Multimedia Tools and Applications (2023) 82:21311–21351
https://doi.org/10.1007/s11042-022-13844-6

Pediatric pneumonia diagnosis using stacked ensemble learning on multi-model deep CNN architectures

J Arun Prakash[1] · CR Asswin[1] · Vinayakumar Ravi[2] · V Sowmya[1] · KP Soman[1]

Received: 30 March 2022 / Revised: 17 June 2022 / Accepted: 6 September 2022 /
Published online: 20 October 2022

Abstract

Pediatric pneumonia has drawn immense awareness due to the high mortality rates over recent years. The acute respiratory infection caused by bacteria, viruses, or fungi infects the lung region and hinders oxygen transport, making breathing difficult due to inflamed or pus and fluid-filled alveoli. Being non-invasive and painless, chest X-rays are the most common modality for pediatric pneumonia diagnosis. However, the low radiation levels for diagnosis in children make accurate detection challenging. This challenge initiates the need for an unerring computer-aided diagnosis model. Our work proposes Contrast Limited Adaptive Histogram Equalization for image enhancement and a stacking classifier based on the fusion of deep learning-based features for pediatric pneumonia diagnosis. The extracted features from the global average pooling layers of the fine-tuned MobileNet, DenseNet121, DenseNet169, and DenseNet201 are concatenated for the final classification using a stacked ensemble classifier. The stacking classifier uses Support Vector Classifier, Nu-SVC, Logistic Regression, K-Nearest Neighbor, Random Forest Classifier, Gaussian Naïve Bayes, AdaBoost classifier, Bagging Classifier, and Extra-trees Classifier for the first stage, and Nu-SVC as the meta-classifier. The stacking classifier validated using Stratified K-Fold cross-validation achieves an accuracy of

✉ Vinayakumar Ravi
vravi@pmu.edu.sa

J Arun Prakash
arun.jayakanthan@gmail.com

CR Asswin
asswin.cr2001@gmail.com

V Sowmya
v_sowmya@cb.amrita.edu

KP Soman
kp_soman@amrita.edu

[1] Center for Computational Engineering and Networking (CEN), Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

[2] Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia

98.62%, precision of 98.99%, recall of 99.53%, F1 score of 99.26%, and an AUC score of 93.17% on the publicly available pediatric pneumonia dataset. We expect this model to greatly help the real-time diagnosis of pediatric pneumonia.

Keywords Pediatric pneumonia · Chest X-rays · Computer-aided diagnosis · Contrast limited adaptive histogram equalization · Deep learning · Transfer learning · Stacking classifier · Stratified K-fold

1 Introduction

Acute diseases have become a reason for concern over the past few years. The exponential rise in such ailments results in high mortality rates across various countries and long-term economic losses [25, 46]. Pneumonia is one such disease that causes respiratory infection; as a result, harming the normal functioning of the human body. Some of the known pathogen that causes pneumonia include viruses and bacteria. A notable agent for the pervasive transmission of viruses and bacteria is the degradation of air quality [35]. The human lungs constitute tiny bags or sacs called the alveoli. These air sacs are responsible for exchanging essential gases, namely oxygen and carbon-di-oxide. When an individual is affected by pneumonia, these sacs fill up with pus and fluid, decreasing the gas exchange between the blood and the lungs. Pneumonia causes difficulty in breathing and other complications like chest pain, cough, vomiting, diarrhoea, and fatigue.

Pneumonia generally affects children under five years, the geriatric population, and people with co-morbidities like diabetes, cardiovascular disorders, and auto-immune disorders [35]. In 2017, over 850,000 lives perished due to pneumonia-induced disorders. The mortality rate is specifically high in South Asian and Sub-Saharan countries [17]. Pediatric pneumonia has accounted for the high mortality in children in the past five years [32]. In 2011, nearly 1.2 million children under the age of five died due to pneumonia [1]. In 2016, more than 800,000 children died of pneumonia, most of whom were no more than two years old, and the death toll was more than the totality of malaria, AIDS, and measles [20]. In 2019, pneumonia killed 740,180 children under the age of five, accounting for 14% of all deaths of children under five. Pediatric pneumonia accounts for 19% of the overall mortality rate in children below five years of age.

Studies indicate that the other causes of pediatric pneumonia are malnutrition, air pollution, and lack of immunity. The ailments of pediatric pneumonia are easily preventable when diagnosed at a very early stage. Different approaches for diagnosing pediatric pneumonia include chest X-rays to find inflammation in the lungs, blood sample test for arterial blood gas analysis, and sputum test. Other diagnostic measures include pulse oximetry, Complete Blood Cell (CBC), chest CT scan, bronchoscopy, and thoracentesis. Even though highly advanced diagnostic measures exist, the trend in mortality does not seem to decrease due to high prices and inaccessibility. The factors mentioned are vital, especially in underdeveloped countries where people have limited access to such resources. Chest X-ray-based diagnosis, being comparatively inexpensive, was adopted as a standardized test for pediatric pneumonia.

The chest X-ray is the commonly used test for diagnosing various lung diseases as it is a painless and non-invasive method. It is fast, easy, and inexpensive when compared to other techniques. Though chest X-ray diagnosis is time and cost-effective, it heavily relies on the diagnostic conditions and the expertise of doctors and radiologists. Sometimes the chest X-rays may appear different due to outdated equipment and tools in clinical practice. In other cases, getting a proper X-ray in an ideal posture might be difficult. Hence, these factors will affect the quality of the image captured and diagnosis. In addition, using lower radiation levels for chest X-rays in children makes pneumonia diagnosis a cumbersome task.

Detecting the cause of pneumonia in children is tedious due to the lack of rapid commercially available accurate lab tests for most of the existing pathogens which cause pediatric pneumonia [19]. According to World Health Organization (WHO), radiologists face difficulty distinguishing between pus-filled lesions and the lesions caused by external diagnosis conditions. Computer-aided systems are, therefore, necessary to help radiologists and doctors diagnose pediatric pneumonia and standardize chest X-ray-based diagnosis. The advancement in artificial intelligence, specifically machine learning, has made drastic improvements in the automated diagnosis of diseases. Computer-aided diagnosis (CAD) systems make the diagnosis of diseases and prediction tasks easier with minor error margins. Several CAD approaches using machine learning have been employed to diagnose major respiratory problems such as pneumonia [26, 63, 67], tuberculosis [22], Parkinson's disease classification [52], and cardiac diseases [64]. A research team used a machine learning approach to detect diabetic retinopathy, and the results showed that their algorithms performed tantamount to ophthalmologists [8]. The conventional supervised machine learning approach requires handcrafted filters to extract the input data based on domain knowledge. Feature engineering is a cumbersome task requiring a tremendous amount of time and expertise to develop custom filters. Thus, the scalability of handcrafter filters is limited. In unsupervised learning, the features extracted from input data are not labelled, and machines have to discover patterns among the samples, which generally requires a more extensive training dataset. Large datasets, especially in the medical domain, are limited due to privacy concerns. These existing methods are highly improbable in medical imaging for an immediate and accurate CAD model.

Therefore, deep learning emerged as a state-of-art method [39], where it learns the features on its own, making it more robust compared to machine learning and thereby adding advantages to computer-aided diagnostic systems. Deep learning started with multiple hidden layers and gradually extended to multiple CNN layers resulting in deep CNN architectures. These architectures are the most commonly used technique for pattern recognition and image classification tasks [24, 40]. CheXNet, which uses a convolutional neural network that contains 121 layers, was authored by researchers at Stanford University to diagnose pneumonia [7]. A significant challenge while developing an accurate and generalized deep learning model is that the dataset should be large and well-curated for training. It has to cover variability in patients spread all over the globe, tools used for imaging, and other metrics. Collecting such a vast dataset is generally impossible for all medical diseases, especially labelling and annotating, which requires time. A modern breakthrough in artificial intelligence called transfer learning serves to overcome this impediment [45]. Based on the literature survey, it was noticed that the existing deep learning architectures perform well; however, their performance is limited. A large number of neurons in deep architectures lead to the problem of overfitting, which limits the generalizability property of models. The literature survey concludes that the existing models are not guaranteed to perform well on unseen data. Our proposal of a simple and easily replicable solution for the pediatric pneumonia diagnosis is summarized as follows:

- A feature concatenation-based stacking ensemble leveraging the strengths of various machine learning classifiers.

- Using CLAHE as an image pre-processing technique for contrast enhancement to capture the fine occlusions in the image.
- Class activation Maps (CAM) to visualize the regions of interest and t-distributed stochastic neighbour embedding (t-SNE) based feature visualization for layman interpretability of the features predicted by the deep CNN architecture.
- A detailed investigation of the proposed architecture's advantages and limitations and an up-to-date comparison with recent works.
- Performance analysis of the proposed models on similar lung disease datasets proves its generalizability and robustness.

The contributions made in the field of pediatric pneumonia diagnosis that motivated us to advance with the idea of feature fusion-based stacked ensemble learning are as follows:

- The initial research on pediatric pneumonia diagnosis with transfer learning and the dataset's open-source availability [38].
- A transfer learning approach using dilated convolutions and residual structures to solve the limited performance of deep convolution layers [20].
- The effects of different spatial domain image enhancement techniques on COVID-19 detection using chest X-ray images [21].
- A fusion technique involving a deep CNN model with PCA and logistic regression [41].
- A weighted average ensemble of deep CNN models incorporating deep transfer learning [9].
- A majority voting ensemble of the predictions from deep CNN models [65].
- CheXNet [7], a DenseNet121 model trained on the ChestX-ray14 dataset whose performance exceeded that of the average radiologist.

The rest of this article is organized as follows: Section 2 describes the literature survey and discusses the existing gap in the literature and how our approach completes it. The proposed approach is discussed in Section 3. Section 4 contains the description of the dataset. The performance metrics used in this study are detailed in Section 5. The experimental results are analysed and discussed with plots in Section 6. Finally, in Section 7, we conclude our work, summarizing the problem and the limitations of our approach, along with the possible future works.

2 Literature survey

Pattern classification, which earlier required specific filters for feature extraction, switched to automatic feature extraction processes using deep learning architectures. Initial studies for pattern classification were mainly based on MLP approaches. The downside to this approach was its inability to capture local information. Convolutional Neural Networks (CNNs) were introduced to solve this issue. CNN works based on the convolution operation done using the convolution layer, followed by max pooling for reduced computations. The beginning of pediatric pneumonia-based classification research started with the availability of the dataset in Kermany et al. [38]. Earlier studies on MLP and simple CNN architectures were introduced in [5]. Saraiva et al. [5] extended the findings using the cross-validation technique for extensive learning on the existing simple CNN architectures in [50]. Stanford's research radiologists validated the authenticity of the highly sensitive pediatric pneumonia dataset, and CheXNet was introduced to the research community [7]. The CheXNet is a deep CNN architecture based on DenseNet121 with a performance better than the average radiologists. Several custom deep CNN models were developed; however, their performance was limited. The limitation was due to the increasing number of layers in deep CNN models. The impact of the loss of spatial information in chest X-rays when passing through increasing convolutional layers was studied by Gaobo Liang et al. [20]. This spatial information is of utmost importance in detecting pediatric pneumonia. Gaobo Liang et al. [20] introduced a transfer learning-based dilated convolution CNN model to increase the receptive field, thereby reducing the risk of spatial information degradation. A purely depthwise separable convolution approach followed this variation in CNNs [51]. Siddiqi et al. [55] proposed a deep sequential CNN for pediatric pneumonia detection. Nahid et al. [54] suggested a novel two-channel CNN architecture for pneumonia diagnosis. Yu et al. [34] introduced a graph-based feature reconstruction for pediatric pneumonia diagnosis called CGNet. The new class of deep learning architectures, Capsule Networks for pediatric pneumonia detection, was introduced by Mittal et al. [68]. Wu et al. [31] proposed a hybrid system for diagnosing pneumonia from chest X-Ray images consisting of an adaptive median filter CNN recognition model based on random forest.

Several research studies introduce architectures with competing performances. Rahman et al. [66] studied the performance of AlexNet, ResNet18, DenseNet201, and SqueezeNet utilizing transfer learning. El Asnaoui et al. [42] compared the results of fine-tuned deep learning architectures for binary classification in chest X-rays. Predefined weights are essential in determining a model's performance in transfer learning. Mahajan et al. [3] analyzed the performance differences between CheXNet weights, ImageNet weights, and random weights for the task at hand. The problem of class imbalance in machine learning is a necessity that must be addressed. For unbiased training, machine learning relies mainly on a balanced dataset. Sampling serves as a solution to this class imbalance problem. Using the Xception network, Luján-Garca et al. [30] investigated random undersampling (RUS) for unbiased training and applied a cost-sensitive learning strategy.

Researchers began to analyze the importance of using spatial domain preprocessing techniques like Histogram Equalization, Adaptive Histogram Equalization (AHE), Contrast Limited Adaptive Histogram Equalization (CLAHE), and Gamma correction. Rubini et al. [37] examined two popular spatial processing approaches for improving MRI images: AHE and CLAHE. In [29], the authors propose using CLAHE as the preprocessing technique and custom CNN architectures for prediction. Nneji et al. [33] propose using dynamic histogram enhancement techniques for pediatric pneumonia detection. Tawsifur et al. [66] detail the effects of HE, CLAHE, image complement, gamma correction, and balance contrast enhancement approaches.

In recent studies, ensemble methods proved to show improved performances. Chouhan et al. [65] analyzed the performance of a majority voting ensemble composed of AlexNet, DenseNet121, Inception V3, GoogLeNet, and ResNet18. Sagar Kora Venu [9] proposed combining these deep CNN models - MobileNetV2, Xception, DenseNet201, ResNet152V2, and InceptionResNet - into a weighted average ensemble. Nahida et al. [41] proposed combining a deep convolutional neural network for feature extraction, principal component analysis for dimensionality reduction, and logistic regression for classification**.** On ensembled features from VGG-19 and CheXNet, Nahida et al. [47] proposed using Random Under Sampling, Random Over Sampling and SMOTE for classification. Islam et al. [10] illustrate using feature concatenations from SqueezeNet and InceptionV3 combined with ANNs to make predictions. El Asnaoui et al. [18] proposed average fusion of prediction made by ResNet50, InceptionResNetV2, and MobileNetV2. Ensemble of RetinaNet and Mask R-CNN for pneumonia detection and localization was proposed by Sirazitdinov et al. [2]. A weighted average ensemble of GoogLeNet, ResNet18, and DenseNet121 with a novel assignment weights criteria was proposed in [57]. Rajaraman et al. [23] compared the performance of iterative model pruning of deep learning architectures with majority voting, weighted average ensemble averaging, and stacking on the COVID-19 dataset.

The lack of transparency is the most significant impediment to a complete transition to artificial intelligence (AI). Explainable AI (XAI), a promising research topic, has recently gained much attention. Nguyen Hai et al. [44] introduced a novel technique incorporating explainability for pneumonia detection. They proposed a mixture of custom CNN architecture with Grad-CAM for pneumonia detection. Liz et al. [36] used the XAI framework and an average ensemble of five custom CNN predictions.

An abundance of research has been done in this field. However, there exist limitations are discussed below:

1. Studies emphasize the use of CheXNet weights for custom CNN training which is a challenging task.
2. Lots of research proposes using custom complex architectures that are not easily replicable and hampers the reproducibility of the work.
3. The absence of exploring ensemble approaches pertinent to pediatric pneumonia diagnosis was observed. The same was witnessed concerning the use of machine learning classifiers.
4. Data sampling methods like RUS, ROS, and SMOTE lead to longer training times and over-fitting.
5. Most of the studies mentioned above failed to cover the aspect of feature visualization. Feature visualization is crucial to ensure the learned features are meaningful for predictions.
6. Most studies fail to prove the generalizability of their proposed models and approaches.

The distinguishing factor in our work compared to the rest of the works is detailed in Table 1. Our work proposes a detection pipeline to bridge the gap in the existing literature. The dataset has been redistributed for unbiased training instead of data sampling methods**.** We propose the use of a contrast enhancement technique for image preprocessing. The proposed methodology is based on the feature fusion of DenseNet121, DenseNet169, DenseNet201, and MobileNet architectures pre-trained on the commonly available ImageNet weights for feature extraction. The extracted feature maps from the global average pooling layer are passed to the t-SNE and CAM for visualization. Stacking ensemble classifier approach with KNN, SVC, Random Forest Classifiers, Nu-SVC, Logistic Regression, Extra-trees Classifier, Bagging Classifier, Gaussian Naïve Bayes, and AdaBoost classifier was used along with Stratified K-Fold cross-validation to overcome overfitting. All additional details are discussed in the forthcoming sections.

3 Proposed approach

This section details the workflow of the proposed pediatric pneumonia detection architecture, from the enhancement of images to the final classification, as illustrated in Fig. 1. The input

Table 1 A summarizing comparison of recent literary works with the current study

Reference	Dataset	Pretrained model	Feature visualization	Feature fusion	Ensemble method
Gaobo Liang et al. [20]	Kermany et al. [38]	Custom CNN based on dilated convolutions	Yes	No	No
Nahida Habib et al. [41]	Kermany et al. [38]	CheXNet (DenseNet121)	No	No	No
Mahajan et al. [3]	Kermany et al. [38]	InceptionV3, DenseNet121	No	No	No
Yu et al. [34]	Kermany et al. [38], CT-Pneumonia dataset	CGNet	No	No	No
Siddiqi et al. [55]	Kermany et al. [38]	Custom deep sequential CNN	No	No	No
Nahida Habib et al. [47]	Kermany et al. [38], COVID-19 database	CheXNet, VGG-19	No	Yes	Ensemble of CNN models
Siddiqi et al. [51]	Kermany et al. [38]	PneumoniaNet	No	No	No
Chouhan et al. [65]	Kermany et al. [38]	AlexNet, DenseNet121, ResNet18, IncpetionV3, GoogLeNet	Yes	No	Majority Voting
Rahman et al. [66]	Kermany et al. [38]	DenseNet201	No	No	No
Mittal et al. [68]	Kermany et al. [38]	CapsNet	No	Yes	Ensemble of convolution capsules
Islam et al. [10]	Kermany et al. [38]	SqueezeNet, InceptionV3	No	Yes	No
Kora Venu Sagar [9]	Kermany et al. [38]	DenseNet201, Xception, InceptionResNet, ResNet152V2, MobileNetV2	No	No	Weighted average ensemble
Wu et al. [31]	Kermany et al. [38]	Adaptive median filter CNN	No	No	Ensemble of trees
Current work	Kermany et al. [38]	MobileNet, DenseNet121, DenseNet169, DenseNet201	Yes	Yes	Stacking ensemble

chest X-ray is preprocessed and sent to pre-trained deep CNN architectures with transfer learning to extract the features from the global average pooling layer. The next step is the column-wise concatenation of the extracted features. The concatenated features are passed through the stacking classifier with level one classifiers as SVC, KNN, Logistic Regression, NuSVC, Random Forest Classifier, AdaBoost Classifier, GaussianNB, Bagging Classifier, and Extra-trees Classifier. The binary predictions from the ML classifiers are sent to the NuSVC meta classifier for the final predictions.

The dataset contains images of varying sizes. In this study, we reshape the images according to the requirement of different deep CNN models. Each image is normalized to bring the pixel values between the range 0–1. Image augmentations were introduced on the fly using the Keras image generator as a necessary part of modeling to prevent over-fitting. The

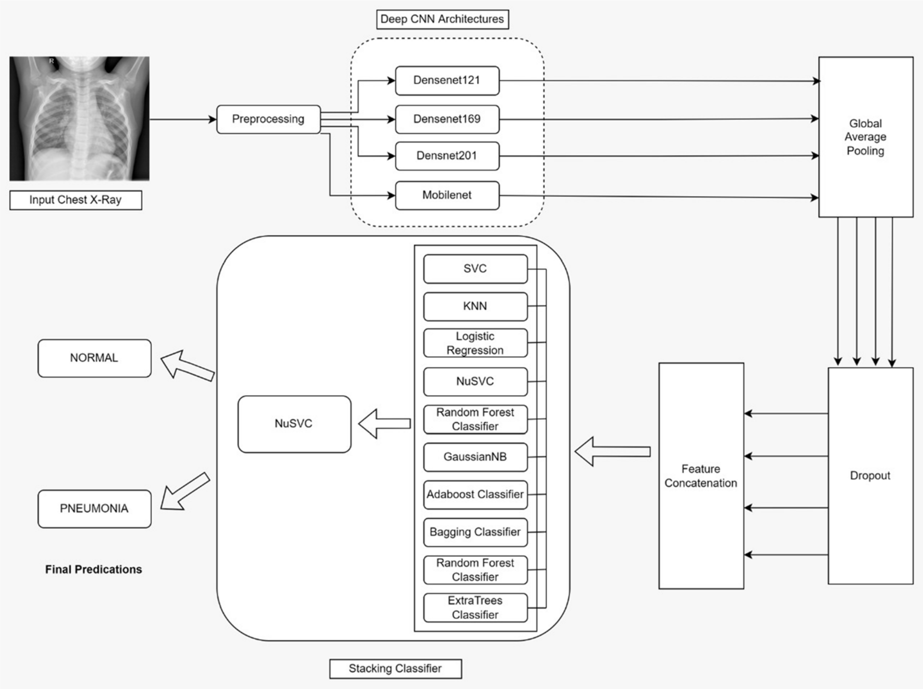


Fig. 1 Proposed architecture for pediatric pneumonia classification

augmentations used include zoom, shear, and flip. Table 2 describes the corresponding augmentation values.

3.1 Image enhancement

Spatial image enhancement techniques are widely used to enhance specific features in an image. Contrast enhancement adjusts the relative difference between darkness and brightness to improve visibility. Contrast enhancement becomes very important in medical image diagnosis, which deals with chest X-rays and CT scans. Several studies strongly recommend using contrast enhancement techniques for better results [21, 28, 49, 53]. Adaptive Histogram Equalization (AHE) is an image processing technique for enhancing contrast.

Table 2 Augmentations used in our study and their corresponding values

Methods	Corresponding Parameters
Rescale	255
Shear	0.2
Zoom	0.2
Horizontal Flip	True

It generates histograms for each image part and uses them to redistribute its values. As a result, this is well suited for improving the local contrast and edge regions in images. However, in relatively homogeneous areas of an image, AHE tends to overamplify noise. Contrast Limited Adaptive Histogram Equalization (CLAHE), a type of adaptive histogram equalization, is thus used to restrict the amplification. CLAHE works on small regions (pixels) in the image known as tiles, where the adjacent tiles are combined using bilinear interpolation. The slope of the transformation function determines the contrast amplification in the neighborhood of a specific pixel value in CLAHE. The Blind/Reference less Image Spatial Quality Evaluator (BRISQUE) metric assessed the image quality after enhancement. The lower the BRISQUE score, the more perceptual quality it retains.

3.2 Transfer learning

The performance of any deep learning model depends on the available training data. Massive datasets are proven to increase the performance metrics of the deep learning models. However, this is not usually the case in the real world due to privacy concerns, especially in medical image data. We use transfer learning to overcome this challenge of limited data. Transfer learning uses the weights pre-trained on a similar dataset; it is then adapted to the new target task by fine-tuning it for the task at hand. We fine-tuned the already pre-trained models on ImageNet in our proposed method.

3.3 Deep learning models

The literature survey discussed earlier ascertained that using pre-trained deep CNN models provided an overall better performance. Existing pre-trained deep CNN models such as VGG16 [48], VGG19 [48], MobileNet [56], MobileNetV2 [56], InceptionResNetV2 [14], DenseNet121 [60], DenseNet169 [60], DenseNet201 [60], InceptionV3 [16], ResNet50 [59], ResNet101 [59], ResNet152 [59], ResNet50V2 [12], ResNet101V2 [12], ResNet152V2 [12], EfficientNetB0 [13] and Xception [61] are trained on the Contrast Limited Adaptive Histogram Equalization (CLAHE) enhanced dataset to select the top four best performing models. The features of the selected architectures are extracted from the global average pooling layer to retrieve a feature map for each image. These extracted features are concatenated for stacked ensemble learning.

3.3.1 MobileNet

MobileNet is a deep convolutional neural network model open-sourced by Google [56], designed for mobile and embedded vision applications. It is a streamlined architecture that uses depthwise separable convolution for lightweight computations. Depthwise separable convolution comprises two factorized convolutions: standard depthwise convolution and pointwise convolution, as shown in Fig. 2. The first phase is the depthwise spatial convolution, where the convolution is done depthwise for each input channel with a single filter. The following phase uses pointwise convolution, which applies 1 × 1 convolution to combine the output of depthwise convolution, thereby reducing the

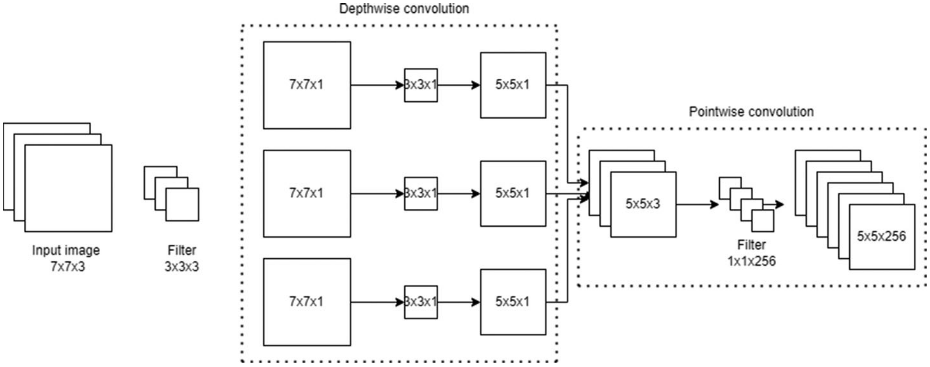


Fig. 2 Illustration of the working of a depthwise separable convolution network

computational cost. This method of factorizing the convolution process into two phases reduces the model's size. MobileNet uses 28 convolutional layers with 3 × 3 depthwise separable convolutions. All layers in the architecture are followed by batch normalization and ReLU nonlinear activation. Figure 3 describes the entire architecture of MobileNet.

3.3.2 DenseNet

Convolutional Neural Networks rely on the gradients in the image for feature retrieval. Increasing convolution layers introduces the vanishing gradient problem, hence explaining the staggering performance with the increasing number of convolutional layers. A unique network architecture called the DenseNet was proposed by Huang et al. [60] as a solution to this vanishing gradient problem. In DenseNet, each layer is interconnected in a feed-forward manner. Features maps of all preceding layers are used as input to each layer, and their feature maps are used as inputs into all subsequent layers. This collective knowledge retained several advantages: an improved flow of information in the network alleviated the problem of relearning redundant features and decreased the number of learnable parameters.

Furthermore, since each layer has direct access to the gradients of any preceding layer, training deep network architectures become much more manageable. In addition, the regularizing effect introduced by the deep connections reduces overfitting when trained on small datasets. Our proposed approach uses DenseNet121, DenseNet169 and DenseNet201. The DenseNet121 was the first model released from the DenseNet family with 121 convolutional layers. After that, researchers began experimenting with added convolutional layers and eventually released DenseNet169 with 169 convolutional layers. The DenseNet201 with 201 layers is the most recent advancement in the DenseNet models and outperforms all other deep CNN models. Figure 4 illustrates the architecture of the DenseNet model.

3.3.3 Hyperparameters

Hyperparameter tuning plays a significant role in increasing the performance of the model. Thus, after extracting the feature maps from global average pooling, we added a dropout rate of 0.4 to avoid overfitting. Next, the feature maps are sent to the softmax layer, which uses the sigmoid activation function for the binary - NORMAL and PNEUMONIA classification.

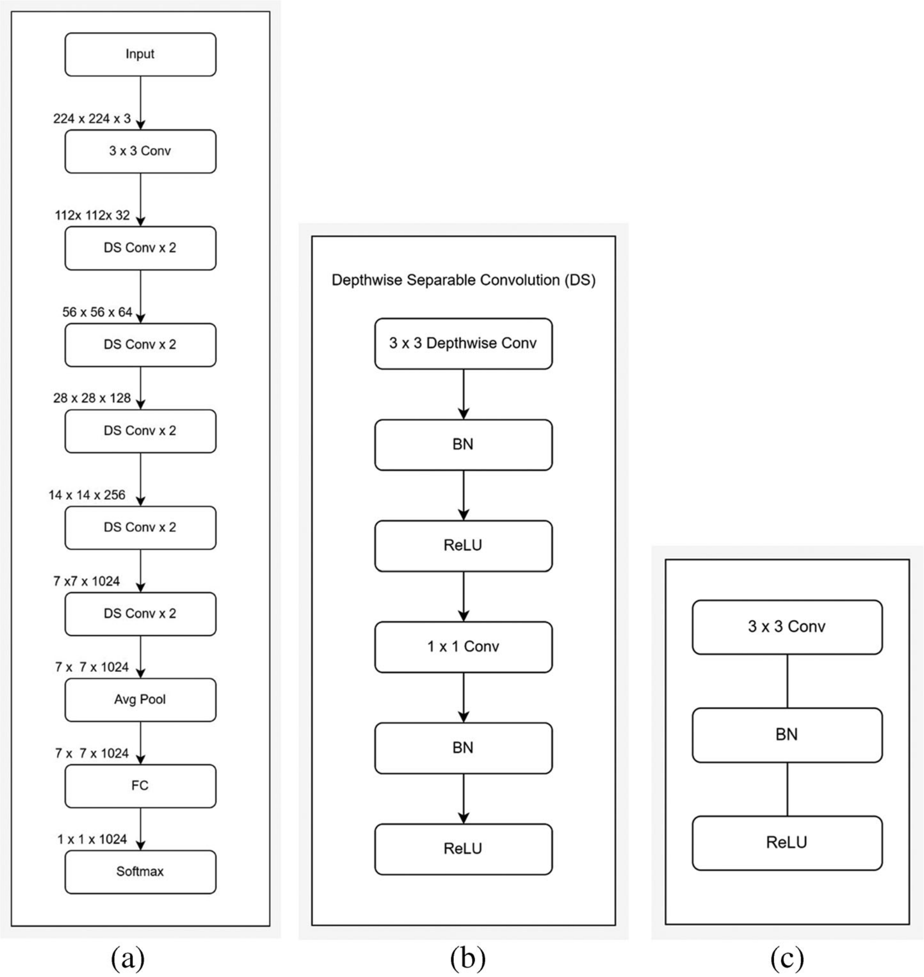


Fig. 3 **a** Entire MobileNet architecture with normal convolutions and depthwise separable convolutions. **b** All the layers following a depthwise separable convolution block. **c** All the layers following a normal (3 × 3) convolution block in MobileNet architecture

Finally, we fine-tuned the models based on different optimizers to find the best hyperparameters, as shown in Table 3.

The Binary Cross Entropy calculates the difference between the expected and actual output. The value for the loss function ranges between 0 and 1 and is given by Eq. 1.

$$\text{Loss} = \sum_{i=0}^{outputsize} y_i * \log\left(\widehat{y}_i\right) + (1-y_i) * \log\left(1-\widehat{y}_i\right) \tag{1}$$

3.4 T-SNE feature representation

t-Distributed Stochastic Neighbour Embedding(t-SNE) is a non-linear dimensionality reduction technique primarily used for visualizing high-dimensional data. This algorithm calculates

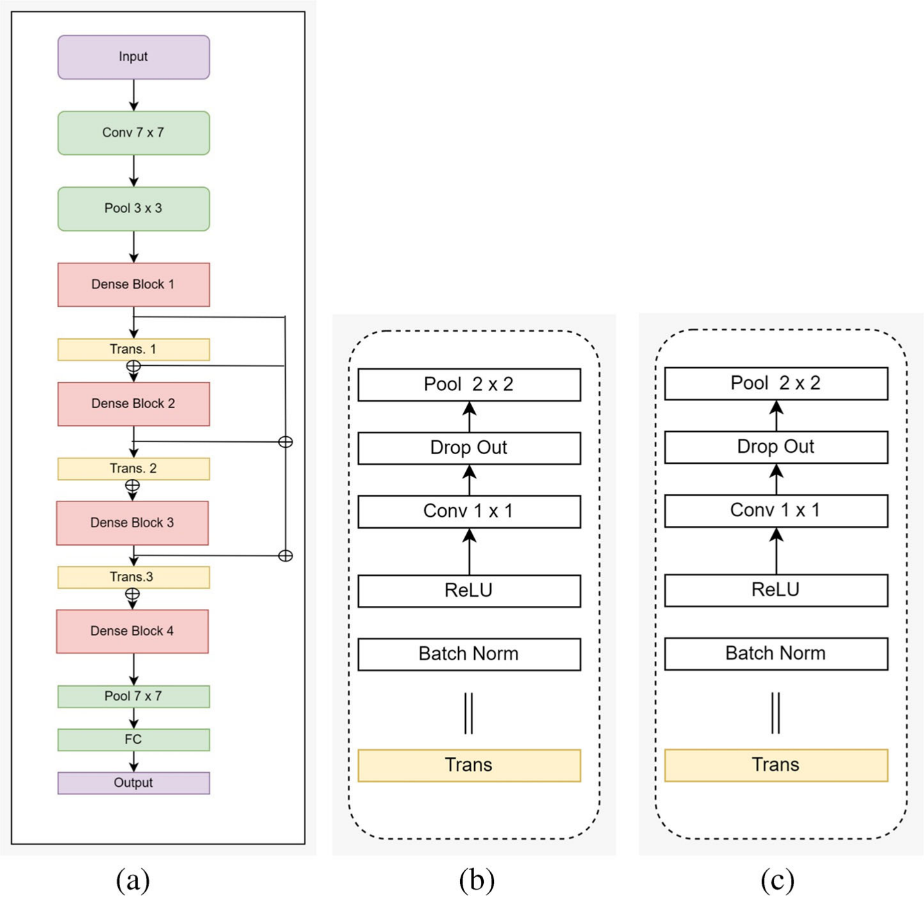


Fig. 4 **a** The entire architecture of DenseNet deconstructed into in **b** Trans block with Dense interconnections and **c** Dense block used in the DenseNet architecture

the resemblance between pairs of instances in high and low dimensional space and then optimizes them using the cost function. The feature maps of the test data are visualized using the t-SNE plot. Unlike PCA, this representation allowed us to visualize the binary classes (normal and pneumonia) as clusters**.** It gives an insight into how well the predictions are made. This visualization gives an insight into the required classifiers for the task.

Table 3 List of hyperparameters and their values used in our study to finalize the perfect combination for the task at hand

Hyper Parameter	Corresponding Values
Optimizer	Adam, SGD, Nadam, RMSprop, Adamax, Adagrad
Learning rate	0.001
Batch size	32

3.5 Stacking classifier

Ensemble learning is the most commonly used technique to increase the performance of a model, thereby making it much more robust. Stacking classifiers is one of the most widely used ensemble techniques for classification or regression. This type of classifier combines different heterogeneous machine learning algorithms such as logistic regression, support vector classifier, and k-nearest classifier, unlike other techniques such as bagging and boosting. The stacking classifier consists of two stages. The first stage consists of several machine learning classifiers stacked together whose predictions are trained on the second stage classifier. The second stage classifier, called the meta classifier, provides the final classification**.**

3.5.1 SVC

The support vector classifier generates decision boundary in such a way that maximizes the margin between the support vectors belonging to different classes. This operation is preceded by mapping to high dimensional spaces with different kernel functions.

3.5.2 KNN

The K-Nearest Neighbour algorithm uses the proximity between k neighbours to classify the data points. The distance between the data points is calculated using different methods namely: hamming, Manhattan and Euclidean. The majority class of the first k distances in the ascending order assign the categorical class value to a given data point.

3.5.3 Logistic Regression

The Logistic Regression is the simplest solution to any binary classification task. It uses a nonlinear log transformation to map and classify points without a linear relationship. If the value of the logistic function is greater than 0.5, it is said to belong to class 1 or else class 0.

3.5.4 NuSVC

The NuSVC is very similar to SVC, except that instead of parameter C, which penalizes the wrong predictions in the optimization algorithm, the NuSVC uses the regularization parameter nu. While C has no direct interpretation, the parameter nu sets the upper bound and lower bound on the fraction of margin errors and the fraction of support vectors, respectively.

3.5.5 Random Forest Classifier

The Random Forest algorithm, used for classification and regression, is one of the most commonly used machine learning algorithms. Decision Trees are the main components of Random Forests. The algorithm becomes more advanced as the number of trees increases. It chooses the best result from the votes gathered by the trees, making it robust. The higher the number of trees in the forest, the more accurate it is, and the problem of overfitting is avoided. It comprises two phases: the first is to combine N decision trees with building a random forest, and the second is to generate predictions for each tree created in the first phase.

3.5.6 Gaussian Naïve Bayes Classifier

The Gaussian Naive Bayes model works based on the Bayes' theorem, assuming that the features are independent of each other. The model learns the conditional probability distribution of the input features such that the likelihood of obtaining the correct class is maximized, as shown in Eq. 2.

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y) \prod_{i=1}^{n} P(x_i|y) \tag{2}$$

3.5.7 AdaBoost Classifier

AdaBoost, also known as Adaptive Boosting, is a machine learning approach utilized as a part of an ensemble method. Decision trees with one level, or decision trees with only one split, are the most popular algorithm used with AdaBoost. This approach creates a model by assigning equal weights to all data points. It then gives points that are incorrectly categorized a higher weight. In subsequent models, points with greater weights are given more importance. This iterative process will continue to train models until a lower error is received.

3.5.8 Bagging Classifier

A bagging classifier is an ensemble meta-estimator technique that fits base classifiers on random subsets of the original dataset and then combines their predictions by voting or average to generate the final prediction.

3.5.9 Extra Trees Classifier

The Extra Trees Classifier uses a meta estimator to get the final prediction based on predictions made by randomized decision trees that fit on various sub-samples of the data.

3.6 Stratified K-fold cross-validation

Cross-validation is the most widely employed technique to estimate the model's performance on unseen data. The performance on unseen data is of utmost importance for real-world deployment. The Stratified K-fold Cross-Validation technique is used when we have an imbalanced dataset. It is an extension of normal K-fold cross-validation. But here, rather than splits being completely random, the ratio of target classes is the same in each fold. Hence this method is used to preserve the class ratio for our target classes and when we have relatively fewer training examples.

4 Dataset description

All the experiments in this study were conducted on the Kermany et al. [38] dataset. The dataset contains 5856 chest X-ray images, which belong to two classes Normal (1583 X-rays) and Pneumonia (4273 X-rays). The existing dataset was imbalanced and was therefore re-redistributed for unbiased training. Table 4 illustrates the new data distribution. The chest X-ray images are from children aged between 1 to 5 years from the Guangzhou Women and

Table 4 Distribution of the dataset for our study

Category	Train	Validation
Normal	1400	183
Pneumonia	1700	2573
Total	3100	2756

Children's Medical Centre. The second row with white patches in the alveolar region of the lungs in the X-ray image shown in Fig. 5 represents the residence of pus and fluid.

4.1 Performance metrics

The critical aspect of evaluation metrics is distinguishing between models' performances. The primary metrics used in this study to evaluate our models are accuracy, precision, recall, F1-score, and the AUC value. The confusion matrix shown in Fig. 6 illustrates the predictions made by the model. The rows and columns of the confusion matrix represent the actual values and the predicted values of the target variable, respectively.

True Positive (TP) - number of pneumonia X-rays correctly predicted as pneumonia
False Negative (FN) - number of pneumonia X-rays wrongly predicted as normal
True Negative (TN) - number of normal X-rays correctly predicted as normal
False Positive (FP) - number of normal X-rays predicted wrongly as pneumonia

The accuracy of a model is calculated as the ratio of correct predictions to the total number of predictions, as shown in Eq. 3. Precision is the ratio of the true positives and the number of predicted positives, as shown in Eq. 4. The recall for a class label is calculated as the ratio between the true positive and the total number of actual positives, as shown in Eq. 5. It measures the

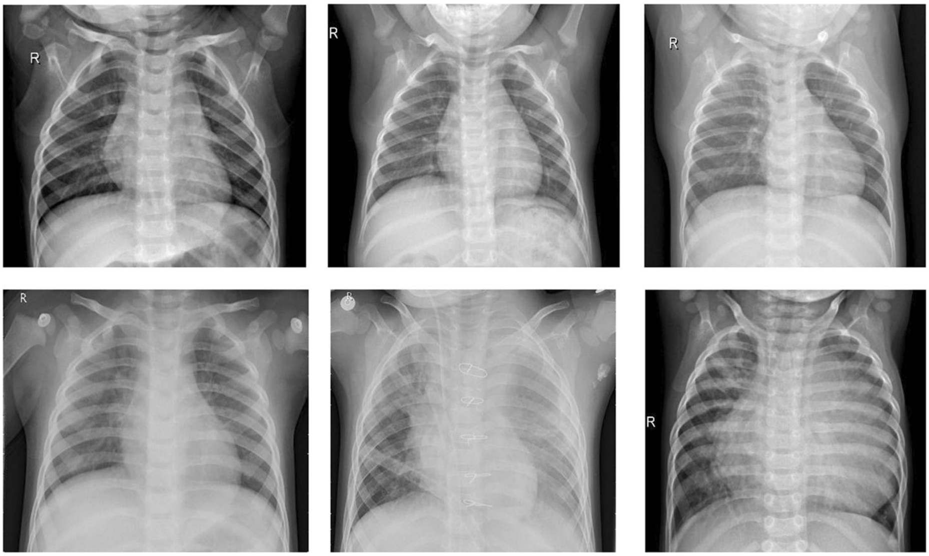
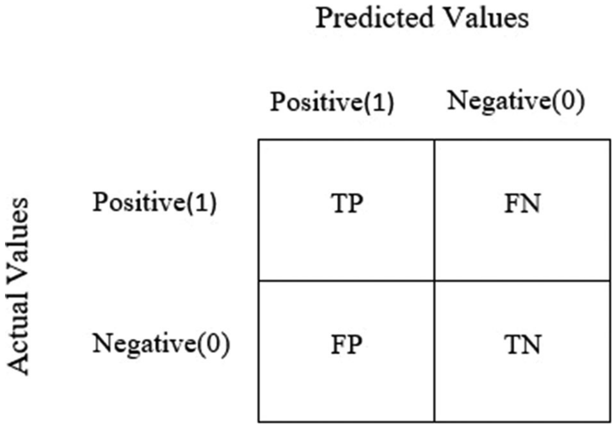


Fig. 5 Samples of Normal x-rays and Pneumonia x-rays from the dataset in the first row and second row respectively



		Predicted Values	
		Positive(1)	Negative(0)
Actual Values	Positive(1)	TP	FN
	Negative(0)	FP	TN

Fig. 6 Confusion matrix

model's ability to detect positive samples. If the model correctly classifies all the positive samples, then the recall will be 1. F1 score is the harmonic mean of precision and recall, as shown in Eq. 6. AUC is a measure to distinguish between classes, i.e., when the AUC value is equal to 1, the classifier will be able to differentiate between children with and without pneumonia.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1 score} = 2\left(\frac{\text{Precision}{*}\text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{6}$$

Table 5 Parameter used for CLAHE image enhancement with the Brisque average calculated over 20 pneumonia samples

Image enhancement used	Clip limit	Tile grid size	Mean Brisque score
CLAHE	0.03	8,8	129.47
	0.02	8,8	129.22
	0.01	4,4	129.26
	0.01	8,8	129.16
	0.01	16,16	133.40
	0.01	32,32	139.53
	0.01	64,64	142.89

Table 6 Fine-tuning information and the number of trainable parameters associated with each model used in our study

Deep CNN model	*Fine-tuned from*	*Total number of trainable parameters*
MobileNetV2	77	2,064,769
MobileNet	50	2,665,473
EfficientNetB0	118	3,700,169
DenseNet121	213	4,632,897
DenseNet169	297	8,544,833
DenseNet201	353	12,741,185
VGG16	9	13,569,793
Xception	66	14,860,313
InceptionV3	155	16,791,489
VGG19	11	17,699,329
ResNet50V2	95	21,352,449
ResNet50	87	21,364,225
ResNet101V2	188	30,625,793
ResNet101	172	30,640,129
ResNet152V2	282	39,836,673
ResNet152	257	39,855,617
InceptionResNetV2	390	41,922,529

5 Results and discussion

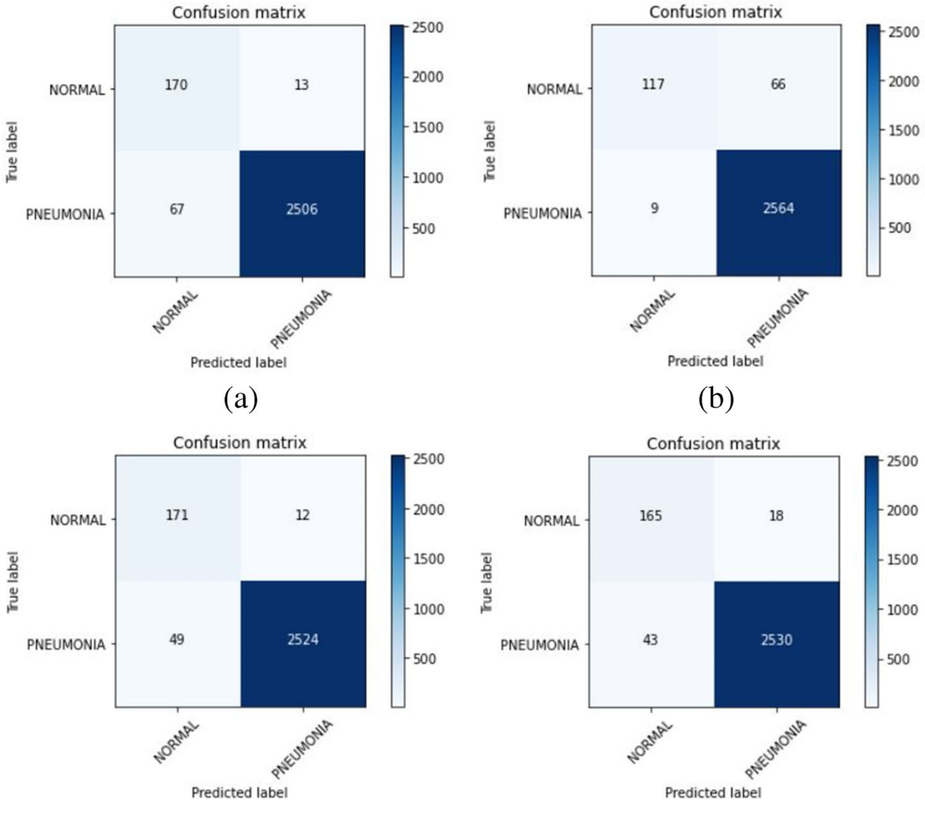
Several deep CNN architectures were trained and validated on the CLAHE enhanced dataset to find the best-performing models suited for the task at hand. The parameters for CLAHE enhancement were selected meticulously to ensure that the crucial features in these chest X-rays do not disappear in the process. Table 5 describes parameters and their BRISQUE scores for the image enhancement techniques used. These scores are calculated as the mean of twenty pneumonia X-rays. The parameters resulting in the least BRISQUE scores are selected for

Table 7 Performance chart of deep learning models with values rounded off to the nearest two decimal positions on the CLAHE enhanced images

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE	AUC
RESNET152	0.47	1.00	0.43	0.60	0.71
MOBILENETV2	0.61	1.00	0.59	0.74	0.79
RESNET101	0.63	1.00	0.66	0.80	0.82
VGG16	0.85	1.00	0.84	0.91	0.90
RESNET152V2	0.86	1.00	0.85	0.92	0.92
INCEPTIONRESNETV2	0.89	0.89	0.88	0.94	0.94
RESNET101V2	0.91	0.96	0.94	0.95	0.71
XCEPTION	0.91	1.00	0.90	0.95	0.97
RESNET50V2	0.92	1.00	0.91	0.95	0.95
VGG19	0.93	0.93	1.00	0.97	0.50
RESNET50	0.93	0.93	1.00	0.97	0.50
EFFICIENTNETB0	0.93	0.93	1.00	0.97	0.50
INCEPTIONV3	0.94	1.00	0.93	0.97	0.96
DENSENET121	0.94	1.00	0.94	0.97	0.97
MOBILENET	0.97	0.99	0.97	0.98	0.95
DENSENET169	0.98	1.00	0.98	0.99	0.96
DENSENET201	0.98	0.99	0.98	0.99	0.94
PROPOSED METHOD	0.99	1.00	0.99	0.99	0.93

enhancement. Looking at the mean BRISQUE values, it is noticed that increasing the tile grid size leads to loss of perceptual quality. The parameters clip limit =0.01, and tile grid size = 8,8 were selected for this study.

Literature survey concludes that several deep CNN architectures achieve competing accuracies for pediatric pneumonia diagnosis. To find the best architectures for the task at hand, the following models are analyzed and compared: VGG16, VGG19, MobileNet, MobileNetV2, InceptionResNetV2, DenseNet121, DenseNet169, DenseNet201, InceptionV3, ResNet50, ResNet101, ResNet152, ResNet50V2, ResNet101V2, ResNet152V2, Xception and EfficientNetB0. Each of the models above were pre-trained on ImageNet weights with the corresponding input image of size 224 × 224, 224 × 224, 224 × 224, 224 × 224, 224 × 224, 224 × 224, 224 × 224, 224 × 224, 299 × 299, 224 × 224, 244 × 244, 224 × 224, 224 × 224, 224 × 224, 299 × 299, 299 × 299, and 224 × 224 respectively. All the architectures were trained for 30 epochs with Adam as the optimization function and the learning rate as 0.001 in Google Colab resourced with K80 GPU and 12 GB RAM. Tensorflow2 and Keras2 were used to build and evaluate the models. The models were fine-tuned from their specified layers, as shown in Table 6. Table 6 also includes the trainable parameters for all the models used in this study.



(a) (b) (c) (d)

Fig. 7 Confusion matrix for (**a**) MobileNet predictions on the test data (**b**) DenseNet121 predictions on the test data (**c**) DenseNet169 predictions on the test data (**d**) DenseNet201 predictions on the test data

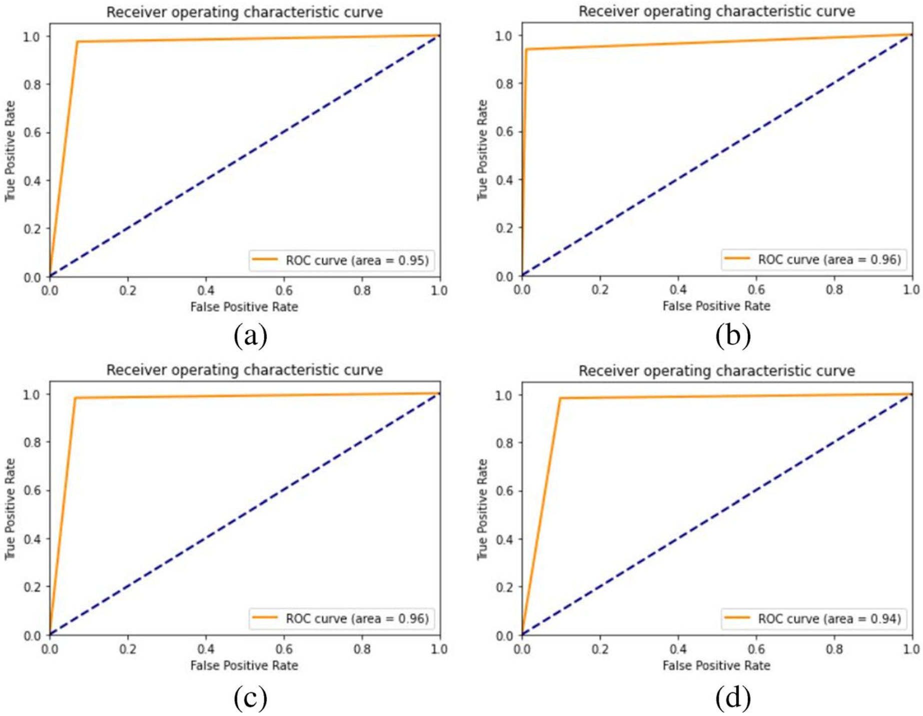


Fig. 8 ROC curve for (**a**) MobileNet predictions on the test data (**b**) DenseNet121 predictions on the test data (**c**) DenseNet169 predictions on the test data (**d**) DenseNet201 predictions on the test data

The results in Table 7 show that DenseNet169 and DenseNet201 are the best performing models. The family of DenseNet models performs consistently well compared to other architectures. DenseNet169 and DenseNet201 achieve the highest accuracy of 97.79%. MobileNet precedes them with an accuracy of 97%, followed by DenseNet121 with an accuracy of 94%. One common attribute in the best-performing models is the residual connection. The residual connections are a key factor that has suppressed over-fitting and thus enabled the above models to perform well on the test data. The thought of collective

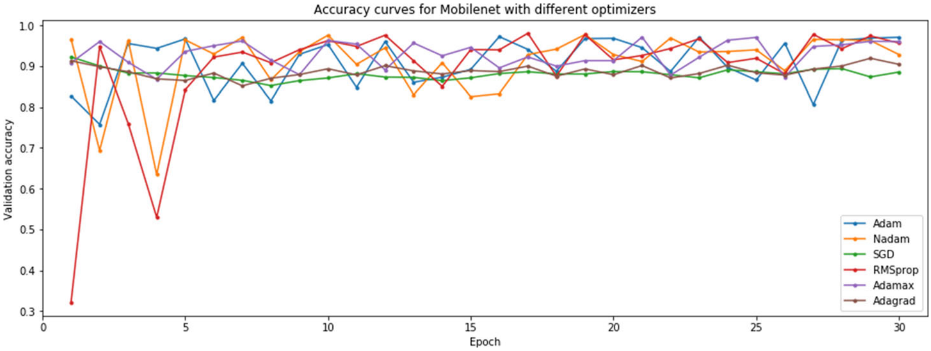


Fig. 9 MobileNet model performance on the validation set using different optimizers

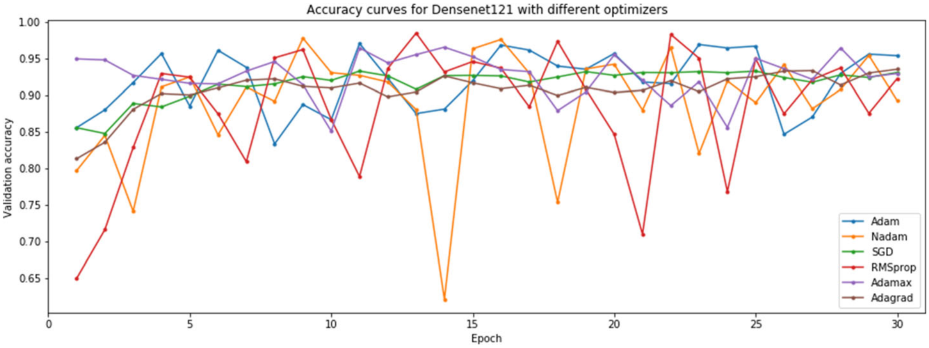


Fig. 10 Densenet121 model performance on the validation set using different optimizers

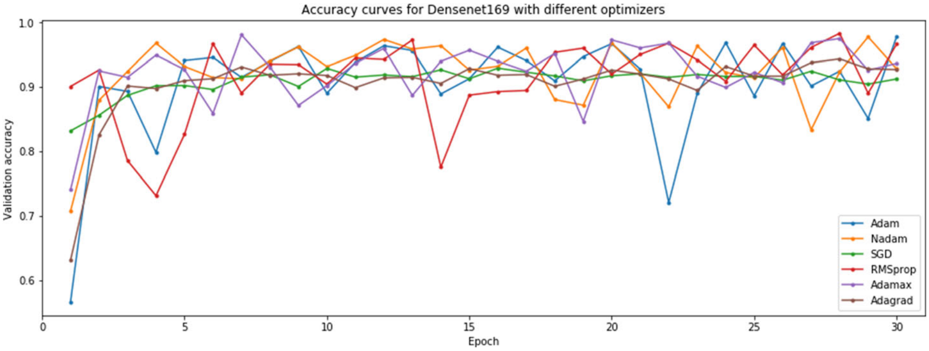


Fig. 11 Densenet169 model performance on the validation set using different optimizers

knowledge and residual connections in Densenets has enabled it to achieve the highest accuracy. The confusion matrix for the test data predictions from MobileNet, DenseNet121, DenseNet169, and DenseNet201 are shown in Fig. 7. MobileNet predicts 67 false negatives and 13 false positives. DenseNet121, DenseNet169 and DenseNet201 predicts 9,49,43 false negatives and 66,12,18 false positives respectively. DenseNet121 has the lowest false-negative

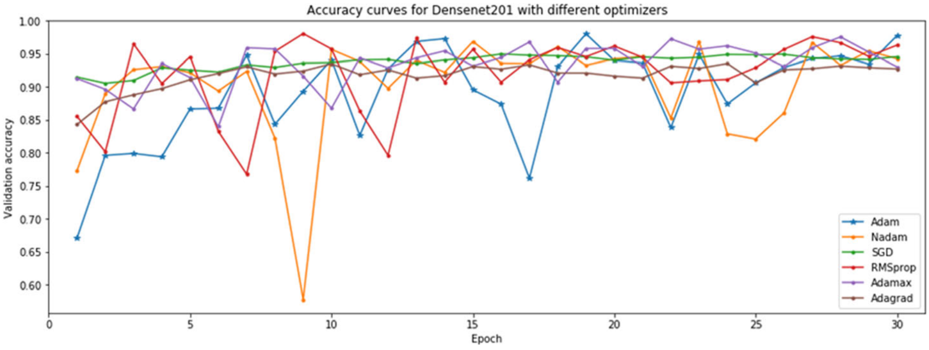


Fig. 12 Densenet201 model performance on the validation set using different optimizers

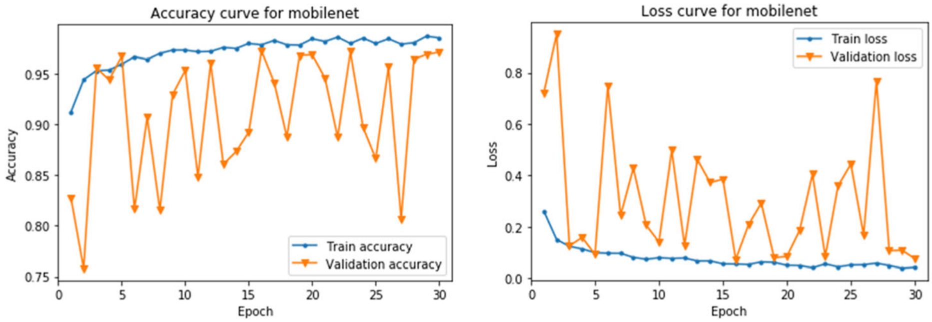


Fig. 13 Training and validation accuracy-loss history of the fine-tuned MobileNet model

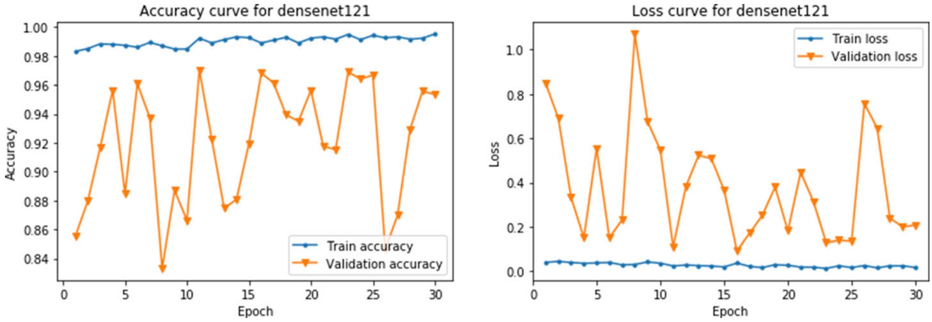


Fig. 14 Training and validation accuracy-loss history of the fine-tuned DenseNet121 model

predictions in the Densenet family. The confusion matrices conclude that the architectures are unable to deal with false predictions. The ROC curves for test data predictions using MobileNet, DenseNet121, DenseNet169, and DenseNet201 are shown in Fig. 8. DenseNet201 has the lowest AUC value of 0.94 compared to the rest. DenseNet121 and DenseNet169 share the highest AUC score of 0.96. Our proposed method achieves the highest accuracy, precision, recall, and f1-score but shows a reduced AUC score.

Extensive experiments were conducted for hyperparameter tuning of each of the best-performing models. The models were fine-tuned on different optimizers to find the best fit for

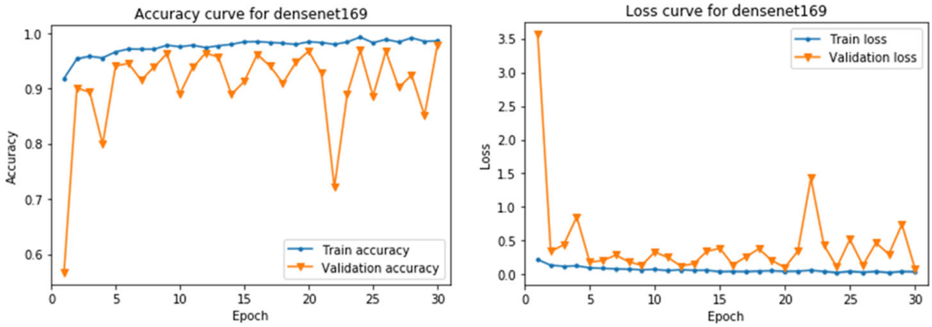


Fig. 15 Training and validation accuracy-loss history of the fine-tuned DenseNet169 model

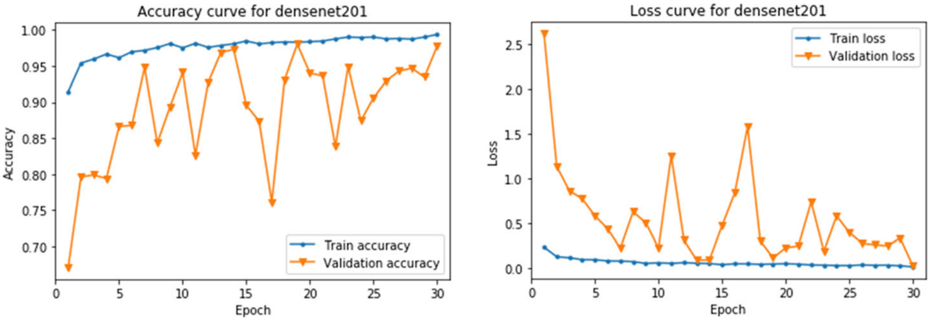


Fig. 16 Training and validation accuracy-loss history of the fine-tuned DenseNet201 model

the task at hand. The Adam optimizer performs best for all models, as seen in Figs. 9, 10, 11 and 12. Based on Figs. 9, 10, 11, and 12, the optimal hyperparameters with the Adam optimizer and a constant learning rate of 0.001 were chosen for feature extraction. The validation loss and accuracy plots of MobileNet, DenseNet121, DenseNet169, and DenseNet201 are shown in Figs. 13, 14, 15 and 16. All the architectures exhibit initial oscillatory behavior. This oscillation is persistent in MobileNet and Densenet121 whereas, in Densenet169 and Densenet201, it gradually reduces with the increasing number of epochs. For the MobileNet model, the validation loss and accuracy are constrained to 1–0 and 0.75–1, respectively. For the DenseNet121 model, the validation loss and accuracy are constrained to 1.2–0 and 0.8–1, respectively. For the DenseNet169 model, the validation loss and accuracy are constrained to 3.7–0 and 0.4–1, respectively. For the DenseNet201 model, the validation loss and accuracy are constrained to 2.8–0 and 0.6–1, respectively.

With the best performing models selected from Table 7, we propose using MobileNet, DenseNet121, DenseNet169, and DenseNet201 for feature extraction. The extracted features

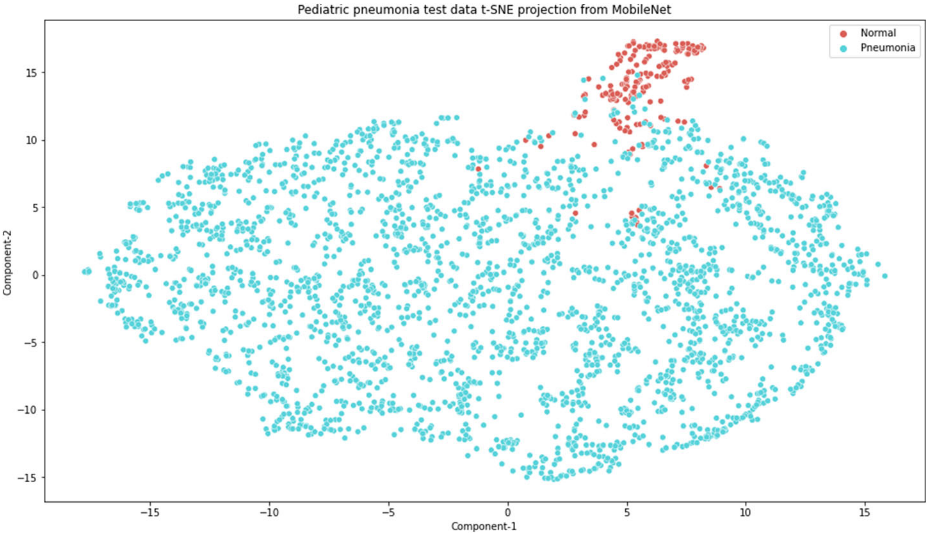


Fig. 17 t-SNE feature representation of the test data extracted from the MobileNet model

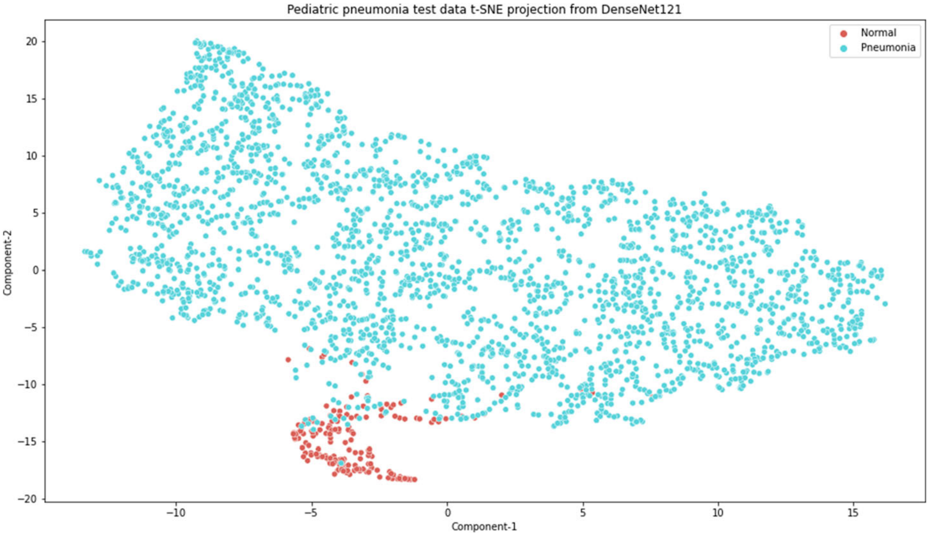


Fig. 18 t-SNE feature representation of the test data extracted from the DenseNet121 model

are concatenated for the final classification. The extracted features are visualized using the t-SNE representation for the layman's interpretability of the features predicted by the models. The parameter values used for visualization are n_components = 2, perplexity = 40, and n_iter = 300**.** Figs. 17, 18, 19, and 20 illustrate the t-SNE plot of the extracted feature maps from each selected architecture.

The t-SNE feature representations conclude that all the models have minor overlapping cluster formations and are non-linearly separable. In the cluster formation of the concatenated

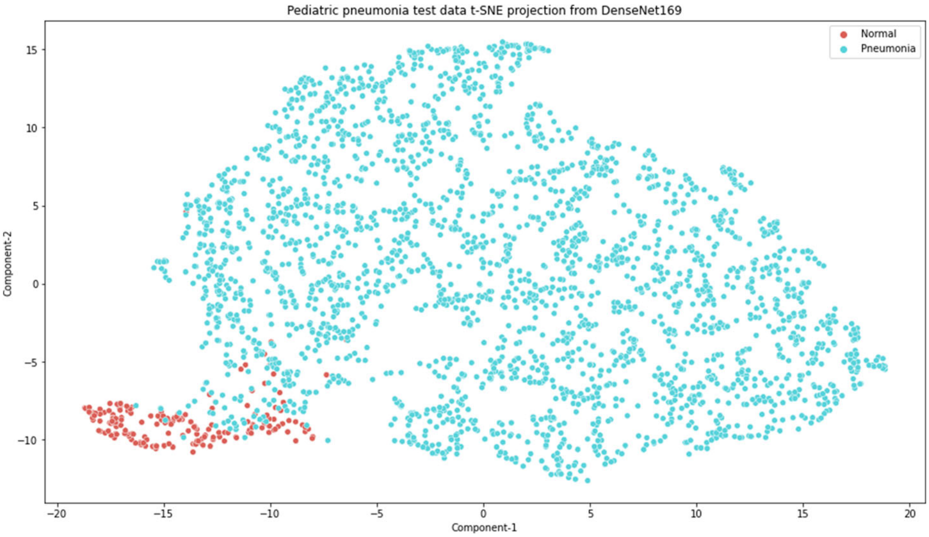


Fig. 19 t-SNE feature representation of the test data extracted from the DenseNet169 model

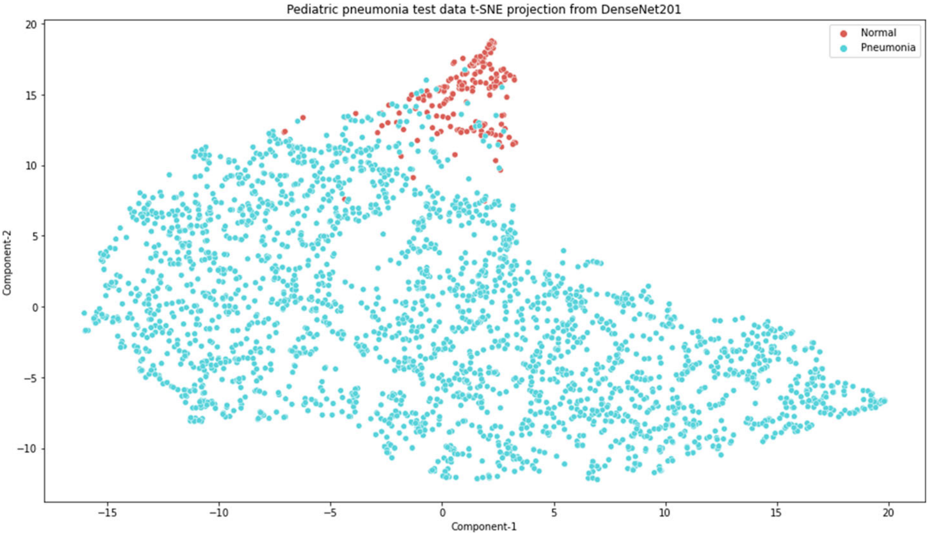


Fig. 20 t-SNE feature representation of the test data extracted from the DenseNet201 model

features, Fig. 21, we notice fewer overlaps than their individual counterparts. The feature representations indicate a possible classifier that can deal with such complexity. This study proposes using the stacking classifier to deal with the non-linearly separable classification.

Class Activation Maps (CAM) are employed to understand the region of interest proposed by these deep CNN architectures. An overview of the regions of interest in the predictions is crucial before real-time deployment as a life-saving resource. The first set of class activation

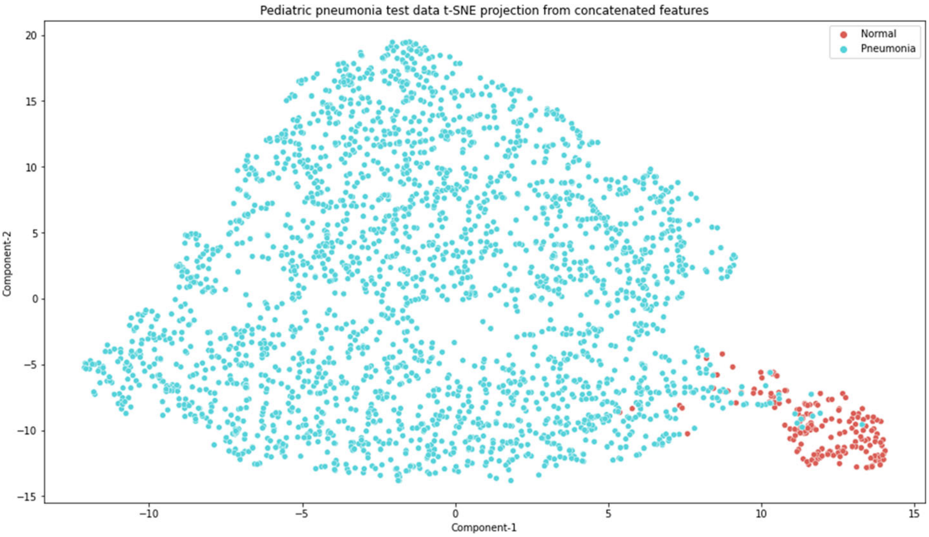


Fig. 21 t-SNE feature representation of the test data from the feature fusion of MobileNet, DenseNet121, DenseNet169, and DenseNet201

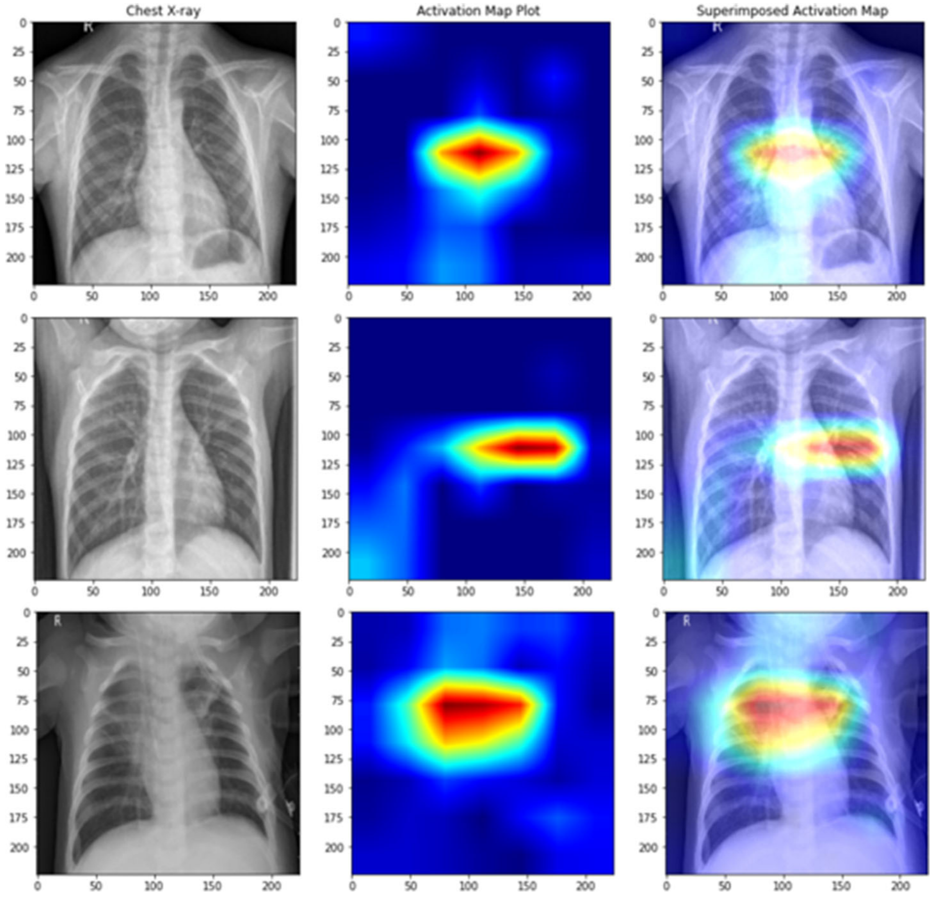


Fig. 22 Class activation maps of misclassified X-rays from MobileNet (row 1: normal classified as pneumonia, row 2: normal classified as pneumonia, row 3: pneumonia classified as normal)

maps for each model misclassified normal as pneumonia, normal as pneumonia, and pneumonia as normal, as shown in Figs. 22, 24, 26, and 28, respectively. The second set of the class activation maps for each model correctly classifies normal and pneumonia X-rays, as shown in Figs. 23, 25, 27, and 29, respectively. From the CAMs of each model, we conclude that the misclassified samples are possibly due to over-fitting. The plots conclude that the accurateness of class activation maps of the concatenated features relies purely on the strength of the individual models. The CAMs of the extracted features from MobileNet, DenseNet121, DenseNet169, and DenseNet201 and their concatenated features are shown in Figs. 30 and 31 for pediatric pneumonia diagnosis for misclassified and correctly classified samples, respectively.

Several machine learning classifiers were trained on the concatenated features and validated against the stacking classifier. Table 8 concludes that the stacking classifier outperforms all machine learning classifiers by leveraging the strength of individual estimators. The first stage in the stacking classifier leverages the RandomForests, Support Vector Classifier, KNeighborsClassifier, GNBClassifier, LogisticRegression, Nu-Support Vector Classifier, ExtraTreesClassifier, AdaBoost Classifier, and a Bagging Classifier. The hyperparameters for

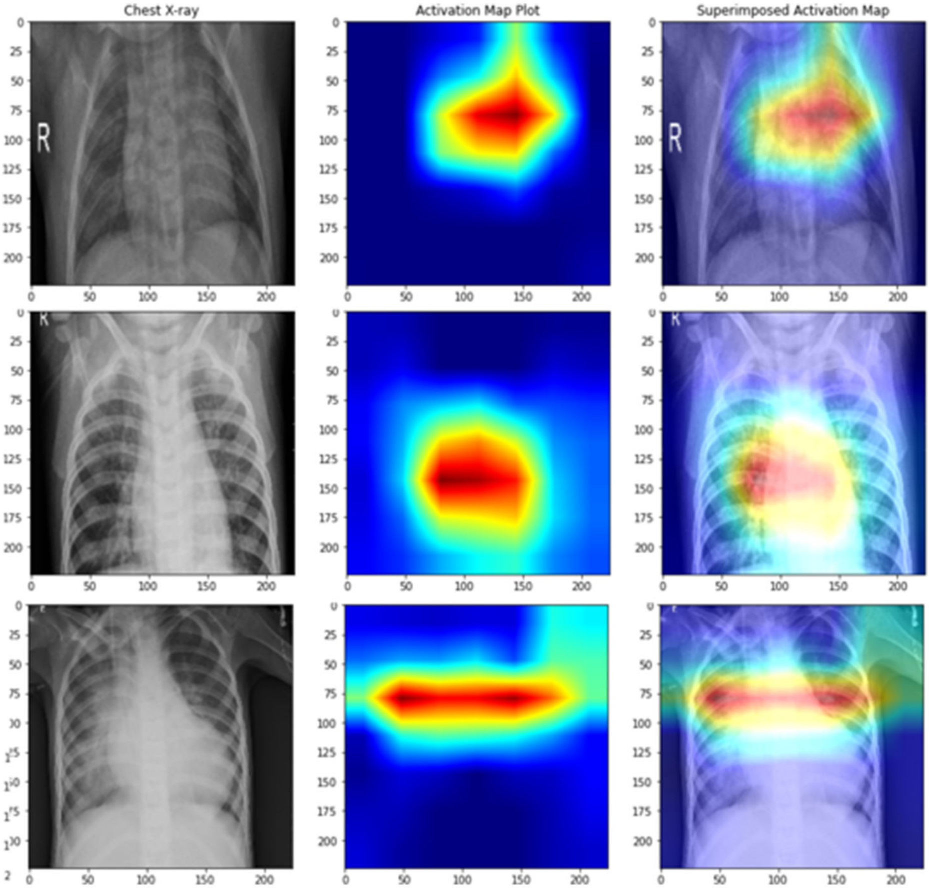


Fig. 23 Class activation maps of correctly classified X-rays from MobileNet (row 1: normal classified as normal, row 2: pneumonia classified as pneumonia, row 3: pneumonia classified as pneumonia)

each of these classifiers were selected using the Bayesian optimization strategy in optune and are detailed in Table 9. Individual predictions from each of the five classifiers are sent to the meta-classifier for the final classification. The meta classifier uses Nu-SVC with nu = 0.5, kernel = “rbf”, degree = 3. Stratified K-Fold cross-validation with n_splits = 30 was employed to help the model learn the most from the existing limited dataset and prevent over-fitting.

The confusion matrix for Stratified K-Fold cross-validation stacking classifier predictions on the test set is shown in Fig. 32. The false predictions are lower in number compared to the raw predictions from the deep CNN architectures due to the collective strength of individual CNN models and the wide range of machine learning classifiers. The perfect combination of base estimators is of utmost importance. Though the accuracy increased by 0.83%, the AUC value has not increased. The number of false-positive predictions is minimal and equal to the number of false positives in DenseNet169. Though the number of false negatives is not as minimal as that of DenseNet121, the proposed approach has overall reduced false predictions. Looking at the confusion matrix, the loss of 1.38% in the model’s accuracy might favorably be due to the imbalanced dataset or insufficient training samples. Another possible reason is the

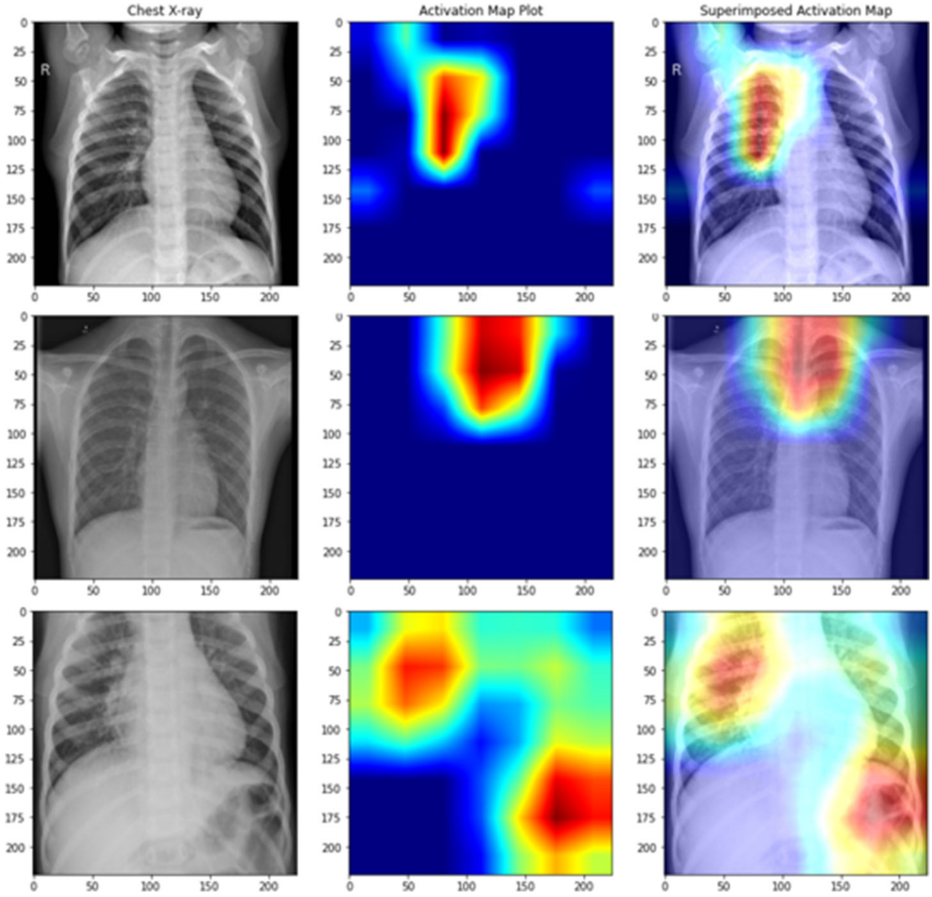


Fig. 24 Class activation maps of misclassified X-rays from DenseNet121 (row 1: normal classified as pneumonia, row 2: normal classified as pneumonia, row 3: pneumonia classified as normal)

use of a simple feature concatenation technique that solely relies on their individual predictors. The proposed method achieves an accuracy of **98.62%**.

Table 10 compares our proposed approach's performance, technique, and classification classes with other recent works. The proposed work exhibits competing performances with other literary works. All the works mentioned in the Table 1 validated their results tested on the Kermany et al. [38] dataset. The advantage of the proposed method compared to other works in Table 10 is in the visualization of features learned by the model using both CAMs and t-SNE plots. Since the feature concatenation of MobileNet, DenseNet121, DenseNet169, and DenseNet201 are used as the feature extractor is based on the commonly available ImageNet weights, reproducibility is easier. In addition to stacking various machine learning classifiers for rich predictions, the proposed models were tested on unseen similar lung disease datasets for generalization and robustness, which was previously absent in recent works.

The proposed model's limitation is its heavy reliance on the individual deep CNN architectures and the correct combination of base classifiers for accurate classification. It was noticed that the meta classifier played a crucial role in determining the performance of

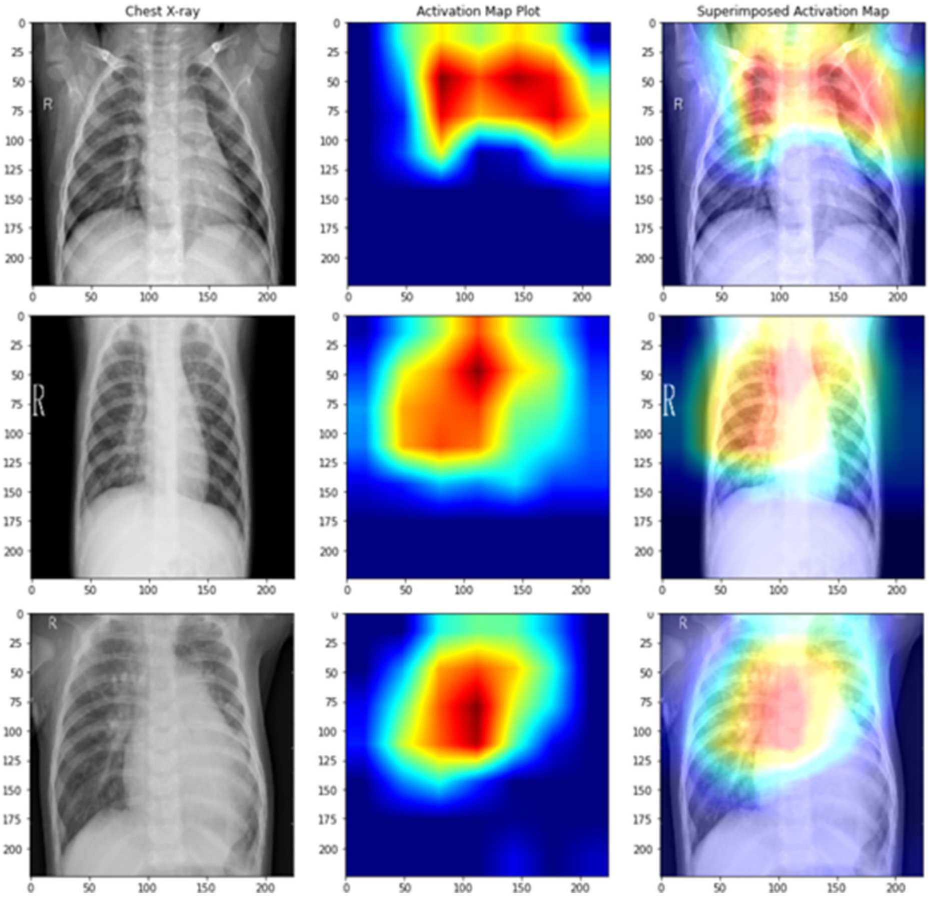


Fig. 25 Class activation maps of correctly classified X-rays from DenseNet121 (row 1: normal classified as normal, row 2: pneumonia classified as pneumonia, row 3: pneumonia classified as pneumonia)

the proposed architecture. While a few experiments resulted in AUC scores reaching 98%, it reduced other metrics by a negligible percentage, and the same was observed vice-versa. The dependence on the fit deep CNN models is a drawback as any one of the models being overfit can affect the final predictions. Though ensemble learning was used to prevent overfitting, the performances indicate the need for cost-sensitive learning-based approaches to learn from the small-scale imbalanced dataset. Other limitations include the inability to subclassify into different stages of pediatric pneumonia like early, latent and severe.

5.1 Robustness and generalization of the proposed approach for lung disease classification

Generalization is an important criterion to be considered before real-time deployment. The proposed deep CNN models trained on the Kermany et al. [38] pediatric pneumonia dataset are tested on similar lung diseases like COVID-19, Tuberculosis, and Pneumonia [6, 15, 27, 43, 62]. The extracted features from each of MobileNet, DenseNet121, DenseNet169, and

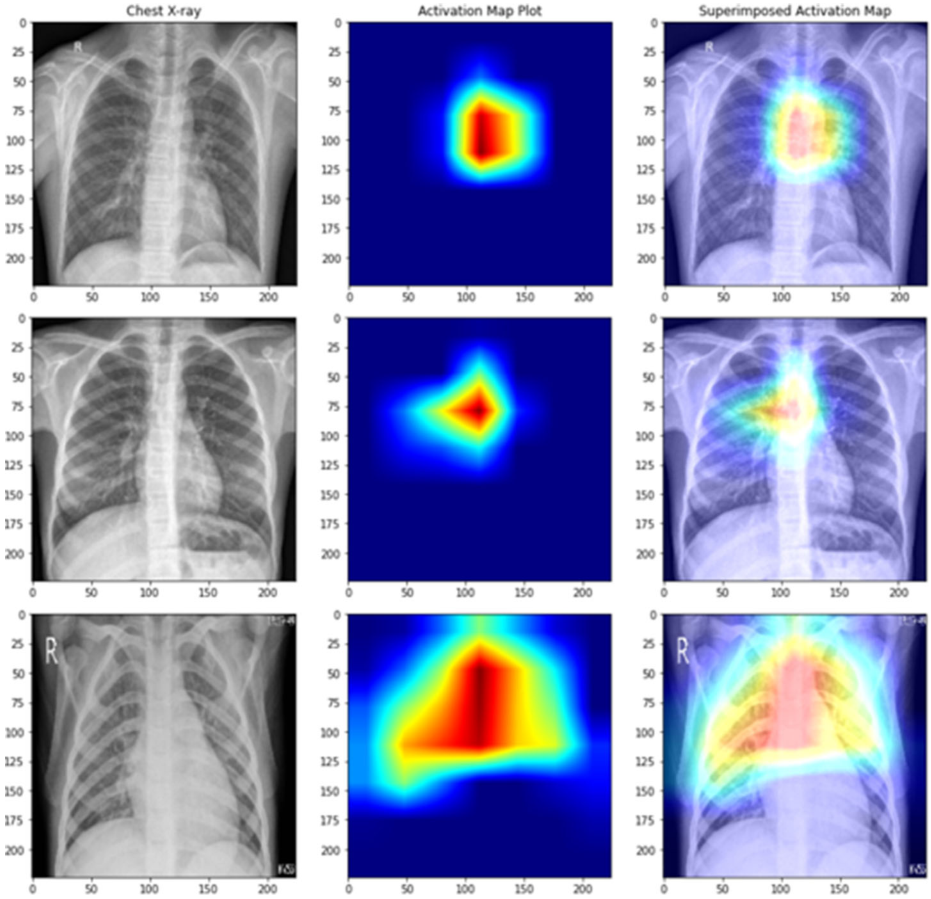


Fig. 26 Class activation maps of misclassified X-rays from DenseNet169 (row 1: normal classified as pneumonia, row 2: normal classified as pneumonia, row 3: pneumonia classified as normal)

DenseNet201 are concatenated and classified using the k-means clustering algorithm. The k-means algorithm was employed to infer the ability to correctly form clusters in the extracted features for reliable classification of similar lung diseases. The results for each dataset are analyzed as classification reports and confusion matrices.

In the COVID-19 vs. normal vs. pneumonia classification dataset [4], the proposed feature fusion achieves an accuracy of 45% with good classification margins for the normal category, as shown in Table 11. The models achieve 1643 correct classifications for normal images. It shows misclassification for the viral adult pneumonia, and COVID-19 predictions as the model are unable to extract the required features for COVID-19 and adult pneumonia detection. The same can be observed in the normal vs. pneumonia vs. tuberculosis classification dataset [4, 58], where the proposed feature fusion achieves an accuracy of 48% with good classification margins for the normal category, as shown in Table 12. The models achieve 1646 correct classifications for normal images. It shows misclassification for the viral adult pneumonia and tuberculosis predictions as the models are unable to extract the required features for the same. The proposed method achieves an accuracy of 60% in the normal vs. pneumonia dataset [43]

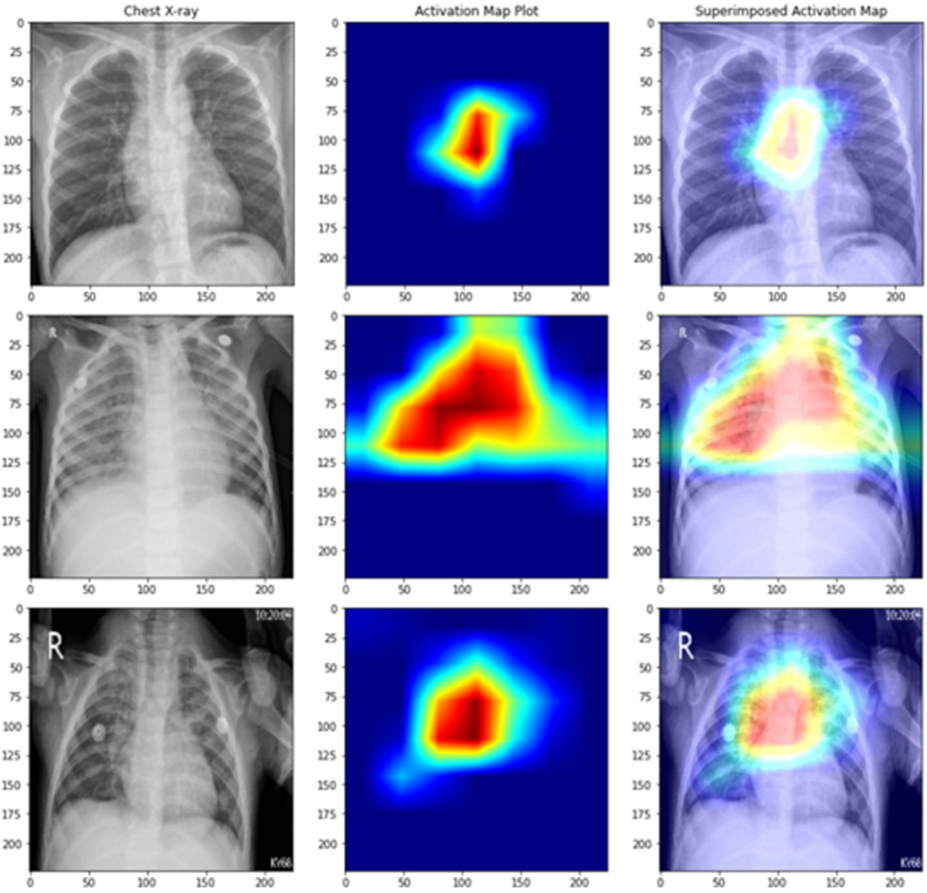


Fig. 27 Class activation maps of correctly classified X-rays from DenseNet169 (row 1: normal classified as normal, row 2: pneumonia classified as pneumonia, row 3: pneumonia classified as pneumonia)

with 80 false negative predictions, as shown in Table 13 and Fig. 33. The results show the proposed models' capability to classify normal chest X-rays from chest X-rays with infections correctly. The results also conclude that though the challenges of pediatric pneumonia diagnosis are characteristically different from adult pneumonia and other similar lung diseases, the proposed method can be extended to aid with diagnosing other lung diseases.

6 Conclusion and future work

This work proposes a computer-aided diagnosis model for detecting pediatric pneumonia using chest X-rays. The use of low radiation levels in chest X-rays for children makes detection a difficult task. Highly trained physicians then diagnose these chest X-rays meticulously to confirm the presence of the acute infection. This process requires loads of time and heavily relies on the availability of experts, which is not always feasible. Other works in the same field include using novel architectures and an ensemble of deep CNN models with the

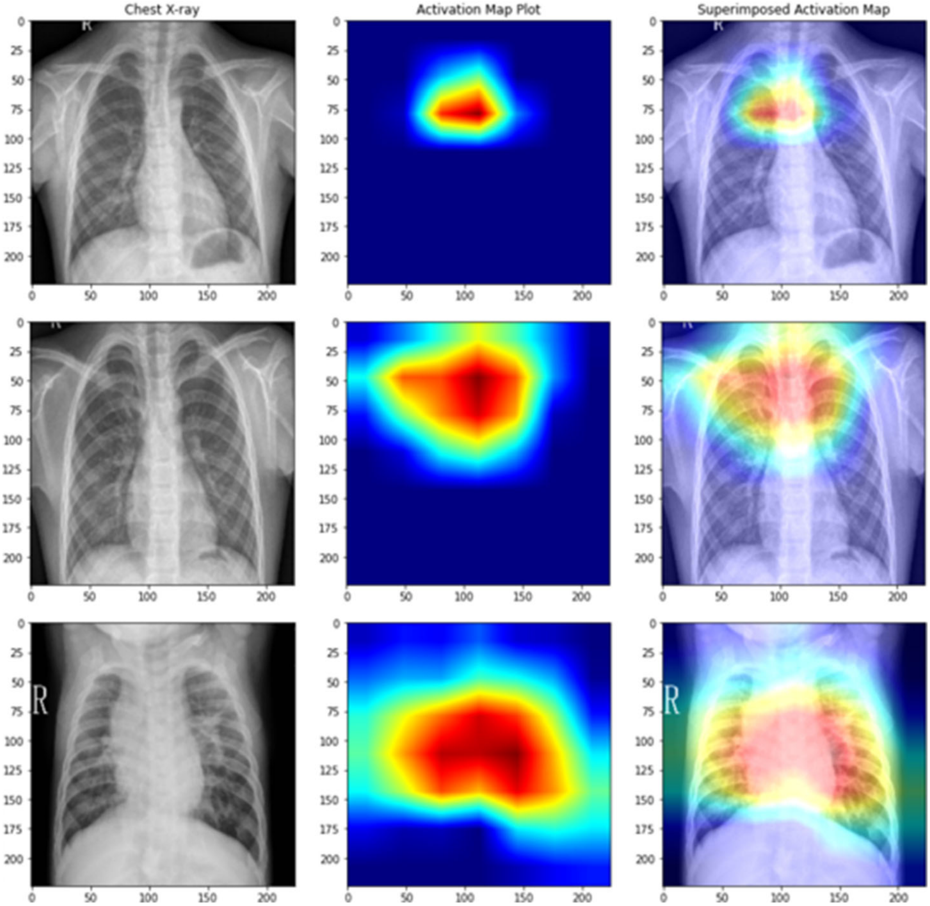


Fig. 28 Class activation maps of misclassified X-rays from DenseNet201 (row 1: normal classified as pneumonia, row 2: normal classified as pneumonia, row 3: pneumonia classified as normal)

added advantage of using an augmented dataset to increase the number of samples in each category. For predictions, the proposed approach uses Contrast Limited Adaptive Histogram Equalization (CLAHE) enhanced chest X-ray images. Our work uses the existing deep CNN models for feature extraction; visualized using t-SNE feature representations and class activation maps**.** The best-performing models' features are concatenated and sent to the stacking classifier for the final - Normal, Pneumonia classification. Redistribution of the dataset instead of added augmentations to ensure unbiased training was the initial dominant factor for reliable performance. Our work uses transfer learning on pre-trained models to compensate for the availability of a limited dataset and introduces data augmentations to prevent overfitting. The features from MobileNet, DenseNet121, DenseNet169 and DenseNet201 are concatenated for stacked ensemble learning. The advantage of the proposed models for this task in specific has been studied in detail. A stacking classifier covering nearly all machine learning models was employed. Stacking classifier with Stratified K-Fold cross-validation results in an accuracy of 98.62%. The proposed models were tested on other lung disease datasets to validate the performance across unseen data for inference.

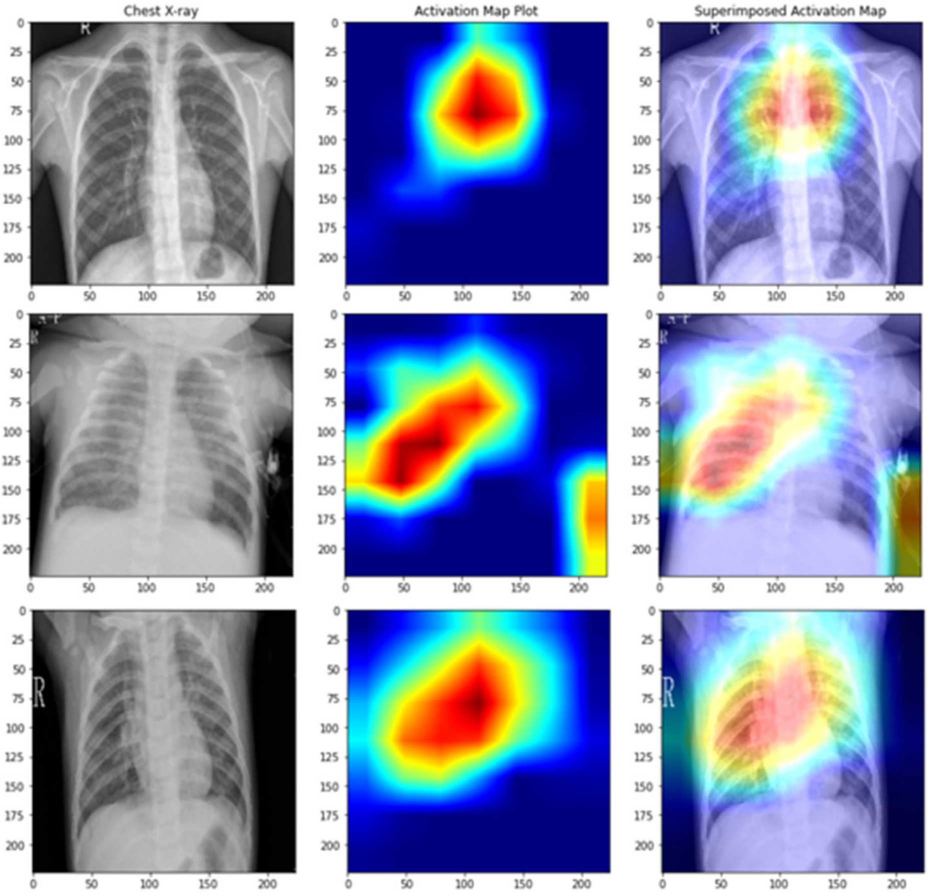


Fig. 29 Class activation maps of correctly classified X-rays from DenseNet201 (row 1: normal classified as normal, row 2: pneumonia classified as pneumonia, row 3: pneumonia classified as pneumonia)

As future work, we would like to explore more about other histogram equalization techniques such as Gamma Correction (GC), Multipeak Histogram Equalization (MPHE), and Multipurpose Beta Optimized Bihistogram Equalization (MBOBHE) for pediatric pneumonia prediction. We notice minor outliers and feature overlap between normal and pneumonia chest X-rays in the t-SNE plots. The t-SNE plots conclude the need for a better architecture to capture many more intrinsic patterns. Introducing attention-based networks and transformers is another possible future direction. Attention-based concatenation instead of simple feature concatenations can be utilized. Augmentations for training might help the model perform much better and reduce the current misclassification rate of 1.38%. With our work performing better CheXNet, which has surpassed normal radiologist level categorization, it will be of immense help to all physicians and radiologists for accurate diagnoses in a matter of seconds. This early detection will help reduce the mortality rate of children suffering from pneumonia.

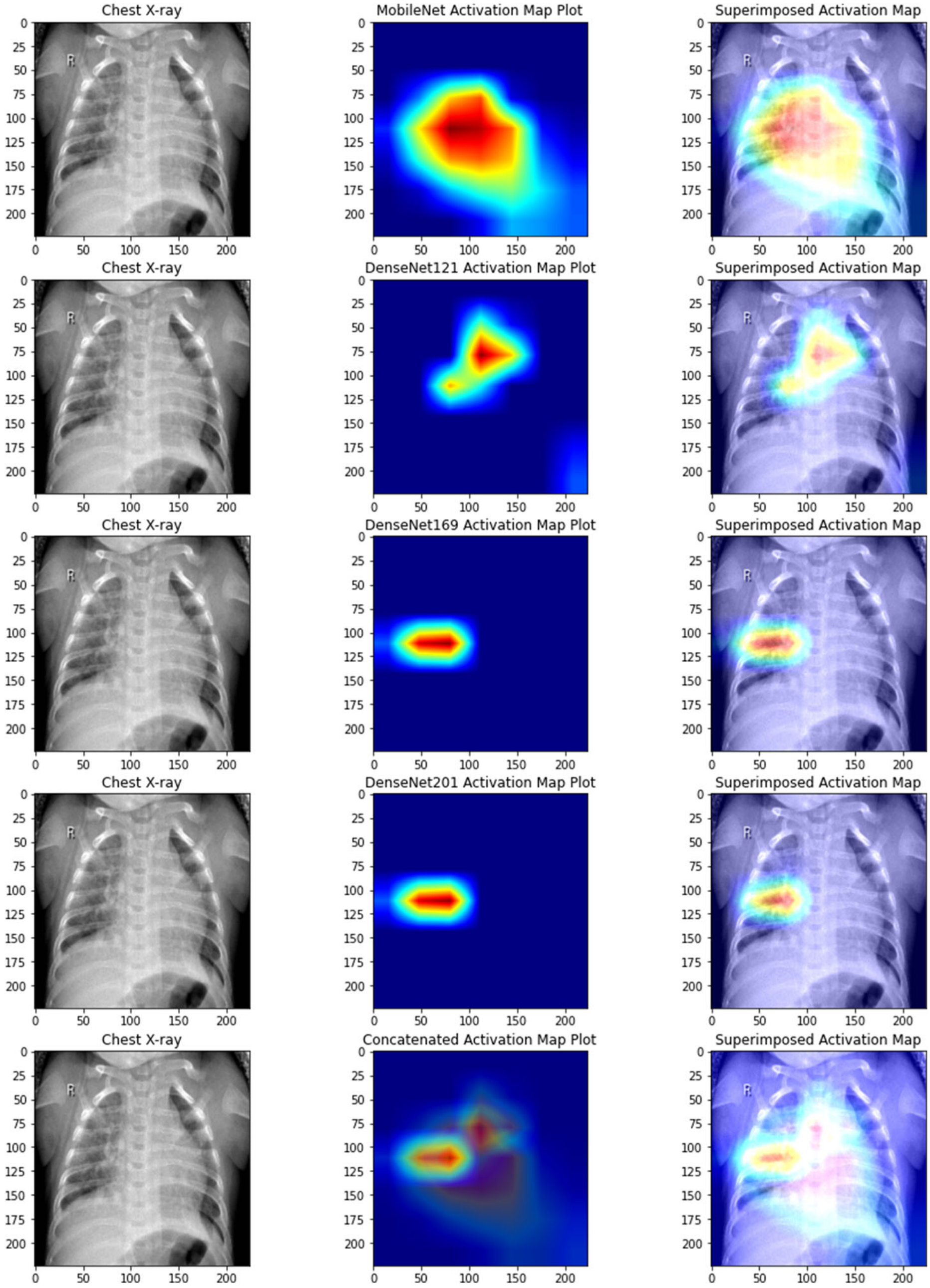

Chest X-ray
MobileNet Activation Map Plot
Superimposed Activation Map
Chest X-ray
DenseNet121 Activation Map Plot
Superimposed Activation Map
Chest X-ray
DenseNet169 Activation Map Plot
Superimposed Activation Map
Chest X-ray
DenseNet201 Activation Map Plot
Superimposed Activation Map
Chest X-ray
Concatenated Activation Map Plot
Superimposed Activation Map
0
25
50
75
100
125
150
175
200
0
50
100
150
200
R

Fig. 30 Class activation maps of an X-ray with pneumonia misclassified as normal from each of the selected architectures. As seen the CAM of the final concatenated feature map is spread over the entire image due to the drawback of using simple feature concatenation techniques

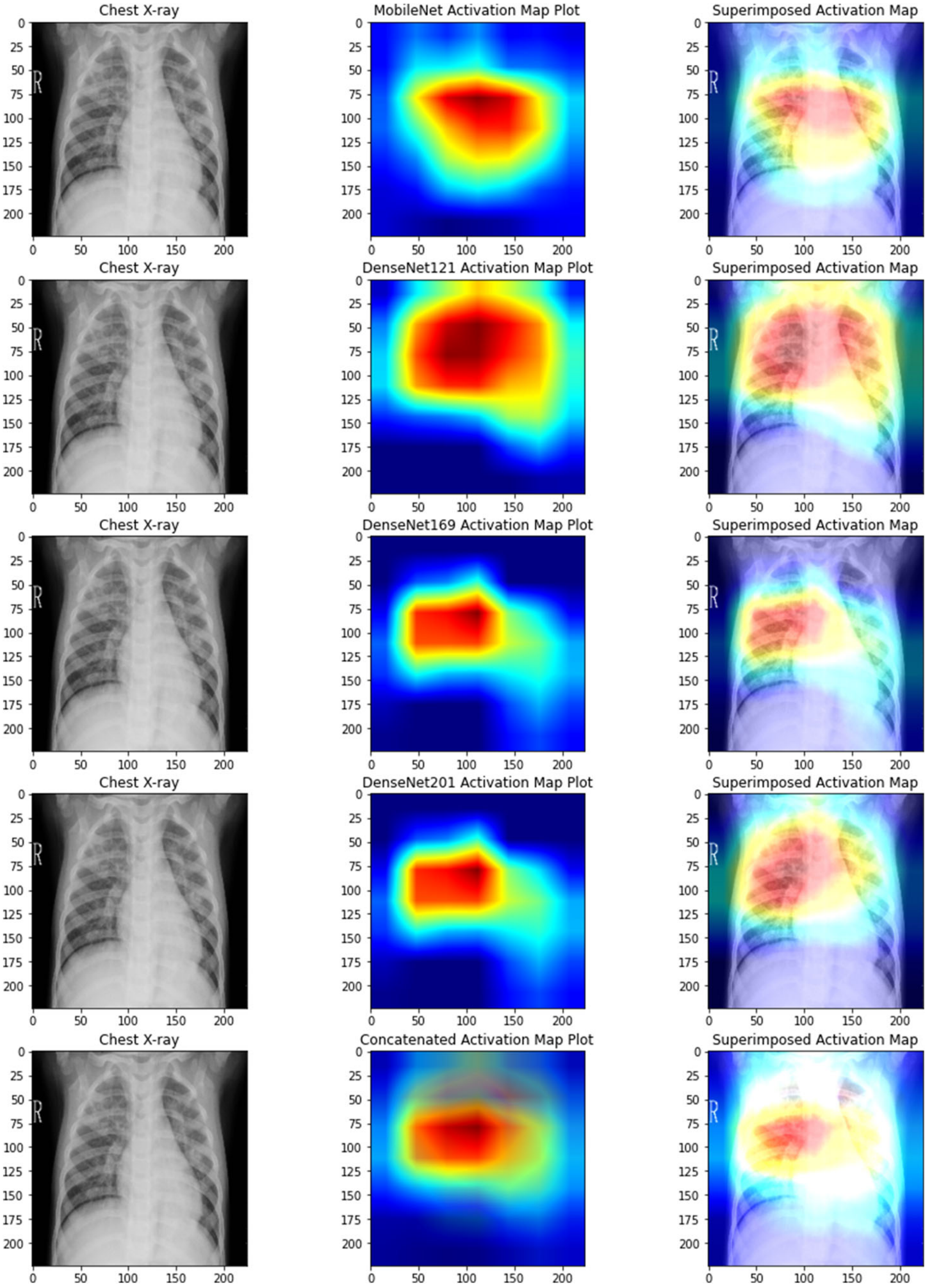

Chest X-ray
MobileNet Activation Map Plot
Superimposed Activation Map
Chest X-ray
DenseNet121 Activation Map Plot
Superimposed Activation Map
Chest X-ray
DenseNet169 Activation Map Plot
Superimposed Activation Map
Chest X-ray
DenseNet201 Activation Map Plot
Superimposed Activation Map
Chest X-ray
Concatenated Activation Map Plot
Superimposed Activation Map
R
0
25
50
75
100
125
150
175
200
0
50
100
150
200

Fig. 31 Class activation maps of an X-ray without pneumonia correctly classified as normal from each of the selected architectures. As seen the CAM of the final concatenated feature map is concentrated to a specific part of the entire X-ray

Table 8 Performance comparison of different machine learning classifiers with the stacking classifier with values rounded off to the nearest two decimal positions

Classifier	Accuracy	Precision	Recall	F1-score	AUC
Logistic regression	97.42	99.96	97.28	98.60	98.37
Support vector classifier	97.75	99.92	97.67	98.78	98.29
Nu- Support vector classifier	98.55	99.69	98.76	99.22	97.19
K-Nearest classifier	97.42	99.92	97.32	98.60	98.11
MLP classifier	97.24	99.96	97.09	98.50	98.27
Gaussian naïve bayes	97.57	99.60	97.78	98.69	96.16
BernoulliNB	96.84	99.60	97.00	98.29	95.77
Gradient boosting classifier	96.99	99.76	97.00	98.36	96.86
XGB classifier	97.28	99.80	97.28	98.52	97.27
DecisionTree classifier	96.52	99.84	96.42	98.10	97.12
RandomForest classifier	97.61	99.88	97.55	98.70	97.96
ExtraTrees classifier	97.71	99.88	97.67	98.76	98.01
Bagging classifier	96.77	99.76	96.77	98.24	96.75
AdaBoost classifier	97.50	99.80	97.51	98.64	97.39
LGB classifier	97.35	99.84	97.32	98.56	97.57
CatBoost classifier	97.53	99.84	97.51	98.66	97.66
HistGradientBoosting classifier	97.17	99.80	97.16	98.46	97.22
Proposed method	98.62	98.99	99.53	99.26	93.17

Table 9 Fine-tuning information and the number of trainable parameters associated with each model used in our study

Classifier	Hyperparameters
Support Vector classifier	C = 2, kernel = ‘poly’, degree = 3, gamma = ‘scale’, coef0 = 0.0, tol = 0.0003
K- Nearest Neighbors classifier	n_neighbors = 175, weights = ‘uniform’, algorithm = ‘brute’,p = 2,leaf_size = 42
Logistic Regression	tol = 3.89253667505987e-05, C = 1, penalty = ‘l2’, max_iter = 102
NuSVC	nu=0.38, kernel=‘rbf’, gamma=‘scale’,tol=2.522320748167019e-05,probability=False
RandomForest	n_estimators = 100, criterion = ‘gini’, min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0, min_impurity_decrease = 0.0, ccp_alpha = 0.0
GaussianNB	var_smoothing = 1e-09
AdaBoostClassifier	base_estimator = DecisionTreeClassifier, n_estimators = 50, learning_rate = 1.0, algorithm = ‘SAMME.R’
BaggingClassifier	base_estimator = DecisionTreeClassifier, n_estimators = 10, max_samples = 1, max_features = 1
ExtraTreesClassifier	n_estimators = 10, criterion = ‘gini’, min_samples_split = 2, min_samples_leaf = 1, max_features = “auto”

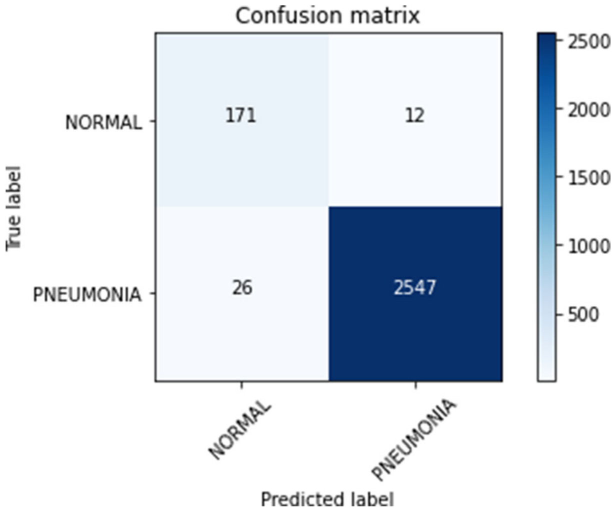


Fig. 32 Confusion matrix for predictions made on the test dataset using the stacked classifier with feature concatenations

Table 10 Performance of other recent works on the Kermany et al. [38] dataset with values rounded off to the nearest two decimal positions

Author	Classes	Technique	Accuracy (%)	Precision (%)	Recall (%)	AUC (%)	Feature Extraction Time
Kermany et al. [38]	Normal and Pneumonia	Inception V3 pretrained CNN model	92.8	90.1	93.2	–	–
Nahida et al. [54]	Normal and Pneumonia	Two-Channel CNN model	97.92	98.38	97.47	97.97	
Stephen et al. [4]	Normal and Pneumonia	Custom CNN model without Transfer Learning	93.73	–	–	–	–
Chouhan et al. [65]	Normal and Pneumonia	Majority voting ensemble model	96.39	93.28	99.62	99.34	–
Rajaraman et al. [58]	Normal and Pneumonia	Custom VGG-16 model	96.2	97.0	99.5	99.0	–
Siddiqi et al. [55]	Normal and Pneumonia	Deep sequential CNN model	94.39	92.0	99.0	–	–
Hashmi et al. [43]	Normal and Pneumonia	Weighted classifier	98.43	–	–	99.76	–
Yu Xiang et al. [34]	Normal and Pneumonia	CGNET	98.72	97.48	99.15	–	–
El Asnaoui et al. [42]	Normal and Pneumonia	Deep CNN model	96.27	98.06	94.61	–	–
Mittal et al. [68]	Normal and Pneumonia	CapsNet architecture	96.36	–	–	–	–
Rahman et al. [66]	Normal and Pneumonia	Deep CNN model	98.0	97.0	99.0	98.0	–
Sagar Kora Venu et al. [9]	Normal and Pneumonia	Weighted average ensemble model	98.46	98.38	99.53	99.60	–

Table 10 (continued)

Author	Classes	Technique	Accuracy (%)	Precision (%)	Recall (%)	AUC (%)	Feature Extraction Time
Toğaçar et al. [11]	Normal and Pneumonia	Deep CNN model	96.84	96.88	96.83	96.80	–
Nahida et al. [47]	Normal and Pneumonia	SMOTE on ensembled features from VGG-19 and CheXNet	98.90	–	–	99.00	–
Islam et al. [10]	Normal and Pneumonia	Feature concatenations with ANN	98.99	99.18	98.90	–	00:04:16
Proposed Work	Normal and Pneumonia	Stacking Classifier based feature concatenations from MobileNet, DenseNet121, DenseNet169 and DenseNet201	98.62	98.99	99.53	93.17	00:01:30

Table 11 Classification report for predictions made on the COVID-19 vs normal vs pneumonia classification dataset [4] using feature fusion from MobileNet, DenseNet121, DenseNet169 and DenseNet201

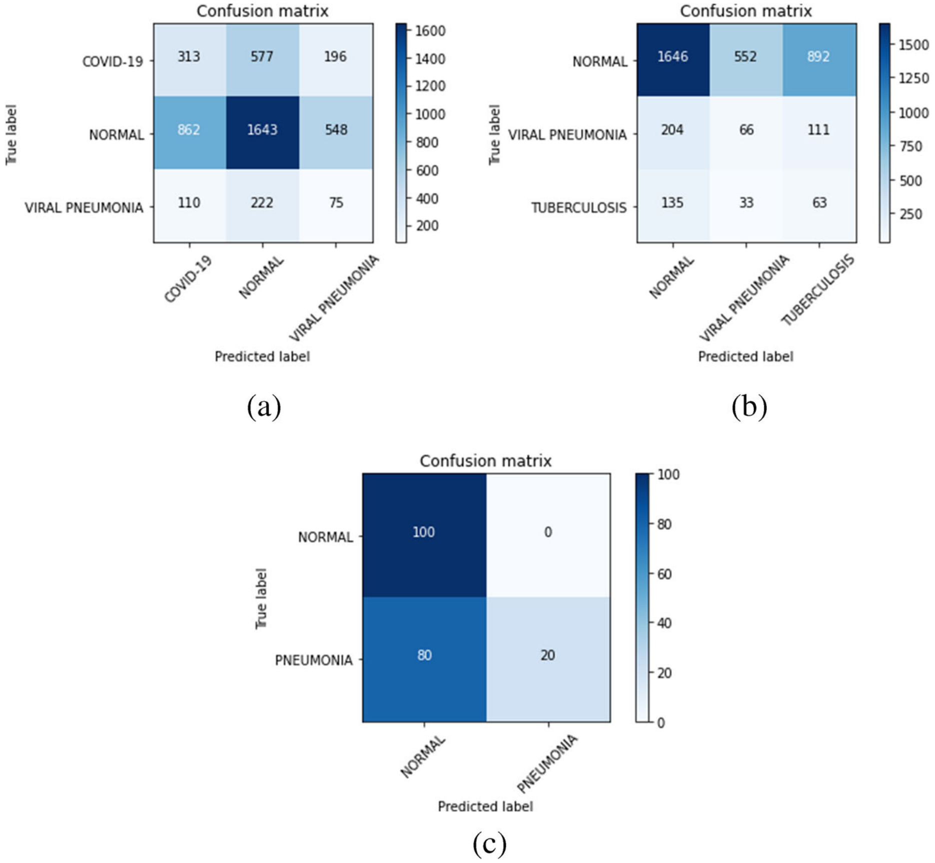
	Precision	Recall	F1-score	Support
COVID-19	0.24	0.29	0.26	1086
NORMAL	0.67	0.54	0.60	3053
VIRAL PNEUMONIA	0.09	0.18	0.12	407
accuracy			0.45	4546
macro avg	0.34	0.34	0.33	4546
weighted avg	0.52	0.45	0.48	4546

Table 12 Classification report for predictions made on the normal vs pneumonia vs tuberculosis classification dataset [4, 27, 58] using feature fusion from MobileNet, DenseNet121, DenseNet169 and DenseNet201

	Precision	Recall	F1-score	Support
NORMAL	0.83	0.53	0.65	3090
VIRAL PNEUMONIA	0.10	0.17	0.13	381
TUBERCULOSIS	0.06	0.27	0.10	231
accuracy			0.48	3702
macro avg	0.33	0.33	0.29	3702
weighted avg	0.71	0.48	0.56	3702

Table 13 Classification report for predictions made on the normal vs pneumonia classification dataset [15, 43] using feature fusion from MobileNet, DenseNet121, DenseNet169 and DenseNet201

	Precision	Recall	F1-score	Support
NORMAL	0.56	1.00	0.71	100
PNEUMONIA	1.00	0.20	0.33	100
accuracy			0.60	200
macro avg	0.78	0.60	0.52	200
weighted avg	0.78	0.60	0.52	200



(a)

(b)

(c)

Fig. 33 Confusion matrix for predictions made on the (**a**) COVID-19 vs normal vs pneumonia classification dataset [4], (**b**) normal vs pneumonia vs tuberculosis classification dataset [4, 58] and normal vs pneumonia classification dataset [43] using the feature fusion approach

Code availability The code is available from the first author upon reasonable request.

Informed consent None.

Data Availability The data that support the findings of this study are available from the first author upon reasonable request.

Declarations

Compliance with ethical standards None.

Conflicts of interest/competing interests The authors declare no conflict of interest.

References

1. Adegbola RA (2012) Childhood pneumonia as a global health priority and the strategic interest of the Bill & Melinda Gates Foundation. Clin Infect Dis 54(suppl_2):S89–S92
2. Asnaoui E, Khalid (2021) Design ensemble deep learning model for pneumonia disease classification. Int J Multimed Inf Retr 10(1):55–68
3. Asnaoui El, Khalid YC, Idri A (2021) "Automated methods for detection and classification pneumonia based on x-ray images using deep learning." Artificial intelligence and blockchain for future cybersecurity applications. Springer, Cham, 257–284
4. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258
5. Chouhan V, Singh SK, Khamparia A, Gupta D, Tiwari P, Moreira C, Damaševičius R, De Albuquerque VHC (2020) A novel transfer learning based approach for pneumonia detection in chest X-ray images. Appl Sci 10(2):559
6. Chowdhury MEH, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, Islam KR, Khan MS, Iqbal A, Emadi NA, Reaz MBI, Islam MT (2020) Can AI help in screening viral and COVID-19 pneumonia? IEEE Access 8:132665–132676
7. Dong S, Wang P, Abbas K (2021) A survey on deep learning and its applications. Comput Sci Rev 40: 100379
8. Gopika P, et al. (2020) "Transferable approach for cardiac disease classification using deep learning."Deep learning techniques for biomedical and health informatics. Academic Press. 285–303
9. Habib N, Hasan Md M, Rahman MM (2020) "Fusion of deep convolutional neural network with PCA and logistic regression for diagnosis of pediatric pneumonia on chest X-rays." Network Biol 76
10. Habib N, Hasan M, Reza M, Rahman MM (2020) Ensemble of CheXNet and VGG-19 feature extractor with random forest classifier for pediatric pneumonia detection. SN Comput Sci 1(6):1–9
11. Hashmi MF et al (2020) Efficient pneumonia detection in chest xray images using deep transfer learning. Diagnostics 10.6:417
12. He, K, et al. (2016) "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition
13. He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. In European conference on computer vision. Springer, Cham, pp 630–645
14. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, ... Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861
15. https://www.kaggle.com/c/detecting-pneumonia-using-cnn-in-pytorch/data?select=chest_xrays. Accessed 10 Feb 2016
16. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition pp 4700–4708
17. Ibrahim AU, et al. (2021) "Pneumonia classification using deep learning from chest X-ray images during COVID-19." Cognit Comput : 1–13
18. Islam KhT, et al. (2020) "A Deep Transfer Learning Framework for Pneumonia Detection from Chest X-ray Images." VISIGRAPP (5: VISAPP)
19. Izadnegahdar R, Cohen AL, Klugman KP, Qazi SA (2013) Childhood pneumonia in developing countries. Lancet Respir Med 1(7):574–584
20. Jadavji T, Law B, Lebel MH, Kennedy WA, Gold R, Wang EE (1997) A practical guide for the diagnosis and treatment of pediatric pneumonia. Cmaj 156(5):703–703
21. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, Dong J, Prasadha MK, Pei J, Ting MYL, Zhu J, Li C, Hewett S, Dong J, Ziyar I, … Zhang K (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172(5):1122–1131
22. Kör H (2022) Hasan Erbay, and Ahmet Haşim Yurttakal. "diagnosing and differentiating viral pneumonia and COVID-19 using X-ray images.". Multimed Tools Appl:1–17
23. Kundu R et al (2021) Pneumonia detection in chest X-ray images using an ensemble of deep learning models. PLoS One 16(9):e0256630
24. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
25. Leung NH (2021) Transmissibility and transmission of respiratory viruses. Nat Rev Microbiol 19(8):528–545
26. Liang G, Zheng L (2020) A transfer learning method with deep residual network for pediatric pneumonia diagnosis. Comput Methods Prog Biomed 187:104964
27. Liu Y, Wu YH, Ban Y, Wang H, Cheng MM (2020) Rethinking computer-aided tuberculosis diagnosis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2646–2655
28. Liz H, Sánchez-Montañés M, Tagarro A, Domínguez-Rodríguez S, Dagan R, Camacho D (2021) Ensembles of convolutional neural network models for pediatric pneumonia diagnosis. Future Gener Comput Syst 122:220–233

29. Luján-García JE et al (2020) A transfer learning method for pneumonia classification and visualization. Appl Sci 10.8:2908
30. Mahajan S, Shah U, Tambe R, Agrawal M, Garware B (2019) Towards evaluating performance of domain specific transfer learning for pneumonia detection from x-ray images. In 2019 IEEE 5th international conference for convergence in technology (I2CT). IEEE, pp 1–6
31. Mittal A et al (2020) Detecting pneumonia using convolutions and dynamic capsule routing for chest X-ray images. Sensors 20.4:1068
32. Muhammad Y, et al. (2021) "Identification of pneumonia disease applying an intelligent computational framework based on deep learning and machine learning techniques." Mob Inf Syst 2021
33. Nafi'iyah N, Setyati E (2021) "Lung X-Ray Image Enhancement to Identify Pneumonia with CNN." 2021 3rd East Indonesia conference on computer and information technology (EIConCIT). IEEE
34. Nahid A-A et al (2020) A novel method to identify pneumonia through analyzing chest radiographs employing a multichannel convolutional neural network. Sensors 20.12:3482
35. Neupane B, Jerrett M, Burnett RT, Marrie T, Arain A, Loeb M (2010) Long-term exposure to ambient air pollution and risk of hospitalization with community-acquired pneumonia in older adults. Am J Respir Crit Care Med 181(1):47–53
36. Nguyen H, Huynh H, Tran T, Huynh H (2020) Explanation of the convolutional neural network classifying chest x-ray images supporting pneumonia diagnosis. EAI Endorsed Trans Context-aware Syst Appl 7(21).
37. Nneji GU, Cai J, Deng J, Monday HN, James EC, Ukwuoma CC (2022) Multi-channel based image processing scheme for pneumonia identification. Diagnostics 12(2):325
38. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
39. Perdomo O, Rios H, Rodríguez FJ, Otálora S, Meriaudeau F, Müller H, González FA (2019) Classification of diabetes-related retinal diseases using a deep learning approach in optical coherence tomography. Comput Methods Programs Biomed 178:181–189
40. Puttagunta M, Ravi S (2021) Medical image analysis based on deep learning approach. Multimed Tools Appl 80(16):24365–24398
41. Rahman T (2021) Et al. "exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images.". Comput Biol Med 132:104319
42. Rahman T et al (2020) Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. Appl Sci 10.9:3233
43. Rajaraman S, Candemir S, Kim I, Thoma G, Antani S (2018) Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. Appl Sci 8(10):1715
44. Rajaraman S, Siegelman J, Alderson PO, Folio LS, Folio LR, Antani SK (2020) Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays. Ieee Access 8:115041–115050
45. Rajpurkar P, et al. (2017) "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." arXiv preprint arXiv:1711.05225
46. Ramezani M, Aemmi SZ, Moghadam ZE (2015) Factors affecting the rate of pediatric pneumonia in developing countries: a review and literature study. Int J Pediatr 3.6(2):1173–1181
47. Rubini C, Pavithra N (2019) Contrast enhancement of MRI images using AHE and CLAHE techniques. Int J Innov Technol Exploring Eng 9(2):2442–2445
48. Sahu S, Singh AK, Ghrera SP, Elhoseny M (2019) An approach for de-noising and contrast enhancement of retinal fundus image using CLAHE. Optics Laser Technol 110:87–98
49. Salem N, Malik H, Shams A (2019) Medical image enhancement based on histogram algorithms. Procedia Comput Sci 163:300–311
50. Saraiva AA, et al. (2019) "Models of Learning to Classify X-ray Images for the Detection of Pneumonia using Neural Networks." Bioimaging
51. Saraiva AA, Ferreira NMF, de Sousa LL, Costa NJC, Sousa JVM, Santos DBS, ... Soares S (2019) Classification of images of childhood pneumonia using convolutional neural networks. Bioimaging 112–119
52. Seshu Babu G, et al. (2021) "Tuberculosis Classification Using Pre-trained Deep Learning Models." Adv Autom Signal Process Instrum Control. Springer, Singapore. 767–774
53. Setiawan AW, Mengko TR, Santoso OS, Suksmono AB (2013) Color retinal image enhancement using CLAHE. In International conference on ICT for smart society. IEEE pp 1–3
54. Siddiqi R (2019) Automated pneumonia diagnosis using a customized sequential convolutional neural network. In Proceedings of the 2019 3rd international conference on deep learning technologies, pp 64–70
55. Siddiqi R (2020) Efficient pediatric pneumonia diagnosis using depthwise separable convolutions. SN Comput Sci 1(6):1–15
56. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

57. Sirazitdinov I, Kholiavchenko M, Mustafaev T, Yixuan Y, Kuleev R, Ibragimov B (2019) Deep neural network ensemble for pneumonia localization from a large-scale chest x-ray database. Comput Electr Eng 78:388–399
58. Stephen O, Sain M, Maduh UJ, Jeong DU (2019) An efficient deep learning approach to pneumonia classification in healthcare. J Healthc Eng 2019
59. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
60. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence
61. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. International conference on machine learning. PMLR
62. Toğaçar M, Ergen B, Cömert Z, Özyurt F (2020) A deep feature learning model for pneumonia detection applying a combination of mRMR feature selection and machine learning models. Irbm 41(4):212–222
63. Trivedi M, Gupta A (2021) A lightweight deep learning architecture for the automatic detection of pneumonia using chest X-ray images. Multimed Tools Appl 81:1–22
64. Veetil IK, et al. (2021) "Parkinson’s Disease Classification from Magnetic Resonance Images (MRI) using Deep Transfer Learned Convolutional Neural Networks." 2021 IEEE 18th India Council International Conference (INDICON). IEEE
65. Venu SK (2020) "An ensemble-based approach by fine-tuning the deep transfer learning models to classify pneumonia from chest X-ray images." arXiv preprint arXiv:2011.05543
66. Wu H, Xie P, Zhang H, Li D, Cheng M (2020) Predict pneumonia with chest X-ray images based on convolutional deep neural learning networks. J Intell Fuzzy Syst 39(3):2893–2907
67. Yadav P, Menon N, Ravi V, Vishvanathan S (2021) Lung-gans: unsupervised representation learning for lung disease classification using chest ct and x-ray images. IEEE Transactions on Engineering Management
68. Yu X, Wang S-H, Zhang Y-D (2021) CGNet: a graph-knowledge embedded convolutional neural network for detection of pneumonia. Inf Process Manag 58(1):102411

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.