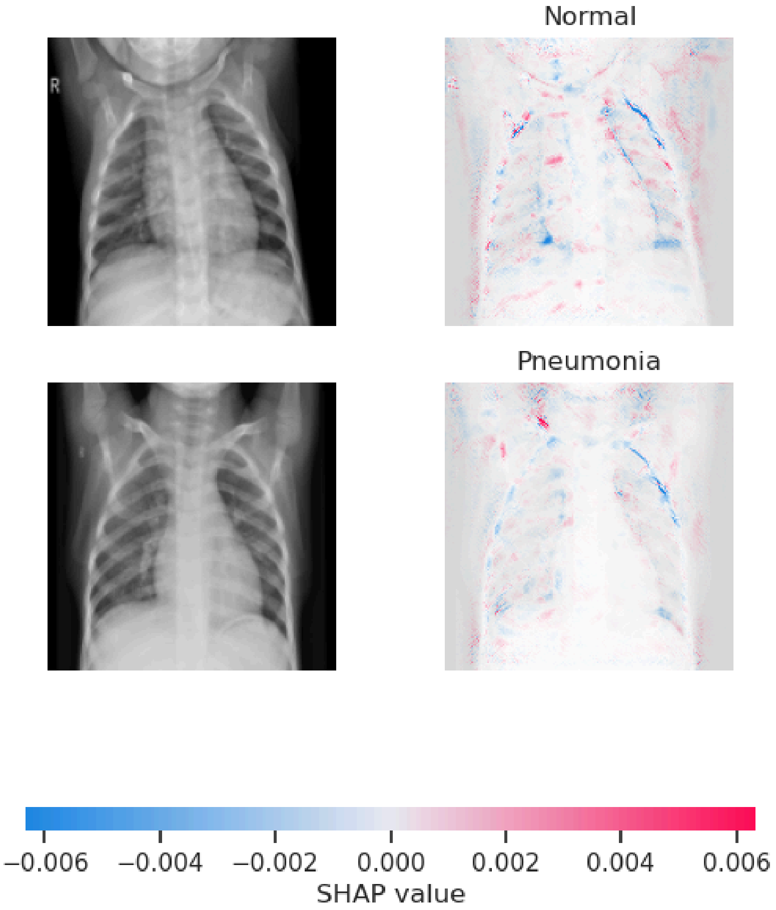
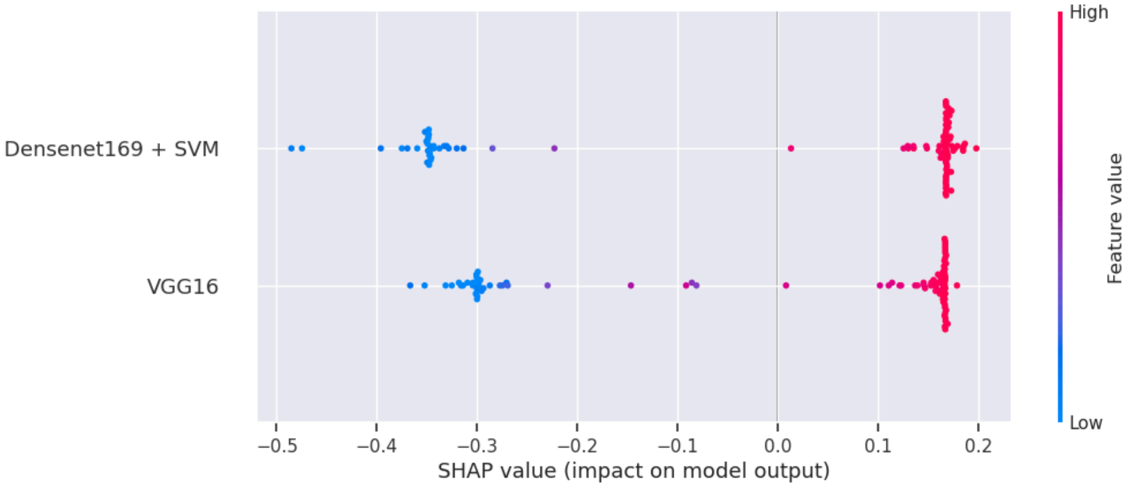


Figure 9. Shapley values of stacking model



5. Conclusions

In conclusion, this project aimed to combat childhood pneumonia, a leading cause of mortality in developing regions, by developing an ensemble deep learning model for chest X-ray analysis. While the stacking model showed superior AUC performance, the VGG16-based model demonstrated better balanced accuracy and F1-score. However, the occasional false positives revealed in the confusion matrix highlight the need for further refinement, particularly in resource-limited settings where precision and recall are vital. Ongoing efforts should focus on optimizing the model for improved sensitivity and specificity, enhancing interpretability, and facilitating seamless integration into real-world healthcare settings. The ultimate goal is to provide a reliable tool for timely pneumonia identification and intervention, contributing to the reduction of childhood mortality in regions with limited healthcare resources.

References

1. Kermany D, Goldbaum M, Cai W et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018; 172(5):1122-1131. doi:10.1016/j.cell.2018.02.010.
2. Cortes O, Diniz J, Silva LDJ (2023) “A novel ensemble CNN model for COVID-19 classification in computerized tomography scans”
3. Biloglav Z, Boschi-Pinto C, Campbell H, Mulholland K, Rudan I (2008) “Epidemiology and etiology of childhood pneumonia”