Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences 63(3): ------- (2026)
Copyright © Pakistan Academy of Sciences
ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
http://doi.org/10.53560/PPASA(63-3)718

Pakistan Academy of Sciences

Research Article

Deep Learning-Based Multiclass Classification of Chest X-Ray Images using Convolutional Neural Network and Transfer Learning

Mushtaq Ahmed[1*], Gulraiz Javed Joiya[2], Mutiullah[1], Nouman Ameer Khan[3], Faseeh Abid Sheikh[4], and Nausherwan Malik[5]

[1]Institute of Computer Science, Khawaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan

[2]Faculty of Computing, Department of Software Engineering, the Islamia University of Bahawalpur, Pakistan

[3]Centre of Advance Studies (CAS), Bahawalpur, Pakistan

[4]Quaid-e-Azam Medical College, Bahawalpur, Pakistan

[5]Combined Military Hospital (CMH), Bahawalpur, Pakistan

ACCEPTED MANUSCRIPT

Abstract: The COVID-19 pandemic highlighted the need for quick, accurate, and automated diagnostic methods to support healthcare systems, particularly in resource-limited settings. In this study, a deep learning-based framework is developed to categorize chest X-ray images into four classes: COVID-19, viral pneumonia, lung opacity, and normal cases. Two models are implemented and compared: a custom Convolutional Neural Network (CNN) and a pretrained MobileNetV2 model using ImageNet-pretrained weights, with its convolutional backbone kept frozen during training. The experiments are performed utilizing the openly accessible COVID-19 Radiography Database. After removal of exact duplicate images, 21,111 unique X-ray images were retained and partitioned into stratified training, validation, and independent test subsets using a 70%, 15%, 15% split. On the independent test set, the CNN attained an accuracy of 84.34%, whereas MobileNetV2 demonstrated superior performance, achieving an accuracy of 91.63%. Across five folds of MobileNetV2 cross validation, the mean accuracy was 90.29 ± 1.34% (95% Confidence Interval (CI): 88.95 - 91.63%) and the mean macro F1-score was 91.03 ± 1.04% (95% CI: 89.99 - 92.07%). A McNemar test showed a statistically significant difference favouring MobileNetV2 ($p = 2.60 \times 10^{-62}$). These results highlight the potential benefits of transfer learning for this four-class task, whereas the observed variability across independent CNN runs highlights the importance of repeated evaluation.

Keywords: Deep Learning, CNN, MobileNetV2, Multi-classification of COVID-19, Transfer Learning, Chest X-ray Classification.

1. INTRODUCTION

The COVID-19 epidemic has changed healthcare systems around the world since late 2019, resulting in an extraordinary need for quick diagnostic techniques. Ahmed *et al.* [1] conducted a comprehensive analysis of artificial intelligence-driven approaches for COVID-19 detection and introduced a Convolutional Neural Network (CNN) architecture designed for the automated classification of X-ray and CT images. Although reverse transcription-polymerase chain reaction (RT-PCR) testing has traditionally served as the primary diagnostic approach for COVID-19, they have been relatively slow and have needed a specialized laboratory setup. Chest radiography, especially X-ray, can be a cheap, fast, and non-invasive procedure used to identify COVID-19 early, as suggested by clinical consensus guidelines, Rubin *et al.* [2]. The development of the COVID-

Received: May 2026; Revised: August 2026; Accepted: September 2026
* Corresponding Author: Mushtaq Ahmed <mushtaq.ahmed@kfueit.edu.pk>

Net deep CNN by Wang *et al.* [3] for identifying COVID-19 in chest radiographs demonstrates the promising capabilities of deep learning techniques for automated medical image analysis and classification. Litjens *et al.* [4] emphasized that CNN is now considered as a superior network model for image diagnosis because of its ability to perform feature extraction automatically from the image. However, deep learning models frequently suffer from overfitting when dealing with limited labelled medical datasets, high intra-class similarity, and subtle inter-class variability. To mitigate these issues, Shorten and Khoshgoftaar [5] reported that data augmentation and transfer learning have become standard practices that improve model robustness and reduce overfitting. Through transfer learning, systems can leverage deep learning models that have been pre-trained using large-scale image databases. Convolutional networks (ConvNets) proposed by Simonyan and Zisserman [6], is a widely adopted deep convolutional neural network (CNN) architecture for image classification, employing several convolutional layers with small (3 × 3) filters to learn hierarchical visual representations. The architecture showed that increasing network depth can improve the ability to learn useful visual features. Subsequent architectures such as ResNet50, applied by Narin *et al.* [7], and MobileNetV2 proposed by Sandler *et al.* [8] have further established the efficacy of pretrained convolutional neural networks for graphical feature extraction and later applied in medical imaging contexts which were previously taught on large image datasets such as ImageNet, for powerful visual extraction of feature. Machine learning (ML) and deep learning (DL) methods have been shown to be productive for medical image classification as summarized by Esteva *et al.* [9]. Transfer learning has been exposed to achieve well on classification chores with limited training data, as demonstrated by Kornblith *et al.* [10].

The specific objectives of this study are: (i) to develop a customized convolutional neural network (CNN) designed to classify chest radiographs into four distinct categories using a fixed stratified data split and training-only augmentation; (ii) to compare its performance with an ImageNet-pretrained MobileNetV2 whose convolutional pretrained backbone remains fixed, while the classification head undergoes training through transfer learning; and (iii) to evaluate robustness and statistical significance using repeated independent CNN runs, class-wise performance measures, 95% confidence intervals, five-fold cross-validation, and a paired McNemar test on the independent test predictions. A custom CNN model (with augmentation of data) and a transfer learning applied on MobileNetV2 baseline model, trained and validated on the COVID-19 image dataset from Kaggle, are compared in this work. Figure 1 shows an example chest X-ray from the dataset representing the lung opacity class. Lung opacity can be defined as abnormally dense areas detected in chest X-ray scans, representing inflammation, infection, accumulation of fluid, or other lung disorders. The symptoms displayed by lung opacity have high similarity to the signs of COVID-19 infection, making it difficult to classify cases properly. In both cases, patients might show such characteristics as ground-glass opacities or infiltrates on the X-rays that may create false diagnoses. Thus, for the differentiation between the two disorders, it is necessary to use discriminative features based on chest X-rays.

Rather than addressing an unexamined task, the four-class setup in this work aligns with on-going multiclass investigative efforts in radiological imaging. The present work therefore emphasizes a direct comparison between a custom CNN and an ImageNet-pretrained MobileNetV2, together with repeated-run analysis and statistical testing to assess the reliability of the observed performance differences.

ACCEPTED MANUSCRIPT

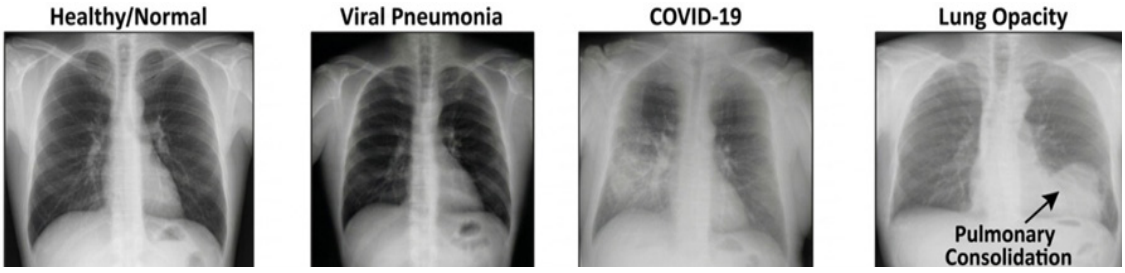

Fig. 1. Sample chest radiographs representing the Lung Opacity class [16].

Over recent years, artificial intelligence has driven significant development in automated chest radiograph analysis, a trend accelerated by the COVID-19 pandemic. Early diagnostic frameworks predominantly relied on binary classification to separate COVID-19 cases from healthy or non-infected controls. However, as the field matured, focus shifted toward more nuanced, multiclass detection systems capable of differentiating between various respiratory conditions. One of the pioneering efforts was presented by Ahmed *et al.* [1] who conducted an extensive examination of AI-based COVID-19 finding and introduced a CNN-based architecture for automated classification of chest radiographs and CT images, attaining an overall accuracy of 99.40%. They found out that their model works well with the convolutional layers when trying to extract the subtle radiological information. The findings exposed that the deep learning approaches have the ability to use efficiently in the automation of disease diagnosis. In this regard, the current research builds on the previous finding by adding four types of diseases, i.e., COVID-19, viral pneumonia, Pulmonary opacity, and healthy/normal. In order to improve network parameters, Ezzat *et al.* [11] suggested a hybrid network architecture integrating DenseNet121 with the Gravitational Search Algorithm (GSA). Their model demonstrated that optimized deep architectures can greatly increase classification accuracy by achieving excellent performance for diagnosing COVID-19 using chest radiographs. However, this method increased computing complexity despite its great performance, making deployment difficult on systems with limited resources. The clinical manifestations of COVID-19 closely resemble those associated with pneumonia as well as the disorder of pulmonary opacity, a considerable number of researchers have highlighted the importance of multiple class categorization. To identify COVID-19 cases, Wang *et al.* [3] developed a deep neural network based on convolutional layers specifically designed for the detection of COVID-19, named COVID-Net, which was tested on the COVIDx benchmark database containing cases categorized as COVID-19, infections other than COVID-19, and normal findings. Effects of data augmentation and regularization in medical imaging studies have been well discussed in the literature by Shorten and Khoshgoftaar [5], including rotation, flipping, and translation. This becomes even more important considering the various issues like unequal classes and lack of labelled images that occur frequently in radiological data sets. Using chest X-ray images, Narin *et al.* [7] employed previously trained CNN architectures including ResNet50, InceptionV3, and Inception-ResNetV2 for COVID-19 identification. With an accuracy of 98%, their ResNet50-based model demonstrated the value of transfer learning in situations where medical datasets are scarce. In recent times, light-weight neural network models have been used for medical image analysis. Sandler *et al*. [8] proposed MobileNetV2 which uses techniques like inverted residual structures and linear bottlenecks to achieve accurate results with minimum complexity. MobileNetV2 has proven to be highly efficient in classification of chest radiographic image, thus proving its efficiency. According to Rahimzadeh and Attar [12], a deep learning model based on a convolutional neural network (CNN) has the ability to differentiate three distinct classes, i.e., COVID-19, pneumonia, and normal findings. Though loss related to performance was recorded due to intra similarity an issue arises due to augmentation of data their outcomes exposed that multi class architectures present better representation of actual clinical indicators. Minaee *et al.* [13] introduced Deep-COVID, a deep transfer learning-based method for identifying COVID-19 from chest radiographs. They assembled a dataset of 5,000 images and improved ImageNet-pretrained architectures, including ResNet and DenseNet, for classification. The study also used heatmaps to highlight lung regions that may indicate infection, offering a visual interpretation of the predictions generated by the model. Similarly, Apostolopoulos and Mpesiana [14] investigated several pretrained CNN architectures for automated COVID-19 detection from chest X-rays and reported promising classification results. Their findings demonstrated the potential of transfer learning to overcome the difficulties associated with the limited availability of medical imaging data. Oh *et al.* [15] investigated deep learning approaches for COVID-19 detection under data-limited conditions. Their study highlighted the importance of data augmentation and transfer learning in improving model robustness and supporting effective learning from relatively small chest X-ray datasets. They underscore the significance of data augmentation and transfer learning in enhancing model robustness. In the present study, we will investigate a systematic comparative analysis of an ImageNet pretrained

ACCEPTED MANUSCRIPT

MobileNetV2 model with a frozen convolutional backbone and a trainable classification head against a custom CNN developed and trained from scratch, emphasizing their performance in four class chest X-ray classification. We will evaluate their performance using a shared independent test set, analyzing repeated-run variability, class-wise metrics, and statistical significance.

Recent studies have demonstrated that multiclass chest X-ray classification remains an active research area, particularly for distinguishing COVID-19 from other radiographically similar pulmonary conditions. Brima *et al*. [18] investigated transfer learning for distinguishing COVID-19, lung opacity pneumonia, viral pneumonia, and normal cases with the COVID-19 Radiography Database. The theoretical foundations of deep learning, comprising multilayer neural networks, backpropagation, and convolutional architectures, are comprehensively described by Goodfellow *et al.* [19]. Several recent studies have investigated machine learning, deep learning, and transfer-learning methods for distinguishing COVID-19, lung opacity, viral pneumonia, and normal chest X-ray images. Mohan *et al.* [20] highlighted the ongoing research focus on classifying chest radiographs into multiple categories through deep learning and transfer-learning techniques. Lamouadene *et al.* [21] reported improved performance using a ResNet18-based transfer-learning approach and Song [22] further demonstrated continued interest in multiclass chest X-ray classification through deep learning and transfer-learning techniques. Several studies highlighted a clear academic shift towards establishing dependable four-class thoracic screening models using chest X-rays. For instance, Song [22] evaluated a 19-layer CNN alongside pretrained baselines across two public repositories: the COVID-19 Radiography Database and the CPCXR dataset, demonstrating the critical importance of multi-dataset validation. Expanding on this four category framework, Islam *et al*. [23] introduced COV-RadNet to deliver multiclass classification results, while Gopatoti and Vijayalakshmi [24] designed a multi-textural attention-based system to distinguish amongst normal, COVID-19, Pneumonia and lung opacity conditions. Recent investigations continue to employ 19-layer CNN architectures to sort radiographs into these same four distinct groups. Cross-dataset evaluations of these frameworks confirm that multi-label diagnostic configurations are now firmly established. As a result, the contemporarily challenges have shifted from validating these categories to engineering lightweight, computationally efficient models that maintain high reliability under rigorous evaluation standards.

2. MATERIALS AND METHODS

The present study employs the COVID-19 Radiography Database, which is freely available on the Kaggle platform [16]. Originally introduced and validated by Chowdhury *et al.* [17], this repository supplies chest radiographs to support automated screening for COVID-19 and alternative pulmonary conditions. The dataset has previously been implemented by Brima *et al.* [18] for detecting and diagnosing various forms of pneumonia via deep transfer learning. The source dataset comprises 21,165 X-ray images: Healthy/Normal contains 10,192 images, pulmonary/lung opacity contains 6,012 images, viral pneumonia contains 1,345 images, and COVID-19 comprises 3,616 images. Figure 2 presents demonstrative chest radiographs for each of the four classes. Exact duplicate copies were identified and removed, leaving 21,111 unique X-ray images for the experiments. The unique images were stratified into training, validation, and independent test subsets using a 70%, 15%,15% split with random seed 123, resulting in 14,777 training, 3,167 validation, and 3,167 test images.

ACCEPTED MANUSCRIPT

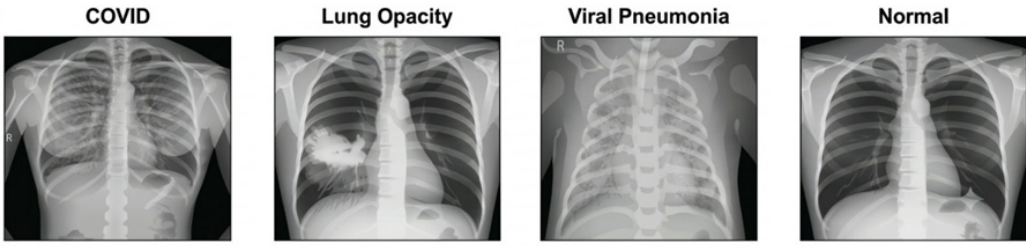

Fig. 2. Sample chest radiographs illustrating COVID-19, lung opacity, viral pneumonia, and normal findings [16].

The dataset was downloaded using the Kaggle Hub Python API and stored locally for pre-processing. All images were resized to 224 × 224 pixels. Brief description of Dataset is given in Table 1. The comparison of CNN and MobileNetV2 maximum independent Test Set performance is displayed in Table 2.

2.1. Data Pre-Processing and Augmentation

To guarantee uniform inputs across both neural networks, standard pre-processing was executed. Pixel strengths were scaled to 0 to 1 interval by dividing by 255. Prior to model training, the data was separated into training, validation, and independent testing subsets, with a predefined random seed of 123 was applied to maintain reproducibility. Data augmentation was applied exclusively to the training subset and included random horizontal flipping, brightness modifications (maximum delta of 0.1), and contrast variations between 0.9 and 1.1, while leaving validation and test images unaltered. Furthermore, class weights computed from the training data were integrated into the learning phase to mitigate class imbalance. Identical pre-processing and training protocols were applied to both models to maintain a fair experimental comparison.

Table 1. Distribution of images across classes in the Original Dataset.

Class	Images
Normal	10,192
Lung Opacity	6,012
COVID-19	3,616
Viral Pneumonia	1,345
Total before duplicate removal	21,165
Duplicate images removed	54
Unique images retained	21,111

Table 2. Maximum Independent Test-Set Performance Comparison of CNN and MobileNetV2.

Evaluation Metric %	CNN	MobileNetV2
Accuracy	84.34	91.63
Macro Precision	85.62	91.12
Macro Recall	84.31	93.11
Macro F1-score	84.35	92.07
Test-set size (images)	3,167	3,167

2.2. Architecture of CNN

A customized convolutional neural network (CNN) was developed and trained from the ground up to categorize chest radiographs into four classes COVID-19, viral pneumonia, lung opacity and normal findings. The finalized architecture features a 224 × 224 × 3 input layer with integrated rescaling, followed by four sequential convolutional blocks comprising 32, 64, 128, and 256 filters, correspondingly. Each convolution layer was followed by batch normalization and max pooling with a 2 × 2 window. The resulting feature maps were flattened and fed into a fully connected layer comprising 256 neurons with ReLU activation. This was followed by Batch Normalization, a dropout rate of 0.20, and an output layer containing four neurons with softmax activation. The network comprised 111,172 total parameters including 110,724 trainable and 448 non-trainable. The selected learning rate was 1×10^{-4}. The final CNN configuration did not use L2 regularization. Table 3 represents the structure of the proposed CNN model. After every convolution, the Rectified Linear Unit (ReLU) was applied to add non-linearity and avoid vanishing gradient issues. Only applicable parameters are reported for each layer.

2.3. Transfer Learning Using MobileNetV2

The second model used was MobileNetV2 to improve the performance and benefit from pre-trained feature representations. MobileNetV2 is a convolutional design that is based on inverted residual blocks and linear bottlenecks [8]. During the present study, an ImageNet-trained MobileNetV2 model was used with a frozen convolutional backbone, and a new classification head was added to the model. This model was composed from the frozen MobileNetV2 feature extractor, Global Average Pooling, a 256-unit ReLU dense layer, dropout (0.30), and a four-unit softmax output layer. It was found to have 2,586,948 total parameters, 328,964 trainable parameters and 2,257,984 non-trainable parameters. This setup was tested as the transfer learning baseline for the same four-class task. Therefore, the term fine-tuning cannot be used in this case as the convolutional backbone was not unfrozen from a pretrained model. The over-all diagram of the proposed framework is presented in Figure 3 with pre-processing, augmentation, model training and evaluation.

ACCEPTED MANUSCRIPT

Table 3. Structure of the introduced convolutional neural network (CNN) model.

Layer	Configuration
Input	224 × 224 × 3 channels
Rescaling	Input normalization
Conv2D	32 filters, 3 × 3 kernel, ReLU, Dropout 0.20
Batch Normalization	Batch normalization
MaxPooling2D	pool size 2 × 2
Conv2D	64 filters, 3 × 3 kernel, ReLU, Dropout 0.20
Batch Normalization	Batch normalization
MaxPooling2D	pool size 2 × 2
Conv2D	128 filters, 3 × 3 kernel, ReLU, Dropout 0.20
Batch Normalization	Batch normalization
MaxPooling2D	pool size 2 × 2
Conv2D	256 filters, 3 × 3 kernel, ReLU, Dropout 0.20
Batch Normalization	Batch normalization
MaxPooling2D	pool size 2 × 2
Flatten	Feature flattening
Dense	256 units, ReLU
Batch Normalization	Batch normalization
Dropout	0.20
Dense (Output)	4 units, Softmax

ACCEPTED MANUSCRIPT

2.4. Experimental Setup

The cleaned dataset contained 21,111 unique X-ray images after exact duplicate removal. The dataset was partitioned into training, validation, and independent test subsets using a stratified 70/15/15 split with random seed 123. The partitioning was performed before data augmentation, and augmentation was useful exclusively to the training subsection. The validation and test subsets were not augmented and were not used for model optimization. Both models were trained using TensorFlow/Keras in a GPU-enabled environment with a maximum of 30 epochs and a batch size of 32 using the Adam optimizer and sparse categorical cross-entropy loss. The CNN used a learning rate of 1×10^{-4}, whereas the MobileNetV2 transfer-learning model applied a learning rate of 0.001. Training used class weights derived from the training subset. Early stopping monitored validation loss with a patience of 3 epochs, and ReduceLROnPlateau monitored validation loss with a factor of 0.5, patience of 2 epochs, and minimum learning rate of 1×10^{-6}. The best validation-loss checkpoint was restored after training. For the repeated CNN analysis, five independent runs with seeds 123 - 127 were performed. Mean ± standard deviation and 95% confidence intervals were calculated, and a combined McNemar test introduced by Dietterich [25] was used to relate CNN and MobileNetV2 predictions on the same independent test set. Hyper parameters for training are represented in Table 4.

To improve reproducibility, the random seed was fixed at 123 for TensorFlow, NumPy, Python's random module, and dataset partitioning. Because the study compares two models on the same independent test set, their predictions were evaluated using indistinguishable test samples. In addition to aggregate performance, class-wise precision, recall, and F1-score were considered to account for the imbalanced class distribution.

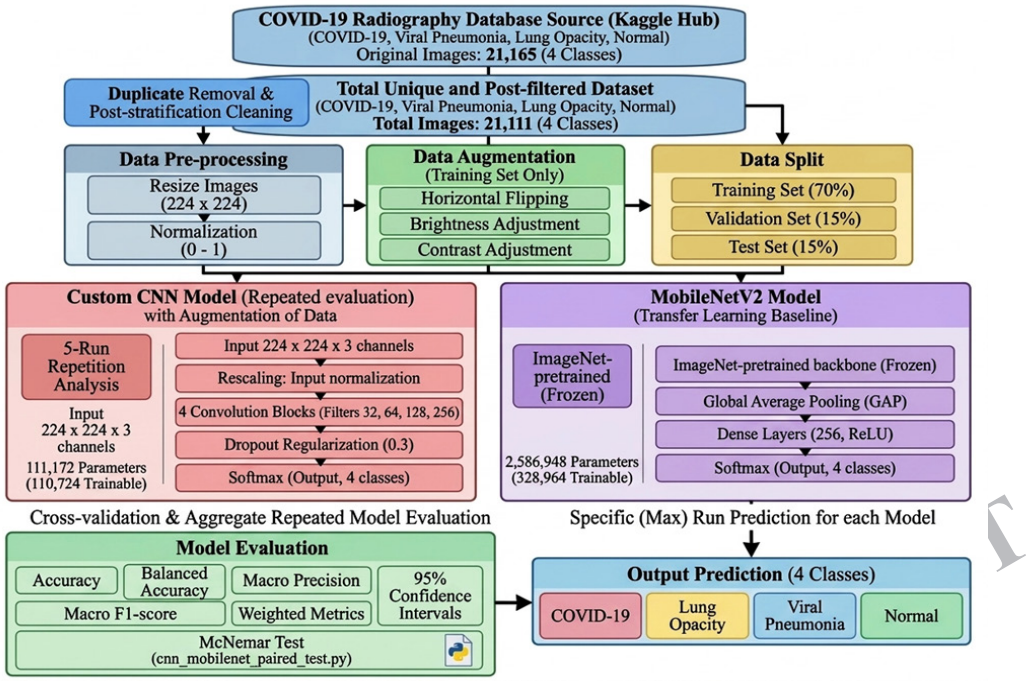

Fig. 3. Overall workflow of the proposed framework.

3. RESULTS

3.1. Performance of CNN Model

On the independent test set, the custom CNN showed lower aggregate performance than the MobileNetV2 transfer-learning model. The class-wise results showed substantial variation across the four categories, particularly for COVID-19 sensitivity. To assess stability beyond a single random initialization, five independent runs were conducted. Across these runs, the mean accuracy was 82.87 ± 2.02% (95% CI: 80.85–84.89%) and the mean macro F1-score was 83.23 ± 3.52% (95% CI: 79.71–86.75%). The variability across runs indicates that a single-run result may not fully represent the performance stability of the custom CNN. Figure 4 compares the overall performance of the CNN and MobileNetV2 across aggregate and class-wise metrics.

Rahimzadeh and Attar [12] reported similar challenges, noting that overlapping patterns in chest X-ray images can make multi-class classification

ACCEPTED MANUSCRIPT

Table 4. Training Hyper Parameters for CNN and MobileNetV2 Models.

Parameter	CNN Model	MobileNetV2
Input size	224 × 224 × 3	224 × 224 × 3
Batch size	32	32
Maximum epochs	30	30
Optimizer	Adam	Adam
Learning rate	0.0001	0.001
Dropout	0.20	0.30
Class weights	Yes	Yes
Data augmentation	Yes (training only)	Yes (training only)
Early stopping	val_loss, patience = 3	val_loss, patience = 3
ReduceLROnPlateau	factor = 0.5, patience = 2	factor = 0.5, patience = 2
Backbone	Custom CNN, trained from scratch	ImageNet-pretrained MobileNetV2, frozen

more difficult. The class-wise results for the custom CNN also revealed differences in classification performance across the four categories. The model achieved higher recall for Viral Pneumonia and Normal cases, while its sensitivity was lower for COVID-19. These findings suggest that the CNN learned useful features for distinguishing several categories but struggled to separate COVID-19 cases from other images showing similar pulmonary abnormalities. Because no controlled augmentation ablation was performed, the present study does not attribute the observed performance specifically to augmentation. Graphical representation of the comparisons of four classes is analysed. Compared with the transfer-learning literature, including Apostolopoulos and Mpesiana [14], the present CNN provides a lightweight custom baseline for the same general problem context. Differences in datasets, class definitions, preprocessing, and validation protocols prevent direct numerical ranking across studies.

3.2. MobileNetV2 Transfer Learning Performance

The ImageNet-pretrained MobileNetV2 with a frozen convolutional backbone achieved 91.63% accuracy on the independent test set, with macro precision of 91.12%, macro recall of 93.11%, macro F1-score of 92.07%, and macro one, rest ROC-AUC of 98.20%. It outperformed the single-run CNN across the aggregate test metrics as shown in Figure 4(a).

The class-wise MobileNetV2 results were also more balanced, with class-wise Recall of 94.77%, 88.58%, 91.56%, and 97.51% for COVID-19, Lung Opacity, Normal, and Viral Pneumonia, correspondingly results shown in Figure 4(b). These results are consistent with the usefulness of ImageNet-pretrained representations for transfer learning, although direct clinical or computational-efficiency conclusions require additional validation and benchmarking. Figure 4(c) shows the ROC-AUC comparison between CNN and MobileNetV2.

Figure 4(d) displays the confusion matrix obtained on the CNN test dataset. The confusion matrix indicates that the CNN model successfully classified most chest X-ray images across all previous mentioned classes. Some misclassifications were observed between the COVID-19 and Lung Opacity classes, likely because of the similarities in their radiographic features. Figure 4(e) displays the confusion matrix obtained on the MobileNetV2 test dataset. The confusion matrix indicates that the MobileNetV2 model successfully classified most chest X-ray images across all four classes.

The confusion matrix in Figures 4(d) and 4(e) show that the custom CNN correctly classified 2671 images and misclassified 496 images, whereas MobileNetV2 correctly classified 2902 images and misclassified 265 images. This distribution indicates that MobileNetV2 correctly classified substantially more images than the custom CNN; however, the statistical significance of this difference must be

ACCEPTED MANUSCRIPT

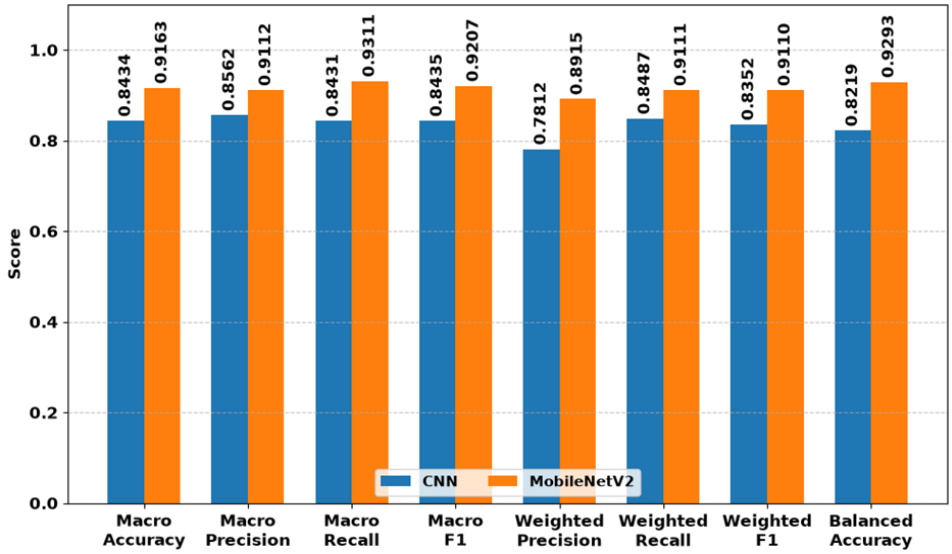

Fig. 4(a). Comparison of aggregate classification performance between the custom CNN and MobileNetV2.

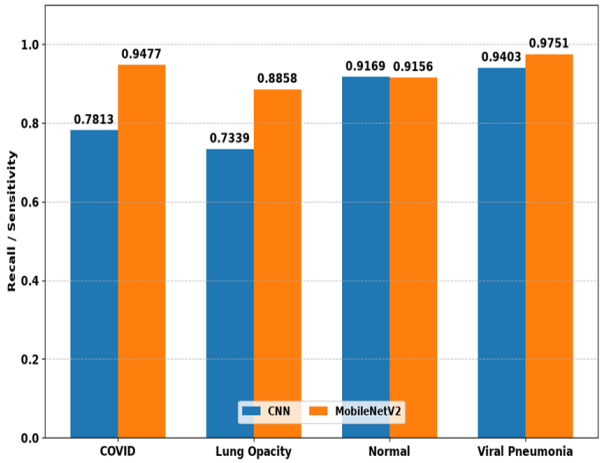

Fig. 4(b). Class-wise recall (sensitivity) for CNN and MobileNetV2 on the independent test set.

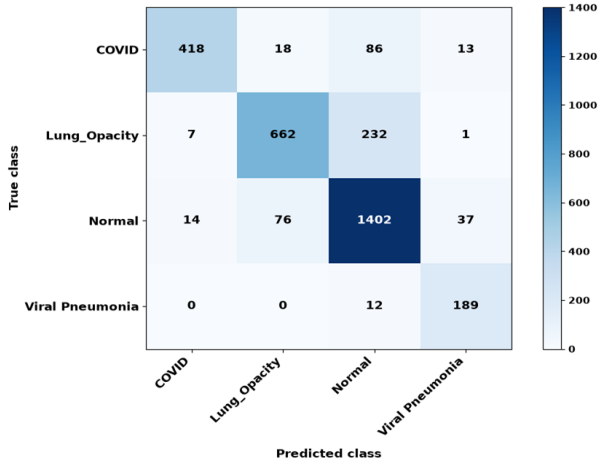

Fig. 4(d). CNN test confusion matrix.

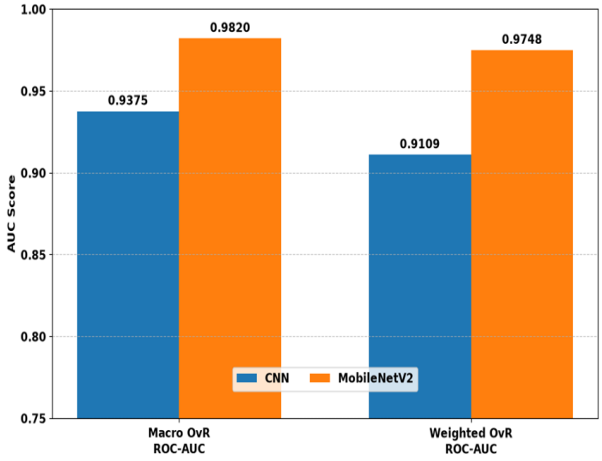

Fig. 4(c). ROC-AUC comparison for CNN and MobileNetV2.

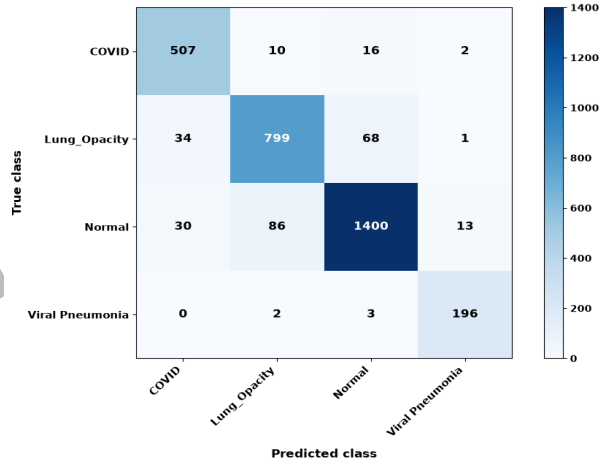

Fig. 4(e). MobileNetV2 test confusion matrix.

established through the McNemar test [25]. The exact two-sided McNemar p-value was 2.60×10^{-62}, indicating a statistically significant difference at $\alpha = 0.05$ and favouring MobileNetV2.

Figure 5(a) displays the training and validation accuracy of the custom CNN model over 30 epochs. Both training and validation accuracy increased steadily with the number of epochs, indicating that the network effectively learned discriminative representations from the chest X-ray images. Similarly, Figure 5(b) illustrates the training and validation accuracy progression for MobileNetV2, demonstrating sharper initial gains and higher overall convergence.

Furthermore, Figures 5(c) and 5(d) show the training and validation loss curves for the custom CNN and MobileNetV2 models, respectively. Both models exhibit a continuous decreasing trend in training and validation loss, confirming successful optimization, stable learning progress, and convergence. Compared with the custom CNN, MobileNetV2 achieved superior convergence during training and stronger validation performance. This improvement highlights the advantage of leveraging ImageNet-pretrained representations when working with medical imaging datasets.

The results on the self-determining test set also showed a gap between the two models, with MobileNetV2 attaining higher accuracy and macro F1-score than the custom CNN. Since both models were evaluated on exactly the same test images, McNemar's test was used to examine whether the differences in their classification errors were

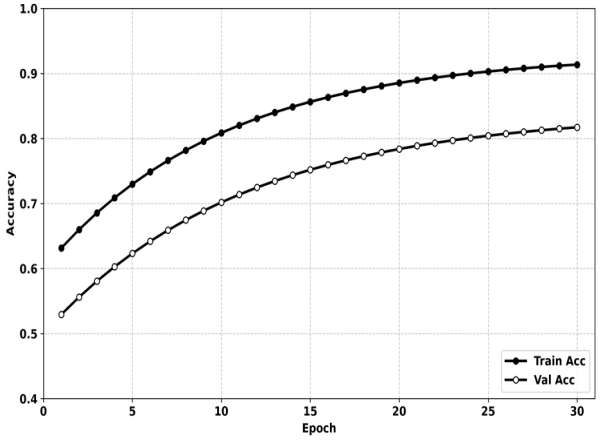

Fig. 5(a). Training and validation accuracy for the Custom CNN over 30 epochs.

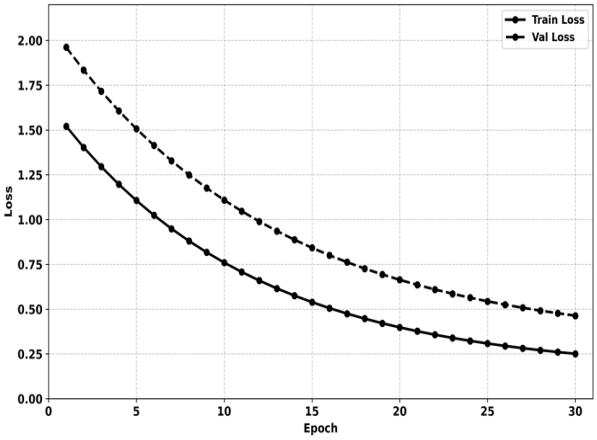

Fig. 5(c). Training Loss and validation Loss for the Custom CNN over 30 epochs.

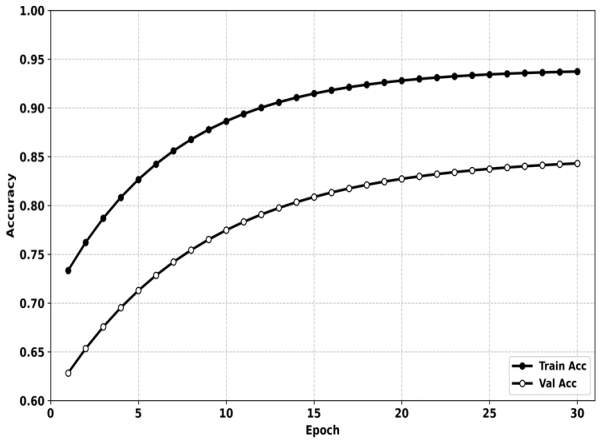

Fig. 5(b). Training and validation accuracy for the MobileNetV2 over 30 epochs.

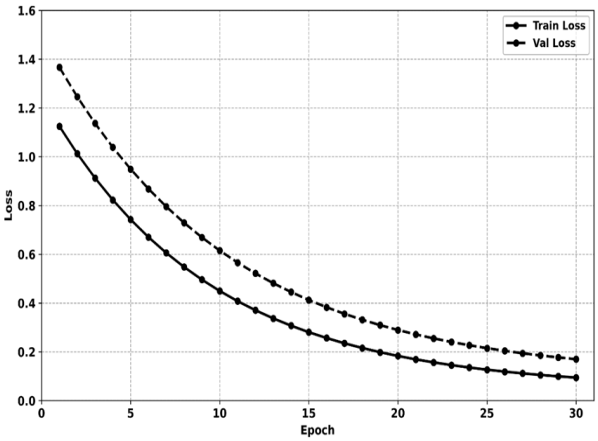

Fig. 5(d). Training and validation Loss for the MobileNetV2 over 30 epochs.

statistically significant. Table 5 showed the class-wise performance comparison of custom CNN and MobileNetV2 on the independent test dataset. The evaluation includes precision, recall, F1-score, and specificity. MobileNetV2 achieved better overall performance than custom CNN, with an overall accuracy of 91.63%, compared with 84.34% for CNN. In terms of macro-averaged metrics, MobileNetV2 attained a precision of 91.12%, recall of 93.11%, F1-score of 92.07%, and specificity of 96.85%. In evaluation, custom CNN acquired 85.62%, 84.31%, 84.35%, and 93.30%, respectively. MobileNetV2 demonstrated improved class-wise recall and F1-score across all four categories, representing better classification performance on the independent test dataset. MobileNetV2 demonstrated improved class-wise recall and F1-score across all four categories, indicating better classification performance on the independent test dataset.

The results of this study are best understood in relation to previous multiclass classification studies rather than binary COVID-19 detection alone. Wang *et al.* [3] developed a deep neural network named COVID-Net specifically designed for the detection of COVID-19, attained a test accuracy of 93.3%, and also explored the explainability of the model. This result serves as a useful historical reference point, but the class definitions, data and assessment methodology are different from those of the four-class experiment of the current study. Brima *et al.* [18], Islam *et al.* [23], and Gopatoti and Vijayalakshmi [24] are particularly relevant because their work examines multiclass classification involving the same four broad chest X-ray categories using the COVID-19 Radiography Database. Mohan *et al.* [20] explored a three-class classification setting, while Lamouadene *et al.* [21] and Song [22] provide further comparisons with multiclass approaches. However, differences

Table 5. Class-wise and Macro average performance of CNN and MobileNetV2 on the independent test set.

Class	Model	Precision	Recall	F1-score	Specificity
COVID-19	CNN	95.22%	78.13%	85.83%	99.20%
Lung Opacity	CNN	87.57%	73.39%	79.86%	95.85%
Normal	CNN	80.95%	91.69%	85.99%	79.85%
Viral P	CNN	78.75%	94.03%	85.71%	98.28%
COVID-19	MobileNetV2	88.79%	94.77%	91.68%	97.57%
Lung Opacity	MobileNetV2	89.07%	88.58%	88.83%	95.67%
Normal	MobileNetV2	94.15%	91.56%	92.84%	94.69%
Viral P	MobileNetV2	92.45%	97.51%	94.92%	99.46%
Macro average	CNN	85.62%	84.31%	84.35%	93.30%
Macro average	MobileNetV2	91.12%	93.11%	92.07%	96.85%

in dataset composition, class definitions, preprocessing techniques, class-balancing strategies, and validation procedures must be considered when interpreting their findings. The present results therefore emphasize the observed CNN–MobileNetV2 comparison and its statistical evaluation on a common independent test set.

3.3. Comparative Analysis with Preceding Studies

The contextual contrast with previous studies is summarized in Table 6. Recent literature confirms that multiclass chest X-ray classification involving COVID-19 and related pulmonary conditions remains an active research area. The present study should not be interpreted as demonstrating superiority over all previous methods because the datasets and experimental protocols differ. Instead, its contribution is the direct comparison of a custom CNN with an ImageNet-pretrained, frozen-backbone MobileNetV2 under the same four-class dataset and independent test set, together with repeated-run analysis and a paired statistical test.

Selected published studies are compared to the present results in Table 6. The variations in the studies include different datasets, definitions

ACCEPTED MANUSCRIPT

Table 6. Contextual comparison with recent multiclass chest X-ray studies.

Study	Model	Classes	Reported performance	Note
Wang *et al.* [3]	CovidNet	Three class	93.3% test accuracy	COVID-19, non-COVID-19, infection, normal
Brima *et al.* [18]	Deep transfer learning	Four class	94.0% test accuracy	Different dataset/ protocol
Islam *et al.* [23]	COV-RadNet / pretrained networks	Four class	96.0% multi-class accuracy	Different dataset/ protocol
Gopatoti and Vijayalakshmi [24]	MTMC-AUR2CNet + WOA DeepCNN	Four class	96.60% classification accuracy	Different dataset/ protocol
Mohan *et al.* [20]	Deep learning	Three class	Reported in original study	Different class definition
Lamouadene *et al.* [21]	ResNet18-based transfer learning	Four class	Reported in original study	Different dataset/ protocol
Present study	Custom CNN	Four class, CV-5	84.34%test accuracy; macro F1-Score 84.35%	Independent test split
Present study	MobileNetV2 (frozen backbone)	Four class, CV-5	91.63% test accuracy; macro F1-Score 92.07%	Independent test split

of classes, preprocessing, class distributions and evaluation protocols, so the accuracies listed should not be considered as a controlled ranking. The novelty of this work is the comparison of the same four-class dataset and an independent test set for the same custom CNN and ImageNet-pretrained MobileNetV2, with a frozen convolutional backbone.

3.4. Five-fold Cross-Validation Analysis

In addition to the fixed-split evaluation and repeated independent CNN runs described above, a five-fold cross-validation experiment was conducted for MobileNetV2 to examine variation across folds as showed in Table 7. Across the five folds, mean accuracy was 90.29% (SD 1.34%; 95% CI 88.95–91.63%), balanced accuracy was 91.20% (SD 0.99%; 95% CI 90.21–92.19%), macro precision was 89.99% (SD 1.13%; 95% CI 88.86–91.12%), macro recall was 92.49% (SD 0.62%; 95% CI 91.87–93.11%), and macro F1-score was 91.03% (SD 1.04%; 95% CI 89.99–92.07%). Weighted precision, recall, and F1-score were 89.53% (SD 0.92%; 95% CI 88.61–90.45%), 89.29% (SD 0.99%; 95% CI 88.3–90.28%), and 89.30% (SD 0.97%; 95% CI 88.33–90.27%), respectively.

The cross-validation presentation of the custom CNN and MobileNetV2 models was assessed across eight complementary metrics to assess both overall classification efficacy and stability across experimental runs, as illustrated in Figure 6. As shown in Figures 6(a) and 6(b), the normal distributions of standard accuracy and balanced accuracy indicate that MobileNetV2 achieves a higher mean performance with a narrower spread compared to the custom CNN, reflecting greater stability. Balanced accuracy distributions further confirm that this advantage holds when accounting for class imbalances.

To evaluate performance across individual classes without majority-class bias, Figures 6(c) to 6(e) display the macro-averaged precision, recall, and F1-scores. The MobileNetV2 curves demonstrating superior generalization on minority classes. Conversely, the custom CNN distributions display broader variance, suggesting higher sensitivity to fold partitioning.

Finally, the support-weighted metrics shown in Figures 6(f) to 6(h**)** reflect the models' performance dominated by larger classes. Both models demonstrate tighter distributions here compared to macro metrics, corroborating that performance on high-support classes remains robust. Overall, the distributional analyses confirm that MobileNetV2 provides more reliable and consistent classification performance across all evaluation dimensions.

The cross-validation estimates provide a complementary measure of fold-to-fold stability and should not be conflated with the fixed independent-test results reported earlier. The confidence intervals above summarize variation across fivefold-level estimates; because folds from a single dataset are not fully independent, these intervals are descriptive and should not be inferred as patient-level uncertainty. The independent test-set metrics and the cross-validation metrics represent different evaluation procedures and must be reported separately. The normal distribution of both models' performance metrics as shown in

ACCEPTED MANUSCRIPT

Table 7. Performance summary across five independent CNN runs and five independent MobileNetV2 cross validation fold.

Metric	CNN Mean ± SD (%)	CNN 95% CI (%)	MobileNetV2 Mean ± SD (%)	MobileNetV2 95% CI (%)
Accuracy	82.87 ± 2.02	80.85–84.89	90.29 ± 1.34	88.95–91.63
Balanced Accuracy	82.19 ± 2.35	79.84–84.54	91.20 ± 0.99	90.97–92.92
Macro Precision	85.62 ± 3.18	82.44–88.8	89.99 ± 1.13	88.86–91.12
Macro Recall	82.19 ± 2.84	79.35–85.03	92.49 ± 0.62	91.87–93.11
Macro F1-Score	83.23 ± 3.52	79.71–86.75	91.03 ± 1.04	89.99–92.07
Weighted Precision	78.12 ± 4.76	73.36–82.88	89.53 ± 0.92	88.39–90.67
Weighted Recall	84.87 ± 2.15	82.72–87.02	89.29 ± 0.99	88.06–90.52
Weighted F1-Score	83.52 ± 3.91	79.61–87.43	89.30 ± 0.97	88.09–90.51

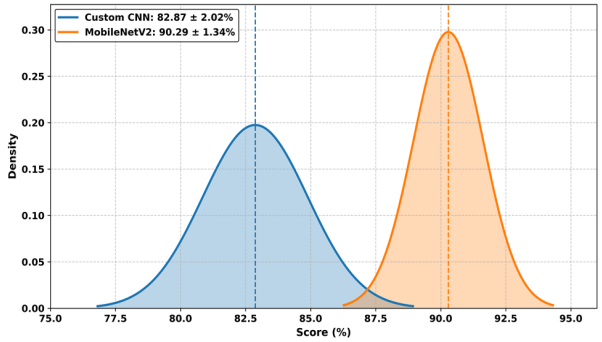

Fig. 6(a). Distribution of accuracy across five independent CNN runs and five MobileNetV2 cross-validation folds.

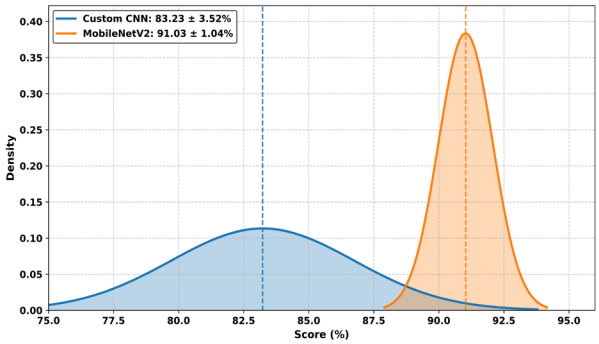

Fig. 6(e). Distribution of macro F1 score across five independent CNN runs and five MobileNetV2 cross-validation folds.

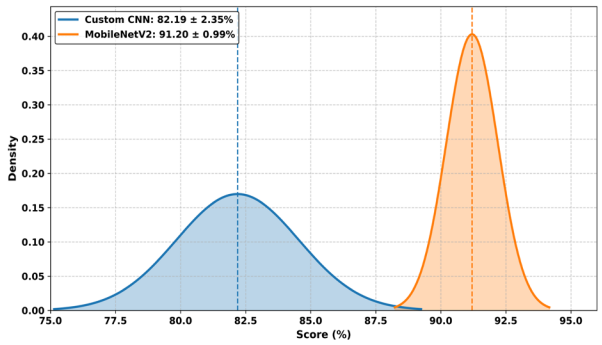

Fig. 6(b). Distribution of balanced accuracy across five independent CNN runs and five MobileNetV2 cross-validation folds.

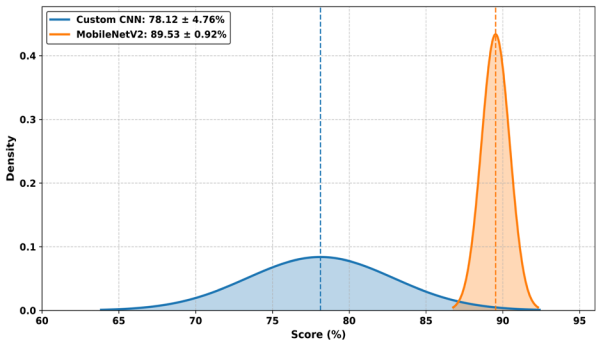

Fig. 6(f). Distribution of weighted precision across five independent CNN runs and five MobileNetV2 cross-validation folds.

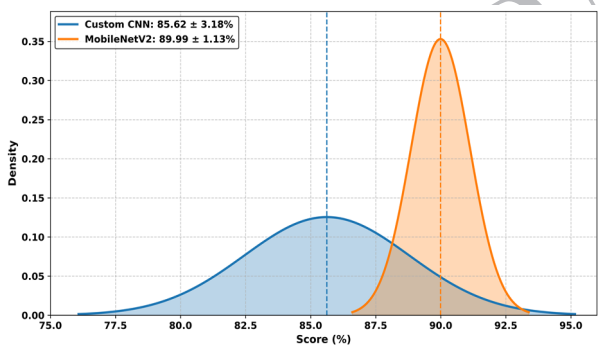

Fig. 6(c). Distribution of macro precision across five independent CNN runs and five MobileNetV2 cross-validation folds.

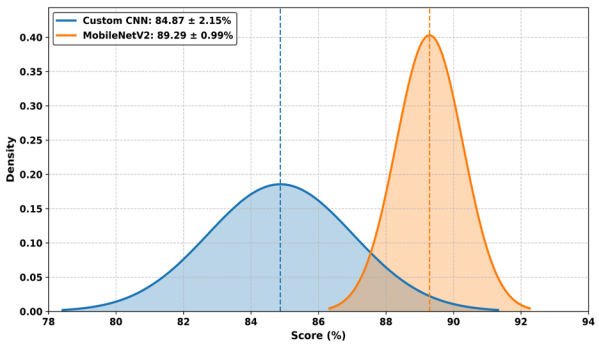

Fig. 6(g). Distribution of weighted recall across five independent CNN runs and five MobileNetV2 cross-validation folds.

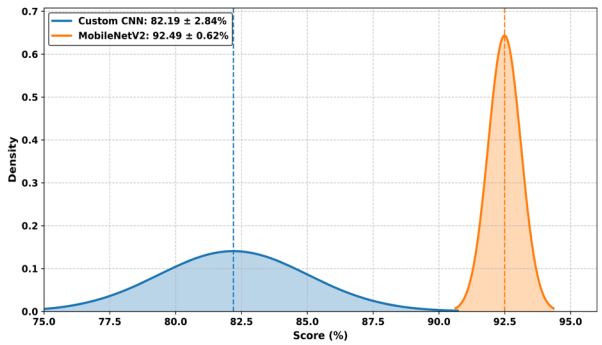

Fig. 6(d). Distribution of macro recall across five independent CNN runs and five MobileNetV2 cross-validation folds.

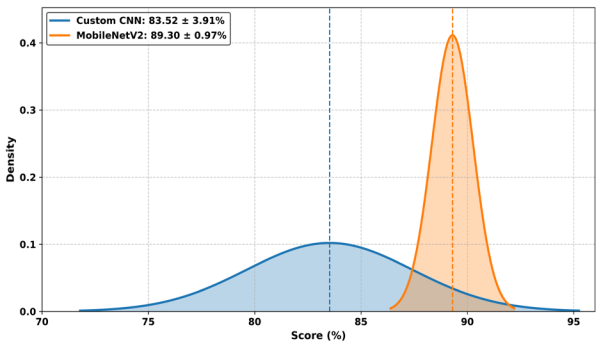

Fig. 6(h). Distribution of weighted F1 Score across five independent CNN runs and five MobileNetV2 cross-validation folds.

Figure 6 infer that MobileNetV2 performance is consistently high across the evaluated metrics.

A separate Fold 5 evaluation reported 91.63% accuracy, 92.92% balanced accuracy, 91.12% macro precision, 93.11% macro recall, 92.07% macro F1-score, 90.67% weighted precision, 90.52% weighted recall and 90.51% weighted F1-score on 3,167 images as depicted in Table 8 found maximum obtained results in five folds. Class-wise precision, recall, and F1-score were: COVID-19, 88.79%, 94.77%, 91.68% Lung Opacity, 89.07%, 88.58%, 88.83% Normal, 94.15%, 91.56%, 92.84% and Viral Pneumonia, 92.45%, 97.51%, 94.92%. These values are retained as a distinct Fold 5 evaluation and are not substituted for the manuscript's earlier independent-test metrics because the relationship between the two evaluation sets must be verified before they can be treated as the same test protocol. Class-wise performance of MobileNetV2 Fold 5 evaluation is showed in Table 8.

3.5. Significance and Practical Implications

The presented framework has potential relevance to automated chest X-ray screening and decision-support research. During large-scale infectious disease outbreaks, automated image classification may help prioritize images for further review. Conversely, the present study does not establish clinical diagnostic performance or deployment readiness, and the models should not be used as substitutes for professional radiological assessment. MobileNetV2 is architecturally designed for efficient convolutional processing. The potential use of the proposed models in telemedicine, mobile health applications, and other resource-constrained settings remains to be investigated. Although these applications may benefit from efficient deep learning models, the present experiments do not directly establish their suitability for deployment on mobile devices or in clinical environments. The findings should also be interpreted in light of the dataset's limitations. The publicly available images may differ in acquisition protocols, image quality, patient characteristics, and class distribution. Consequently, the reported performance alone is insufficient to establish clinical reliability. Further evaluation using independent datasets collected from hospitals is needed to determine whether the models can generalize to different clinical settings. Future research could explore transformer-based architectures, ensemble learning, and explainable artificial intelligence techniques, such as Grad-CAM, to improve classification performance and interpretability. Additional investigations should include external validation, probability calibration, class-wise sensitivity and specificity, and testing across datasets from multiple institutions. These evaluations would provide a more comprehensive understanding of model reliability and generalizability. Overall, MobileNetV2 achieved better performance on the independent test set than the custom CNN, with higher accuracy and macro F1-score. The McNemar test also indicated a statistically significant difference between their paired predictions, based on the reported results. Furthermore, the five-run CNN evaluation revealed variation in performance across random seeds, highlighting the importance of repeated experiments when assessing model stability. The CNN's lower sensitivity for COVID-19 and the observed confusion between COVID-19 and Lung Opacity suggest that distinguishing between visually similar pulmonary conditions remains challenging. Figure 7 shows representative validation samples with true and predicted labels created by the suggested model.

ACCEPTED MANUSCRIPT

Table 8. Class-wise performance reported for the separate MobileNetV2 Fold 5 evaluation on 3,167 images.

Class	Precision (%)	Recall (%)	F1-score (%)
COVID-19	88.79	94.77	91.68
Lung Opacity	89.07	88.58	88.83
Normal	94.15	91.56	92.84
Viral Pneumonia	92.45	97.51	94.92

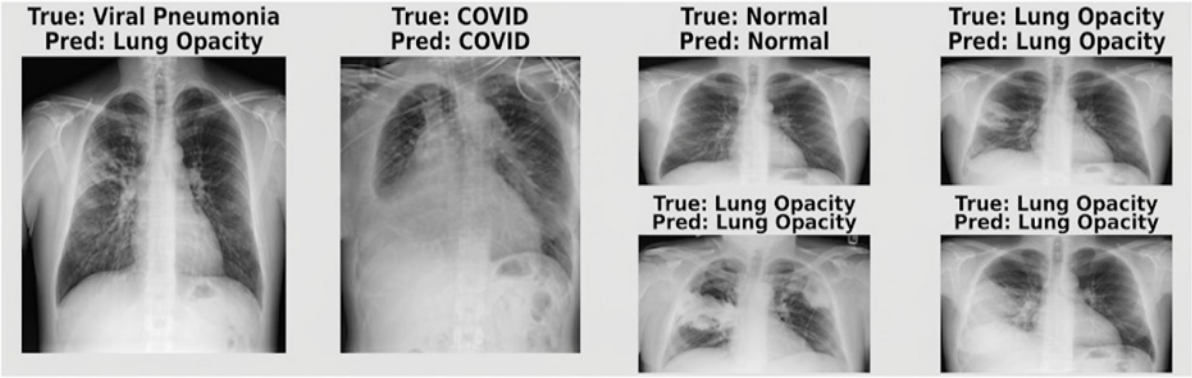

Fig. 7. Representative validation samples showing the true and predicted class labels generated by the proposed model.

4. CONCLUSIONS

The present research evaluated a deep learning based framework designed to categorize chest radiographs into four distinct classes using the COVID-19 Radiography dataset. A custom CNN was compared with an ImageNet-pretrained MobileNetV2 with a frozen convolutional backbone for distinguishing. On the independent test set, the CNN attained 84.34% accuracy, whereas MobileNetV2 attained 91.63% accuracy. The CNN showed variability across five independent runs, with a mean accuracy of 82.87 ± 2.02% and a mean macro F1-score of 83.23 ± 3.52%. The MobileNetV2 showed variability across five-fold cross-validation, with a mean accuracy of 90.29 ± 1.34% and a mean macro F1-score of 91.03 ± 1.04%. A paired McNemar test demonstrated a statistically significant difference between the two models, favouring MobileNetV2. The use of ImageNet-pretrained features provided a stronger transfer-learning baseline for this four-class task. However, the findings should not be inferred as sign of clinical deployment readiness or proven computational efficiency because external hospital validation and direct latency/FLOPs/model-size benchmarking were not performed. Future research should focus on external validation, calibration, multicentre evaluation, controlled augmentation ablation, and computational profiling before deployment-oriented conclusions are made. Future research will focus on external validation using independent clinical datasets, repeated cross-validation, ensemble and transformer-based architectures. Further investigation of calibration, class-wise performance, cross-dataset generalization, and measured computational requirements will be important before considering real-world deployment.

5. ACKNOWLEDGEMENTS

The authors would like to acknowledge the publicly available COVID-19 Radiography Dataset provided through Kaggle.

6. ETHICAL STATEMENT

Because this research relied entirely on open-access datasets without involving human subjects or animal test subjects; therefore, formal ethical approval was not required.

7. FUNDING

No funding was received to conduct this research.

8. CONFLICT OF INTEREST

The authors declare no conflict of interest.

9. AUTHORSHIP CONTRIBUTION STATEMENT

Mushtaq Ahmed: Conceptualization, methodology, software implementation, experimentation, data curation, and manuscript preparation. Gulraiz Javed Joiya: Project administration and manuscript editing. Mutiullah: Formal analysis, validation, visualization, and manuscript review. Nouman Ameer Khan: Resources, data acquisition, investigation, and manuscript review. Faseeh Abid Sheikh: Medical domain expertise, clinical validation, radiological interpretation, and assistance in medical image identification. Nausherwan Malik: Clinical validation, radiological assessment, medical image identification.

ACCEPTED MANUSCRIPT

10. DECLARATION

The authors declare that the results presented in this manuscript are original and that the same material has neither been published nor is under consideration for publication elsewhere. AI tools, including ChatGPT, were used only for English-language refinement, readability, and editorial improvement. All scientific content, results, analyses, and conclusions were independently reviewed and verified by the authors, who remain fully responsible for the manuscript. Approval of all authors has been obtained. In the event of acceptance, copyright of the article will be assigned to the Pakistan Academy of Sciences.

11. REFERENCES

1. M. Ahmed, G. Gilanie, M. Ahsan, H. Ullah, and F.A. Sheikh. Review of artificial intelligence-based COVID-19 detection and a CNN-based model to detect COVID-19 from X-rays and CT images. *VFAST Transactions on Software Engineering* 11(2): 100-112 (2023). https://doi.org/10.21015/vtse.v11i2.1460
2. G.D. Rubin, C.J. Ryerson, L.B. Haramati, et al. The role of chest imaging in patient management during the COVID-19 pandemic: A multinational consensus statement from the Fleischner Society. *Radiology* 296(1): 172-180 (2020). https://doi.org/10.1148/radiol.2020201365
3. L. Wang, Z.Q. Lin, and A. Wong. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports* 10(1): 19549 (2020). https://doi.org/10.1038/s41598-020-76550-z
4. G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W.M. van der Laak, B. van Ginneken, and C.I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42: 60-88 (2017). https://doi.org/10.1016/j.media.2017.07.005
5. C. Shorten and T.M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data* 6(1): 60 (2019). https://doi.org/10.1186/s40537-019-0197-0
6. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations (ICLR)*, *7-9 May 2015, San Diego, CA, USA* (2015). https://arxiv.org/abs/1409.1556
7. A. Narin, C. Kaya, and Z. Pamuk. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern Analysis and Applications* 24(3): 1207-1220 (2021). https://doi.org/10.1007/s10044-021-00984-y
8. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.C. Chen. MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 18-23 June 2018*, *Salt Lake City, UT, USA* pp. 4510-4520 (2018). https://doi.org/10.1109/CVPR.2018.00474
9. A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean. A guide to deep learning in healthcare. *Nature Medicine* 25(1): 24-29 (2019). https://doi.org/10.1038/s41591-018-0316-z
10. S. Kornblith, J. Shlens, and Q.V. Le. Do better ImageNet models transfer better? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15-20 June 2019, Long Beach, CA, USA* pp. 2656-2666 (2019). https://doi.org/10.1109/CVPR.2019.00277
11. D. Ezzat, A.E. Hassanien, and H.A. Ella. An optimized deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization. *Applied Soft Computing* 98: 106742 (2021). https://doi.org/10.1016/j.asoc.2020.106742
12. M. Rahimzadeh and A. Attar. A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2. *Informatics in Medicine Unlocked* 19: 100360 (2020). https://doi.org/10.1016/j.imu.2020.100360
13. S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G.J. Soufi. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. *Medical Image Analysis* 65: 101794 (2020). https://doi.org/10.1016/j.media.2020.101794
14. I.D. Apostolopoulos and T.A. Mpesiana. COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine* 43(2): 635-640 (2020). https://doi.org/10.1007/s13246-020-00865-4
15. Y. Oh, S. Park, and J.C. Ye. Deep learning COVID-19 features on chest X-ray using limited training data sets. *IEEE Transactions on Medical Imaging* 39(8): 2688-2700 (2020). https://doi.org/10.1109/TMI.2020.2993291
16. Kaggle. COVID-19 Radiography Database.

ACCEPTED MANUSCRIPT

Kaggle Datasets Repository (2021). https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database

17. M.E.H. Chowdhury, T. Rahman, A. Khandakar, M.A. Ayari, A.Z. Khan, M.S. Khan, N. Al-Emadi, M.B.I. Reaz, M.T. Islam, and S.H.M. Ali. Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access* 8: 132665-132676 (2020). https://doi.org/10.1109/ACCESS.2020.3010287
18. Y. Brima, M. Atemkeng, S.T. Djiokap, J. Ebiele, and F. Tchakounté. Transfer Learning for the Detection and Diagnosis of Types of Pneumonia including Pneumonia Induced by COVID-19 from Chest X-ray Images. *Diagnostics* 11(8): 1480 (2021). https://doi.org/10.3390/diagnostics11081480
19. I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA (2016).
20. G. Mohan, M.M. Subashini, S. Balan, et al. A multiclass deep learning algorithm for healthy lung, COVID-19 and pneumonia disease detection from chest X-ray images. *Discover Artificial Intelligence* 4: 20 (2024). https://doi.org/10.1007/s44163-024-00110-x
21. H. Lamouadene, M. El Kassaoui, M. El Yadari, A. El Kenz, A. Benyoussef, A. El Moutaouakil, and O. Mounkachi. Detection of COVID-19, lung opacity, and viral pneumonia via X-ray using machine learning and deep learning. *Computers in Biology and Medicine* 191: 110131 (2025). https://doi.org/10.1016/j.compbiomed.2025.110131
22. X. Song, A 19-layer convolutional neural network for accurate COVID-19 detection in chest X-ray images: Comparative analysis with pretrained networks. *Scientific Reports* 16: 25141 (2026). https://doi.org/10.1038/s41598-026-49315-3
23. M.K. Islam, S.U. Habiba, T.A. Khan, and F. Tasnim. COV-RadNet: A deep convolutional neural network for automatic detection of COVID-19 from chest X-rays and CT scans. *Computer Methods and Programs in Biomedicine Update* 2: 100064 (2022). https://doi.org/10.1016/j.cmpbup.2022.100064
24. A. Gopatoti and P. Vijayalakshmi. MTMC-AUR2CNet: Multi-textural multi-class attention recurrent residual convolutional neural network for COVID-19 classification using chest X-ray images. *Biomedical Signal Processing and Control* 85: 104857 (2023). https://doi.org/10.1016/j.bspc.2023.104857
25. T.G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* 10(7): 1895-1923 (1998). https://doi 10.1162/089976698300017197

ACCEPTED MANUSCRIPT