Comparative Study of Deep Learning Models for Pneumonia Classification

Betzy Babu Thoppil
Department of CSE
Amal Jyothi College of Engineering, Autonomous
Kanjirappally, Kottayam, India
betzybt@gmail.com

Midhun P Mathew
Department of CSE
Amal Jyothi College of Engineering, Autonomous
Kanjirappally, Kottayam, India
midhunpmathew@amaljyothi.ac.in

Sania Elsa Regi
Department of CSE
Amal Jyothi College of Engineering, Autonomous
Kanjirappally, Kottayam, India
saniaelsaregi@gmail.com

Nazreen Shanavaz
Department of CSE
Amal Jyothi College of Engineering, Autonomous
Kanjirappally, Kottayam, India
nazreenshanavaz2004@gmail.com

Unnimaya V Ashok
Department of CSE
Amal Jyothi College of Engineering, Autonomous
Kanjirappally, Kottayam, India
unnimayavashok@gmail.com

Nila S
Department of CSE
Amal Jyothi College of Engineering, Autonomous
Kanjirappally, Kottayam, India
nilasj21@gmail.com

***Abstract*—Deep learning is a powerful method for analyzing medical data such as detecting pneumonia and other respiratory diseases from chest X-rays. This paper presents a comparative analysis of a few of the most prominent convolutional neural network (CNN) architectures. These CNNs include VGG16, VGG19, DenseNet121, DenseNet201, MobileNetV1, MobileNetV2, InceptionV3, and Inception-ResNetV2.This study also explores a hybrid VGG19–Transformer architecture to enhance pneumonia detection by combining CNN-based spatial feature extraction with transformer-based global context learning. Each of these models was evaluated on a chest X-ray dataset while measuring a set of prediction performance metrics, namely, accuracy, precision, recall, and the F1 score. The results are heterogeneous with respect to the different models, and the highest levels of test accuracy were, however, 82.71% for VGG19 and 81.53% for MobileNetV1. Other architectures such as DenseNet and Inception variant models were noted to have competitive accuracy, but these models were significantly weak for the more difficult problem of class imbalance, particularly distinguishing bacterial from viral pneumonia. The trade-offs of different architectures are discussed, underscoring the trade-off of merely model accuracy for class-for robustness. These outcomes represent a critical foundation for further work aimed at the improvement of deep learning systems focused on the practical and validated clinical detection of pneumonia, and other respiratory diseases.**

***Index Terms*—Comparative Study of Deep Learning Models for Pneumonia Classification**

I. Introduction

A. Clinical Context and Motivation

Pneumonia is one of the most common illnesses and the most common cause of death globally, impacting most, young children, the elderly, and the immunocompromised. For effective treatment, accurate and timely diagnosis is necessary, but analyzing chest X-rays is an arduous and error-prone process. This is one of the reasons purely automated systems are in increasing demand, and will be useful for both low and high resource healthcare systems.

Medical image processing has been revolutionized using deep learning algorithms, especially Convolutional Neural Networks. VGG, DenseNet, MobileNet, Inception and a myriad of architectures are CNNs that can be and have been used for the classification of chest X-rays.The paper further introduces a hybrid deep learning framework, namely VGG19-Transformer, which combine convolutional neural networks with transformer encoders to simultaneously detect local spatial features and understand global contextual dependencies in chest X-ray images. There are, of course, varying degrees of success that depend on architectural choices, resources available, and the binary classification of pneumonia being bacterial or viral.

This research highlights the most clinically relevant CNNs, most of which have been used in chest X-ray classification, and their strengths and weaknesses. This, in turn, should improve automated detection of pneumonia in Real Time, and thus improve healthcare.

B. The Role and Limitations of Radiography

Chest X-rays (CXR) are efficient, cost-effective, and accessible imaging methods for the diagnosis of respiratory

diseases, leading to their frequent use in the field [1], [3], [6], [7], [9], [15]. Despite the rise of advanced imaging techniques such as Computed Tomography (CT), CXRs are almost always the first choice for pneumonia screening, as they are still the industry standard for primary assessment. Typical agricultural radiology signs such as opacities and infiltrates are frequently visible. [9].

As helpful as they may be, CXRs are notoriously difficult to interpret. The signs of pneumonia are visible, albeit subtle, and are often mistaken for other conditions, including lung cancer and benign lesions [1], [6]. This increases variability and the risk of misdiagnosis [6], [9], [15]. The issue of differentiating between bacterial and viral pneumonia, and between COVID-19 and other viruses, is exacerbated when the only tools available are X-rays [8], [9]. These obstacles only increase the importance of available and dependable automated Computer-Aided Diagnosis (CAD) systems.

C. Deep Learning as a Solution

Recent advances in Artificial Intelligence, in particular the domain of Deep Learning, have facilitated and improved the analysis of medical imaging (MI). Models based on convolutional neural networks (CNNs) have been shown to autonomously learn and hierarchically extract complex features, accomplishing a level of performance in various diagnostic tasks that rivals and even surpasses that of expert professionals [1], [3], [6], [12]. Their level of performance in visual pattern extraction and refinement makes them especially suited for the task of computer-aided pneumonia diagnosis from chest X-ray imaging.

The objective of the current study is to appraise and benchmark the classification performance of several pre-trained CNN architectures across various pneumonia imaged datasets to assess their output in terms of the performance and impact of potential diagnostic tools in healthcare systems.

II. Literature Review

A. Foundational Deep Learning Concepts

1) The Architecture of Convolutional Neural Networks (CNNs): The Structure Of Convolutional Neural Network (CNN): A deep neural network that handles pixel data is called CNN, and it is very important and common in medical image analysis. [3], [6]. A basic CNN contains the following:

- **Convolutional Layers:** Building blocks within modular neural networks are convolutional layers which are the most critical and primary components where image features such as edges and textures are detected and mapped using feature maps and filters (kernels). [1].
- **Activation Function:** This model uses an activation function like ReLU (f(x) = max(0,x)) to introduce the nonlinearity required to learn more complicated patterns. [3].
- **Pooling Layers:** This reduces the feature map size and the amount of calculations that need to be done for max pooling which is the most used type of pooling layer [6]. [6].
- **Fully Connected Layers:** After feature extraction, the layers are converted into a flattened vector which is then passed into the fully connected layers to facilitate the last classification. [1].

In this work, these building blocks form the basis of the pre-trained architectures evaluated, such as VGG, DenseNet, MobileNet, and Inception.

2) Transfer Learning is Key: With deep Learning CNNs, training them on domain-specific data is costly, and in medicine, data is usually very limited. Transfer learning becomes the standard approach [1], [9] - [12]. This is done by tuning a CNN model that has already been trained on a large, non-medical dataset, usually ImageNet, to a specific medical task. [9].

The initial layer focus on various features that are relevant in different areas. Subsequently added layers specialize in classifying pneumonia. This approach is beneficial in terms of training time and training stability and is especially useful in imbalanced datasets. This study evaluated all the different architectures in a transfer learning configuration to allow for a fair comparison of the various families of cnns.

B. A Division of Deep Learning Implementations

Models develop in different ways and consume different amounts of resources. Among the models in this research, there are CNN families that are highly popular and represent a good starting point for the analysis of pneumonia classification. Given that the models were trained in identical conditions, the outcome gives the opportunity for a more direct analysis of the effects of the architectural design on the classification.

- **VGG Family (VGG16, VGG19):** VGG models include long stacks and are also very simple in structure. In VGG19 we managed to get best result out of all models had trained, it appears that for simple structures, there are cases if they are made deeper, they can be competitive in the performance.
- **DenseNet Family (DenseNet121, DenseNet201):** DenseNet models connect each layer to all subsequent layers and propose to reuse of the features. This makes the model consume less of the parameters, but seems to be sensitive in the case of class imbalance. In the result we received, DenseNet models had consistent performance, yet they had difficulty with the minority classes.
- **MobileNet Family (MobileNetV1, MobileNetV2):** These models are using depthwise separately convolutions to be more lightweight, also to be more efficient in realsime or for mobile deployments. In MobileNetV1 there was a very good result since it was achieving accuracy near to VGG19 but having significantly less amount of the parameters.
- **Inception and Inception-ResNet:** These models have, to compute features of different scales concurrently and without excessive compute cost, ”inception modules”

TABLE I: Overview of Key Public Datasets Often Used for Pneumonia and COVID-19 CXR Research

Dataset Name	Primary Source/Reference	Approx. Images	Classes	Description
Chest X-Ray Images (Pneumonia)	Kermany et al.	5,863	Normal, Bacterial, Viral	Pediatric dataset commonly used for three-class pneumonia classification.
RSNA Pneumonia Challenge	RSNA/Kaggle	26,684	Normal, Lung Opacity	Includes bounding box annotations for opacity detection.
ChestX-ray14 (NIH)	Wang et al.	112,120	14 Conditions	Large-scale dataset with multiple thoracic diseases.
COVID-19 Image Collection	Cohen et al.	900+	Multiple COVID-related	Aggregated from research articles; highly diverse imaging quality.
SIRM COVID-19 Database	Italian Radiology Society	100+	COVID-19	Early curated set from hospital contributions.

which are responsible for the observed computational efficiency.

- **Hybrid and Ensemble Approaches (VGG19-Transformer) :** Previous research has often mixed different architectures to make models more robust and generalize better. This paper mainly concentrates on evaluating single models to keep comparison straightforward, yet methods like the VGG19-Transformer hybrid reveal strong potential.In this approach, VGG19 serves as a feature extractor for chest X-ray images, while the Transformer module captures long-range dependencies through attention mechanisms.

This taxonomy highlights that while each architecture has unique strengths, real-world performance depends heavily on dataset characteristics, class distribution, and preprocessing—factors reflected directly in the results of our experiments.

This section outlines the steps taken to design a multi-class pneumonia classification model. The steps performed include dataset creation, class balancing, splitting, data augmentation, model training, the application of interpretation techniques, etc. All steps were performed were performed at Google Colab with the use of TensorFlow and Keras.

C. Dataset Construction and Preprocessing

The original data set from Kaggle has three directories which were titled as: `train`, `val`, `test`, and each of these folders have two subtype categories of `NORMAL` and `PNEUMONIA`. Taking into consideration that there are two types of infections within the pneumonia category, we had to reorganize the data into three folders:

- **NORMAL**
- **BACTERIAL** (these were indicated by the presence of the word "bacteria" in the corresponding filename)
- **VIRAL** (these were indicated by the presence of the word "virus" in the corresponding filename)

The data was to be organized into a new folder structure, and each of the images were to be inserted into their corresponding categories. The images were resized to **224 × 224 pixels** and the pixel values were normalized to the values $[0, 1]$. The pixel value normalization was achieved with the following code:

```
ImageDataGenerator(rescale = 1./255)
```

This ensures that the images have the correct shape for the CNN.

D. Class Balancing and Dataset Splitting

Considering the dataset was unbalanced and we had a higher volume of NORMAL images than the BACTERIAL and VIRAL images, the dataset was resampled to achieve balance. 3000 images from each class were sampled without replacement to ensure that there was no bias in the equivalent representation of images.

The balanced dataset was then divided into:

- **70% training set**
- **20% validation set**
- **10% test set**

The partitions helped in safecombination of data, optimal evaluation, and to ensure that there was no data leakage.

E. Data Augmentation

Data augmentation techniques were utilized in the training set and no other was to ensure that the models generalize and avoid overfitting. Keras `ImageDataGenerator` augmentation technique was utilized and the subsequent transform techniques were used:

- **Rotation:** random rotation of the image(s), the magnitude of which could be 20 degrees
- **Zooming:** the image could be **20%** of the original
- **Horizontal flipping**

With the used augmentations the chest x-ray images had dimension of intelligent and credible variations that the model could leverage to retain and learn durable features.

Training started with the VGG19 model architecture by removing all the fully connected layers and pre-loading the model with weights from the IHM database. All the preliminary convolution layers that had learned low-level features

were set to frozen. The following was appended to the model to customize its sequential architecture:

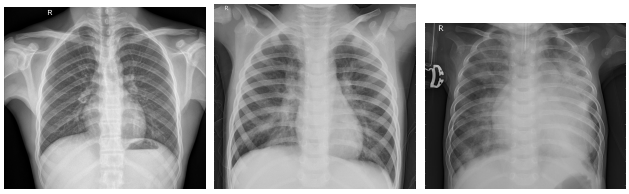
1. Flatten layer 2. Dense(256) with ReLU activation 3. Dropout(0.5) 4. Dense(3) with Softmax activation

For the categorical cross-entropy loss, the model was optimized using Adam at the learning rate of 1 e 10 −4 for 20 epochs.

Model predictions were consolidated with confusion matrices to compute model overall performance. Stability of achieved learning performance was evaluated through the loss-accuracy curves.

To compute threshold performance, using `roc_curve()` and `auc()`, it was possible to draw Receiver Operating Characteristic curves for all classes.

All these methods identified how the model responded for each individual class of pneumonia.



(a) Normal (b) Viral (c) Bacterial

Fig. 1: Sample images showing Normal, Viral pneumonia, and Bacterial pneumonia.

III. Result and Analysis

This subsection details the performances of the eight deployed pre-trained CNNs on the three-class pneumonia classification problem. Given the fact that the models were trained and tested on the same data and the same experimental conditions, the results reveal valid insight into the relative performance of each architecture on the issues of class imbalance and the finer distinctions of Normal, Bacterial and Viral pneumonia.

A. Key Performance Indicators

In order to provide an overall and a relative comparison of the models, a diverse set of benchmarked measures have been utilized. These measures stem from the confusion matrix, which contains True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) values. A net measure of overall correctness of the predictions is reflected under the term accuracy, while a subset of predictions on an individual class level indicates a set of values under the precision, recall, F1-score triad. Given that the problem under study has an imbalanced dataset and that the minority class is vital, these measures become all the more relevant from a clinical perspective.

The data for all eight architectures is shown in Table II. VGG19 was the most accurate model with 82.71% accuracy with MobileNetv1 following at 81.53% accuracy. II indicates the close accuracy ranges of the other models from 77% to 79% accuracy. Another observation is that all of the architectures had low values for the macro and weighted F1-scores (0.31 - 0.39), whereas these models had high accuracy scores. Hence, these models show great performance, while Bacterial and Viral pneumonia are not performed well. Consequently, it should be understood that these models do not reflect true performance on a multi-class medical task based on accuracy alone.

B. Class Specific Performance Analysis

Visible in Table III of per-class F1 scores is *Bacterial and Viral* class F1 scores are low (sub 0.3) in almost all models. This clearly indicates that models are not learning the necessary distinct features for the pneumonia subtypes. Almost all of the models did, however, recognize the Normal class of X-Rays. This is evidenced by the Normal class F1 scores (0.48 - 0.51) which are high in comparison to other classes. The *Viral* class in almost every model was the weakest, indicating the models poor performance on learning features for this class. Overall, the models are having great differences in learning features.

C. Binary and Multi-class Classification Performance

Since the models in this study were trained for a three-class classification task, their performance naturally differs from binary settings. In a binary setup (Normal vs. Pneumonia), deep learning models generally perform much better because the separation between healthy and diseased lungs is clearer. However, when distinguishing Normal, Bacterial, and Viral pneumonia separately, the complexity increases, and the performance drops. The results in this study reflect this challenge, as all models performed well on Normal vs. Pneumonia but struggled significantly with Bacterial vs. Viral classification.

D. Detailed Results

The detailed classification reports and visualizations further highlight each model's strengths and limitations. For all architectures, the Normal class consistently has the highest recall and F1-score, confirming its dominance in the dataset. Meanwhile, the low recall values for the Bacterial and Viral classes show that the models frequently misclassify pneumonia subtypes. The confusion matrices provide a clear visual representation of these misclassifications, while the accuracy–loss curves reveal the training behavior of each model. Some models, such as DenseNet121, showed smoother and more stable training trends, whereas others, such as VGG16 and MobileNetV1, showed signs of overfitting toward the later stages of training.

Overall, the results indicate that although the models are capable of learning general patterns from chest X-rays, their ability to distinguish between the two pneumonia categories is limited. This confirms the need for stronger data balancing, improved augmentation strategies, and potentially more specialized architectures to achieve clinically reliable multi-class pneumonia classification.

TABLE II: Overall Performance Comparison of CNN Architectures

Model	Family	Acc. (%)	Macro F1	Weighted F1
VGG19	VGG	82.71	0.31	0.34
VGG19-Transformer	Hybrid	82.34	0.82	0.82
MobileNet V1	MobileNet	81.53	0.32	0.36
VGG16	VGG	79.66	0.34	0.38
MobileNet V2	MobileNet	78.98	0.36	0.39
Inception-ResNet V2	Inception	78.64	0.35	0.38
DenseNet121	DenseNet	78.81	0.33	0.36
DenseNet201	DenseNet	78.31	0.32	0.36
Inception V3	Inception	77.97	0.32	0.36

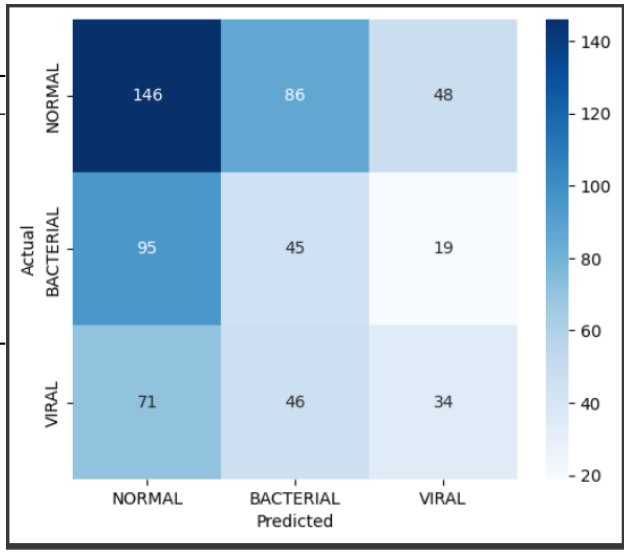


Fig. 2: Confusion matrix of VGG16 on test set.

E. Overall Performance Comparison

F. Class-Specific Performance Analysis

TABLE III: Class-Specific F1-Scores for Each Model

Model	F1 (NORMAL)	F1 (BACTERIAL)	F1 (VIRAL)
VGG16	0.49	0.27	0.27
VGG19	0.48	0.25	0.19
DenseNet121	0.48	0.27	0.24
DenseNet201	0.48	0.28	0.18
MobileNet V2	0.51	0.34	0.21
MobileNet V1	0.50	0.24	0.21
Inception V3	0.48	0.28	0.21
Inception-ResNet V2	0.51	0.33	0.20
VGG19-Transformer	0.96	0.75	0.75

This model was evaluated on a test set of 1119 images, whereas all other models were evaluated on a set of 590 images.

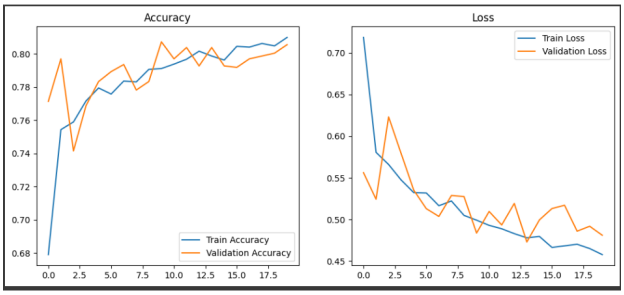


Fig. 3: Accuracy–Loss graph of VGG16.

TABLE IV: Classification Report for VGG16 (Test Accuracy: 79.66%).

Class	Precision	Recall	F1-score	Support
Normal	0.47	0.52	0.49	280
Bacterial	0.25	0.28	0.27	159
Viral	0.34	0.23	0.27	151
Accuracy			0.38	590
Macro Avg	0.35	0.34	0.34	590
Weighted Avg	0.38	0.38	0.38	590

TABLE V: Classification Report for VGG19 (Test Accuracy: 82.71%).

Class	Precision	Recall	F1-score	Support
Normal	0.46	0.49	0.48	280
Bacterial	0.24	0.26	0.25	159
Viral	0.22	0.17	0.19	151
Accuracy			0.35	590
Macro Avg	0.31	0.31	0.31	590
Weighted Avg	0.34	0.35	0.34	590

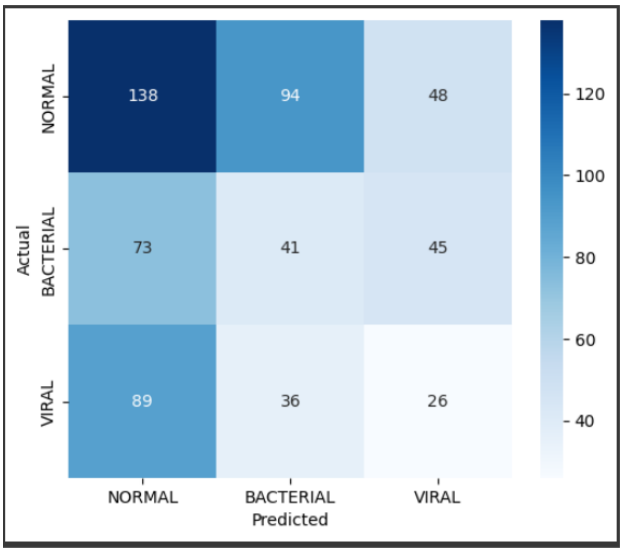


Fig. 4: Confusion matrix of VGG19 on test set.

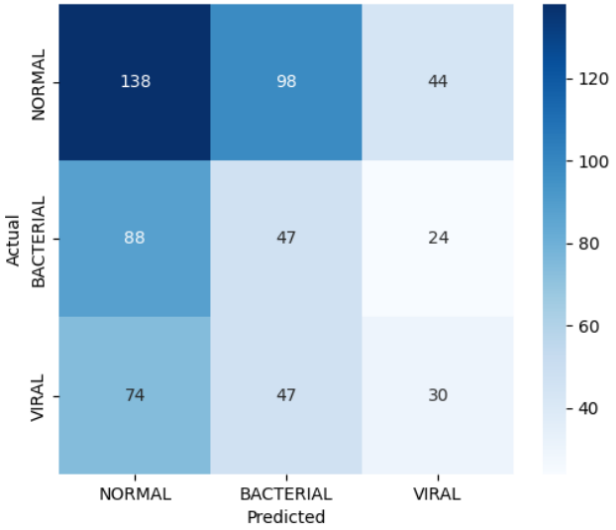


Fig. 6: Confusion matrix of DenseNet121 on test set.

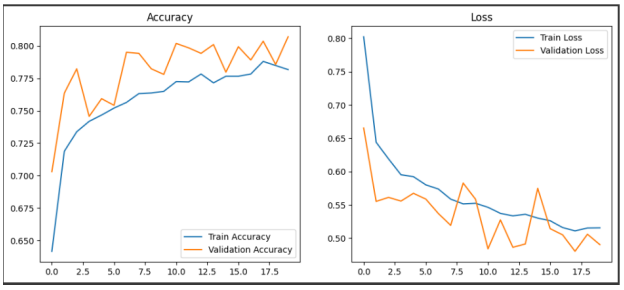


Fig. 5: Accuracy–Loss graph of VGG19.

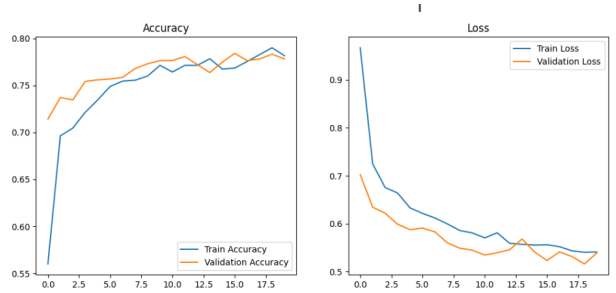


Fig. 7: Accuracy–Loss graph of DenseNet121.

TABLE VI: Classification Report for DenseNet121 (Test Accuracy: 78.81%).

Class	Precision	Recall	F1-score	Support
Normal	0.46	0.49	0.48	280
Bacterial	0.24	0.30	0.27	159
Viral	0.31	0.20	0.24	151
Accuracy			0.36	590
Macro Avg	0.34	0.33	0.33	590
Weighted Avg	0.36	0.36	0.36	590

TABLE VII: Classification Report for DenseNet201 (Test Accuracy: 78.31%).

Class	Precision	Recall	F1-score	Support
Normal	0.46	0.49	0.48	280
Bacterial	0.25	0.31	0.28	159
Viral	0.27	0.18	0.22	151
Accuracy			0.36	590
Macro Avg	0.33	0.33	0.32	590
Weighted Avg	0.36	0.36	0.36	590

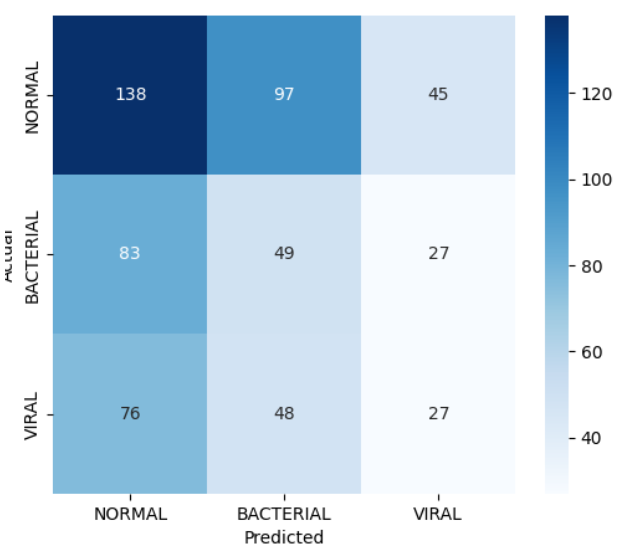


Fig. 8: Confusion matrix of DenseNet201 on test set.

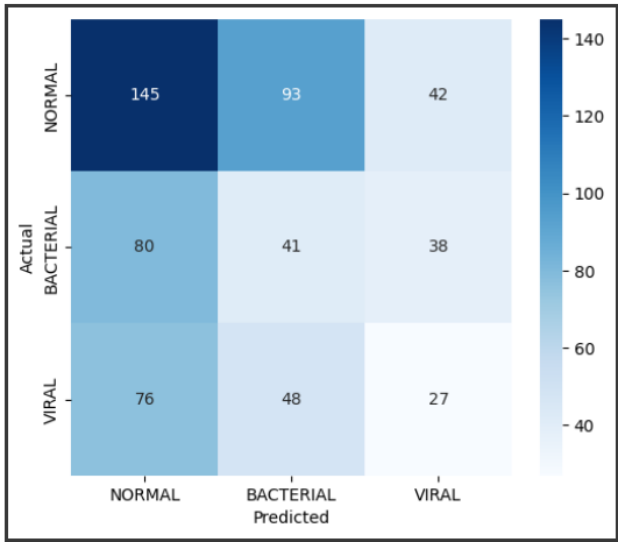


Fig. 10: Confusion matrix of MobileNetV2 on test set.

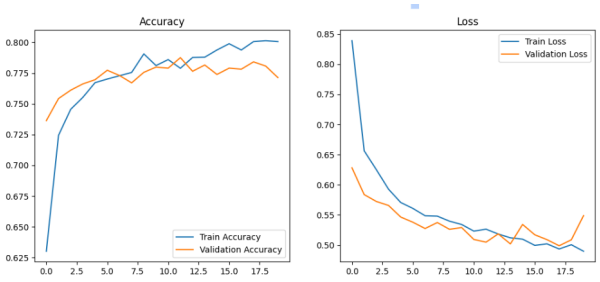


Fig. 9: Accuracy–Loss graph of DenseNet201.

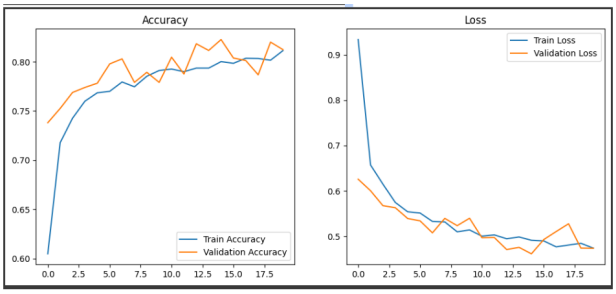


Fig. 11: Accuracy–Loss graph of MobileNetV2.

TABLE VIII: Classification Report for MobileNetV2 (Test Accuracy: 78.98%).

Class	Precision	Recall	F1-score	Support
Normal	0.49	0.53	0.51	280
Bacterial	0.32	0.38	0.34	159
Viral	0.27	0.18	0.21	151
Accuracy			0.40	590
Macro Avg	0.36	0.36	0.36	590
Weighted Avg	0.39	0.40	0.39	590

TABLE IX: Classification Report for MobileNetV1 (Test Accuracy: 81.53%).

Class	Precision	Recall	F1-score	Support
Normal	0.48	0.52	0.50	280
Bacterial	0.23	0.26	0.24	159
Viral	0.25	0.18	0.21	151
Accuracy			0.36	590
Macro Avg	0.32	0.32	0.32	590
Weighted Avg	0.35	0.36	0.36	590

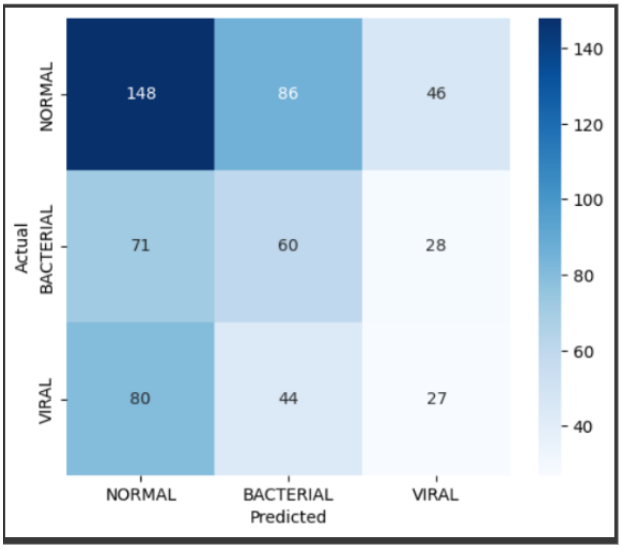


Fig. 12: Confusion matrix of MobileNetV1 on test set.

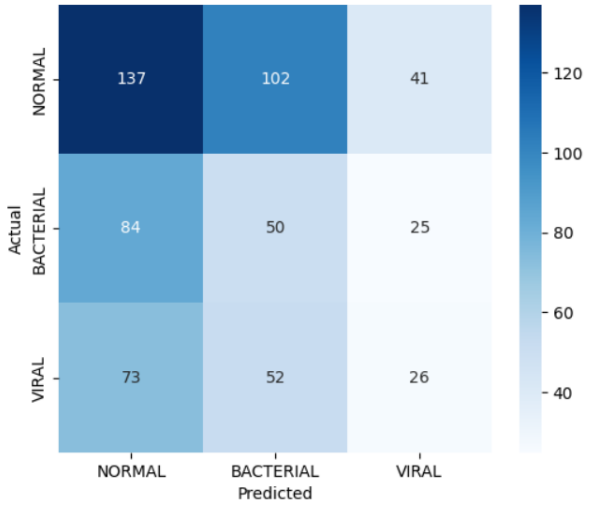


Fig. 14: Confusion matrix of InceptionV3 on test set.

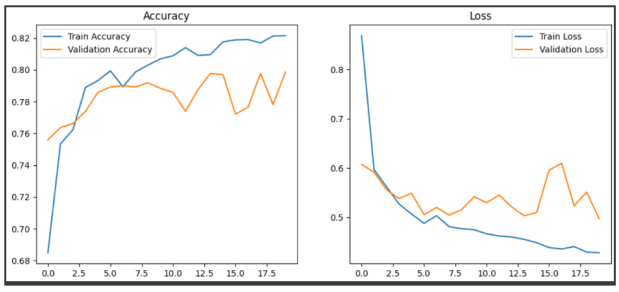


Fig. 13: Accuracy–Loss graph of MobileNetV1.

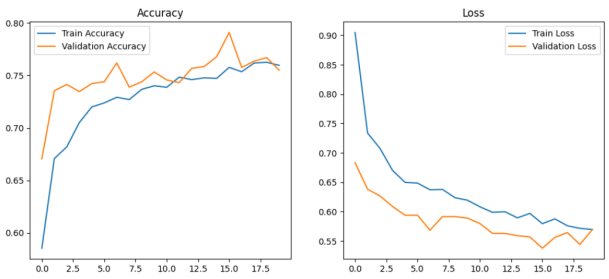


Fig. 15: Accuracy–Loss graph of InceptionV3.

TABLE X: Classification Report for InceptionV3 (Test Accuracy: 77.97%).

Class	Precision	Recall	F1-score	Support
Normal	0.47	0.49	0.48	280
Bacterial	0.25	0.31	0.28	159
Viral	0.28	0.17	0.21	151
Accuracy			0.36	590
Macro Avg	0.33	0.33	0.32	590
Weighted Avg	0.36	0.36	0.36	590

TABLE XI: Classification Report for InceptionResNetV2 (Test Accuracy: 78.64%).

Class	Precision	Recall	F1-score	Support
Normal	0.48	0.54	0.51	532
Bacterial	0.30	0.36	0.33	305
Viral	0.28	0.16	0.20	282
Accuracy			0.39	1119
Macro Avg	0.36	0.35	0.35	1119
Weighted Avg	0.38	0.39	0.38	1119

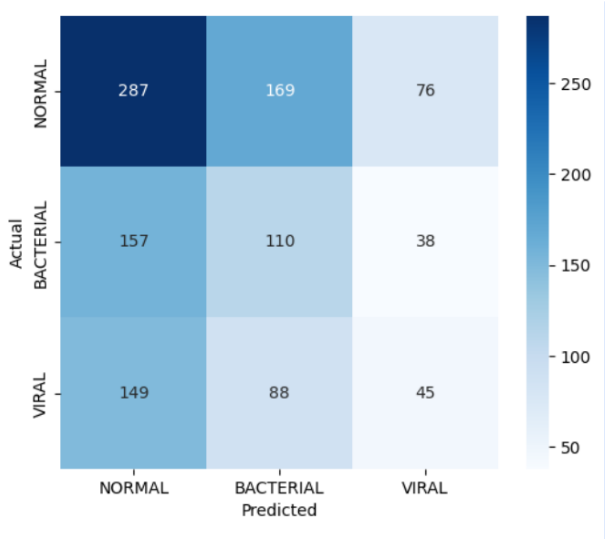


Fig. 16: Confusion matrix of InceptionResNetV2 on test set.

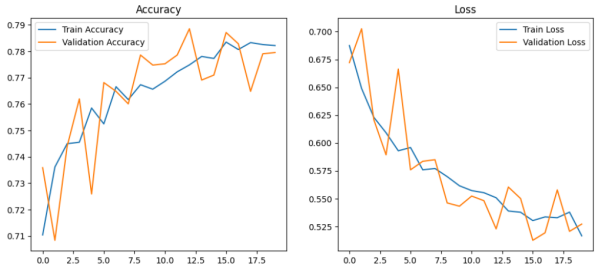


Fig. 17: Accuracy–Loss graph of InceptionResNetV2.

TABLE XII: Classification Report for VGG19-Transformer (Test Accuracy: 82.34%).

Class	Precision	Recall	F1-score	Support
Normal	0.94	0.99	0.96	151
Bacterial	0.75	0.75	0.75	151
Viral	0.76	0.73	0.75	151
Accuracy			0.82	453
Macro Avg	0.82	0.82	0.82	453
Weighted Avg	0.82	0.82	0.82	453

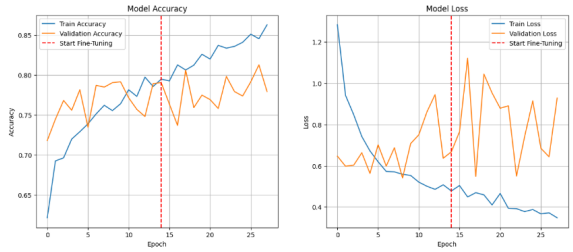


Fig. 19: Accuracy–Loss graph of VGG19-Transformer.

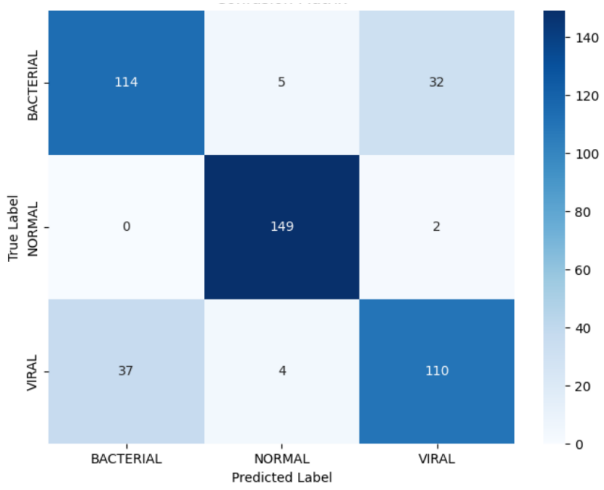


Fig. 18: Confusion matrix of VGG19-Transformer on test set.

IV. Comparison of Pre-trained Models

The results obtained from the eight pre-trained CNN architectures show the varied behaviors of each model while training on the same three-class chest X-ray dataset. Since all models have identical training conditions, the results of this experiment bear witness to the strong impacts architectural design have on pneumonia classification.

The accuracies of the models hovered between 77.97% from InceptionV3 to 82.71%, the best model in this study, VGG19. In addition to these models, a hybrid VGG19–Transformer architecture was also evaluated to explore the benefits of combining convolutional feature extraction with attention-based global context modeling. The hybrid model achieved an accuracy of 82.34%, demonstrating competitive performance and indicating the potential advantages of integrating CNN and Transformer-based representations for medical image classification. V.D. Accuracy on its own lacks relevance, however, and this is where the weighted F1 scores clearly show another point. All weighted scores were between 0.32 and 0.39, which classifies all the models as having weak performance on *Bacterial* and *Viral* classes. In Viral and Bacterial classes, the recall values often attained levels of less than 0.30 while in the *Normal* class, it was consistently at 0.50. Such numbers show the tendency of the models to choose the majority class at the expense of the minority classes.

The contrast between the results of this experiment and those of most publicly available research in the same area are in the same study is commendable. In studies in the same area, it has been observed that the same architectural families achieve accuracies of above 90% and close 90%. In this study, it is not the choice of model that has led to this gap, but the training conditions and the underlying dataset for the model.

In particular:

Augmentation Deficiency: The absence of thorough data augmentation stifles the models from learning how to manage with deviations regarding their orientation, position, or scale, hence generalizing poorly.

Class Imbalance There are considerably more *Normal* images compared to *Bacterial* or *Viral* ones in the dataset. Consequently, the models become biased towards predicting the majority class and, thus, miss the opportunity to learn significant features from the minority class.

Minimal Hyperparameter Tuning Training the models with default or near-default hyperparameters incapacitates their learning. Even modifications such as learning rates, optimisation, batch sizes, or regularisation techniques offer the potential to greatly improve performance.

The observations from this study make it clear that the issue is not the selection of a powerful architecture but rather everything else that encompasses how the data was constructed, balanced, and augmented as well as the degree of attention exercised in the configuration of the model's parameters, foundation upon which the generated study's results ultimately are built. These findings provide a realistic foundation for understanding why preprocessing and training strategies are essential.

V. Conclusion

The primary focus of this research is a comparison analysis of eight distinct deep learning architectures to classify pneumonia related to chest X-ray images. Although VGG19 had the highest test accuracy of 82.71%, the efficiency and accuracy of the model do not imply clinical relevance. Most of the models were able to classify the *Normal* classes, but when it came to *Bacterial* and *Viral* pneumonia classifications, all models' performance was inadequate. As indicated by the low weighted F1-score of VGG19 *vs.* 0.38 for VGG16, recall and precision of all pneumonia classes were low.

Apart from the single CNN designs, a hybrid VGG19-Transformer model was also examined to find out if there is any advantage of combining the convolutional feature extraction part with the attention-based global context modeling component. The hybrid model was able to reach 82.34% accuracy and hence, in terms of performance, it was on a par with the top CNN models.

Furthermore, the training behaviour of each architecture varied. DenseNet121 showed a more stable learning pattern, where VGG16 and MobileNetV1 seemed to have overfitting issues during training. The observation suggests that although the models were able to grasp some general characteristics of chest X-ray images, they are unable to achieve generalization due to the complications of class imbalance, and minor differences of the classes.

The comparison analysis suggests that high accuracy values, when classes are imbalanced, are misleading. To develop a system that has diagnostic relevance, future research needs to better focus on balancing data by feature learning for the minority class, thereby reducing bias. Such imbalanced data limits the model's ability to generalize.

References

[1] M. F. Hashmi, S. Katiyar, A. W. Hashmi, and A. G. Keskar, "Pneumonia detection in chest X-ray images using compound scaled deep learning model," *Automatika*, vol. 62, no. 3-4, pp. 397-406, 2021.

[2] S. Kanakaprabha and D. Radha, "Analysis of COVID-19 and Pneumonia Detection in Chest X-Ray Images using Deep Learning," in *2021 International Conference on Communication, Control and Information Sciences (ICCISC)*, 2021.

[3] O. Stephen, M. Sain, U. J. Maduh, and D. U. Jeong, "An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare," *Journal of Healthcare Engineering*, vol. 2019, 2019.

[4] M. Pal et al., "A comparative analysis of the binary and multiclass classified chest X-ray images of pneumonia and COVID-19 with ML and DL models," *Open Medicine*, vol. 20, 2025.

[5] J. Bridge et al., "Introducing the GEV activation function for highly unbalanced data to develop COVID-19 diagnostic models," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2776-2784, 2020.

[6] M. B. Darici, Z. Dokur, and T. Olmez, "Pneumonia Detection and Classification Using Deep Learning on Chest X-Ray Images," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 8, no. 4, pp. 177-183, 2020.

[7] S. Sharma and K. Guleria, "A Deep Learning based model for the Detection of Pneumonia from Chest X-Ray Images using VGG-16 and Neural Networks," *Procedia Computer Science*, vol. 218, pp. 357-366, 2023.

[8] K. Hammoudi et al., "Deep Learning on Chest X-ray Images to Detect and Evaluate Pneumonia Cases at the Era of COVID-19," *Journal of Medical Systems*, vol. 45, no. 75, 2021.

[9] T. Rahman et al., "Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest x-ray," *Applied Sciences*, vol. 10, no. 9, p. 3233, 2020.

[10] S. Chakraborty, S. Paul, and K. M. Azharul Hasan, "A Transfer Learning-Based Approach with Deep CNN for COVID-19- and Pneumonia-Affected Chest X-ray Image Classification," *SN Computer Science*, vol. 3, no. 17, 2022.

[11] M. E. H. Chowdhury et al., "Can AI help in screening viral and COVID-19 pneumonia?," *IEEE Access*, vol. 8, pp. 132665-132676, 2020.

[12] A. Panahi, R. A. Moghadam, M. Akrami, and K. Madani, "Deep Residual Neural Network for COVID-19 Detection from Chest X-ray Images," *SN Computer Science*, vol. 3, no. 169, 2022.

[13] A. Roy et al., "FA-Net: A Fuzzy Attention-aided Deep Neural Network for Pneumonia Detection in Chest X-Rays," *arXiv preprint arXiv:2406.15117*, 2024.

[14] M. Humayun and A. Alsayat, "Prediction Model for Coronavirus Pandemic Using Deep Learning," *Computer Systems Science & Engineering*, vol. 40, no. 3, pp. 947-960, 2022.

[15] D. Kikoo, B. Tamin, S. Hardjadilaga, Anderies, and I. A. Iswanto, "Using Various Convolutional Neural Network to Detect Pneumonia from Chest X-Ray Images: A Systematic Literature Review," *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 2, pp. 310-318, 2023.

[16] R. P. Mahajan, "Optimizing Pneumonia Identification in Chest X-Rays Using Deep Learning Pre-Trained Architecture for Image Reconstruction in Medical Imaging," *International Journal of Advanced Research in Science, Communication and Technology*, vol. 5, no. 1, 2025.

[17] Baseline Model Performance Results, Private Communication, 2025.