ISSN: 2093-5374 / E-ISSN: 2093-5382

An Optimized Hybrid Deep Ensemble Multi Classifier Model for Lung Disease Classification

V.G. Sreena[1*], Dr.D. Narain Ponraj[2], P.L. Deepa[3], Dr. Xiao-Zhi Gao[4], and Dr. Tony Jose[5]

[1*]Research Scholar, Department of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India; Assistant Professor, Department of Electronics and Communication Engineering, Marian Engineering College, Trivandrum, Kerala, India. sreenavg.ec@marian.ac.in, https://orcid.org/0000-0002-5910-079X

[2]Assistant Professor, Department of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India. narainpons@karunya.edu, https://orcid.org/0000-0001-5531-8313

[3]Assistant Professor, Department of Electronics and Communication Engineering, Mar Baselios College of Engineering and Technology, Trivandrum, Kerala, India. deepa.pl@mbcet.ac.in, https://orcid.org/0000-0002-6028-9374

[4]Professor, School of Computing, University of Eastern Finland, Kuopio, Finland. xiao-zhi.gao@uef.fi, https://orcid.org/0000-0002-0078-5675

[5]Assistant Professor, Department of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India. tonyjose@karunya.edu, https://orcid.org/0000-0001-8458-7029

Received: February 14, 2026; Revised: March 23, 2026; Accepted: May 11, 2026; Published: June 30, 2026

Abstract

Evolution of artificial intelligence has significant impact on medical field, especially in disease classification. Lung disease often referred as respiratory infections, sound severe threat to human life. Early interpretation is very much essential for patient survival. Deep network models, especially the technique of ensemble aid the practitioners in automating disease prediction process. In this work, various ensemble strategies such as average, weighted average and stacked ensemble model for four class lung disease classification are proposed. Deep convolutional neural network (CNN) models namely Resnet101, Inception V3 and Mobilenet V2 are used as first level or base models for the ensemble. To study the potency of different meta learners on ensemble stacking, five machine learning models namely Logistic Regression, Support Vector Machine (SVM), Linear Discriminant Analysis, Decision Tree and Naive Bayes are opted as the meta learner models. Here, QaTa-COV 19 dataset is utilized which comprises of 21165 chest x-ray images under four classes - covid19, lung opacity, healthy and viral pneumonia. The experimental results demonstrated that proposed average, weighted average and deep stacked ensemble models outperform, the base predictors and among best, stacked ensemble with support vector machine as meta learner with an overall model accuracy of 98.6%. The proposed model could reduce the variations in heterogeneous base model

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 2 (June-2026), pp. 324-344. DOI: *10.58346/JOWUA.2026.I2.018*

*Corresponding author: Research Scholar, Department of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu, India; Assistant Professor, Department of Electronics and Communication Engineering, Marian Engineering College, Trivandrum, Kerala, India.

predictions and resulted in a highly efficient stacking ensemble model with better generalization capability for lung disease prediction.

Keywords: Classification, Stacked Ensemble, Lung Disease, Deep Learning, Hybrid Model, Average and Weighted Average Ensemble.

1 Introduction

Lungs are the vital part of complex respiratory system which help us to breathe every day. Infections caused to the lungs may be because of many reasons, especially by smoking, forms of air pollution and microorganisms. Various parts of the lungs such as airways, air sacs, interstitium, pleura, blood vessels may be affected by infections ranging from a common cold to deadly diseases. Lung infections like Tuberculosis, Pneumonia, Lung cancer, COVID-19, Chronic obstructive pulmonary disease (COPD) have always been a threat to human life. According to WHO (World Health Organization), respiratory illness is one among the leading causes of death worldwide. The crisis is intensified in developing countries due to the lack of adequate healthcare funding and supporting medical infrastructure. The world has undergone a pandemic crisis due to COVID-19, which accounts for 6.98 million deaths globally.

Conventional methods like Pulmonary function tests, Bronchoscopy, Blood test, CT scan, Chest X-Ray etc. can be used for identification of lung diseases. The efficiency of the test result always depends upon the quality of the sample taken and sometimes yield to false results. Chest X-Ray (CXR) has been used mostly for diagnosis due to its convenience, cost effectiveness as well as low radiation dose as compared to other imaging methods. Expert radiologists are needed to analyse and detect anomalies from the Chest X-Ray report and to provide timely treatment. Interpreting lung disorders correctly from Chest X-Ray is a major challenge to medical practitioners. Deep learning is considered as a sub-field of Machine learning that works under the principle of Artificial Neural Networks (ANN), designed to perform specific tasks. Invention of computationally excellent machines resulted in developing complex deep architectures that can mimic human brain. Some of the well-known deep learning initiatives include Autonomous cars, Virtual assistants, Translators, Face recognition systems, Disease diagnosis etc.

A powerful technique called transfer learning, allows the researchers to reuse the knowledge procured from already built deep pre-trained models for a relatively different task. It also helps to reduce overall training time and improves the model's generalization capability. Negative transfer and over fitting are the limitations of transfer learning. These challenges can be well addressed, and make transfer learning a breakthrough in the field of deep learning. Researchers also demonstrated the benefit of ensemble learning that integrates well performing pretrained models and achieves high accuracy. The outputs from multiple deep learning models at various stages can be combined to form a best performing ensemble model, that produce best results than parent models. Ensemble reduces the variance in the prediction errors caused by individual models. Bagging (Eg. Random Forest, Bootstrap), Boosting (Eg. Gradient boosting, AdaBoost) and Stacking are the common ensemble techniques adopted in research works for reducing the model error and improving the generalization, thus creating a stronger model that accelerates the predictions by individual model. Maximum voting, Averaging and Weighted averaging can be used for aggregating the individual predictions.

The taxonomy for lung disease classification is represented in figure. 1, which describes a three- stage pipeline for analysing the chest x-ray images and an insight to various chest x ray datasets. Modality specific knowledge transfer and ensemble learning demonstrated worthier results as compared

to other techniques and resulted in solving issues related to overfitting and improved generalization capability. The proposed work demonstrates the efficiency of hybrid deep ensemble model, integrating deep learning base learners and machine learning meta learners, for Chest X-Ray based lung disease classification.

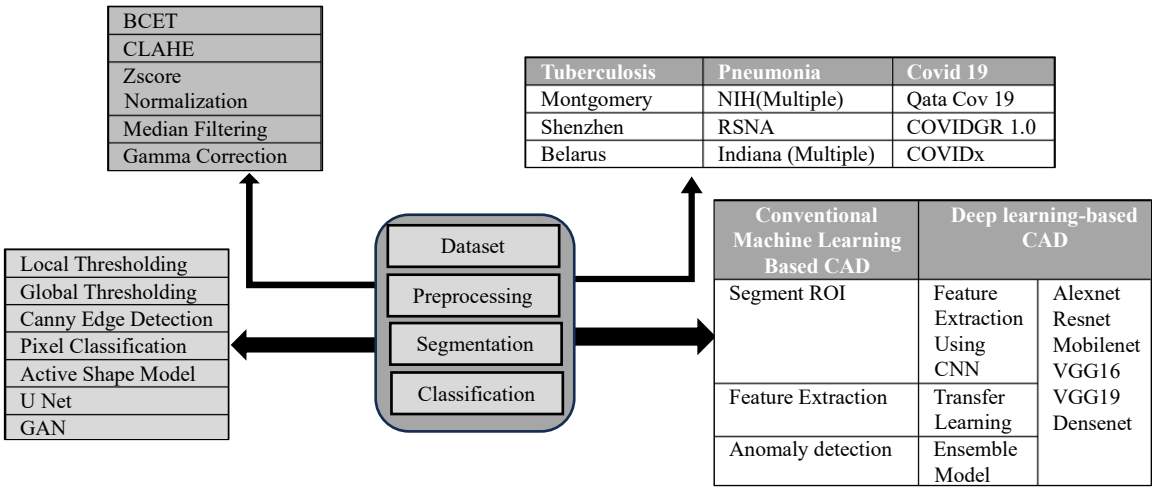


Figure 1: Taxonomy for deep learning based lung disease classification system

The key highlights of the proposed work include

i. Efficient ensemble methods for multi class lung disease classification.

ii. State of the art deep and machine learning models are used as first level (base) and second level (meta) models in ensemble.

iii. Meta models are trained with diversified feature vectors which ensures efficient generalization capability of the model.

iv. Quantitative analysis demonstrates that the proposed model outperforms various state of the art deep neural network models in multi class lung disease classification.

This paper is structured as follows. Section 2 outlines a comprehensive review of related literature. Section 3 demonstrates the proposed average, weighted average and stacked ensemble methodologies, Section 4 details the experimental results and section 5 discuss limitations and future scope and section 6 describes conclusion.

2 Literature Review

Hybrid models incorporating deep learning and machine learning (ML) based feature extractors and classifiers are proposed by researchers (Alshmrani et al., 2023; Emara et al., 2023). There might be performance variations among deep neural networks on same dataset. The technique of ensemble learning helps to combine multiple base learners with varying performance, and suitably selecting a meta learner for enhanced prediction ability. Diverse base model predictions are thus integrated into a single meta model prediction. Two ensemble strategies opted are homogeneous and heterogeneous (Mohammed & Kora, 2023). Homogeneous ensemble aggregates same kind of classifiers built on different data, whereas heterogeneous ensemble utilizes different kind of classifiers built on same data.

Francesco Prinzi et al conducted a detailed study on different deep and shallow classifiers for medical image analysis (Prinzi et al., 2024). Rianne Kablan et al. explored the efficiency of heterogeneous

stacked ensemble with machine learning based base and meta learners for clinical patient data (Kablan et al., 2023). A deep CNN based stacked ensemble model is used for paediatric pneumonia classification, employing kernel principal component analysis for dimension reduction of deep feature vector and stratified k-fold cross validation (Prakash et al., 2023).

A multi-channel Efficientnet is utilized for lung disease detection, which marked high efficiency when detecting tuberculosis, covid-19 and paediatric pneumonia against healthy cases (Ravi et al., 2023). There might be high variance in the performance of base learner models on unseen data. Jayant Bokefode et al. and Weiqiu Jin et al., demonstrated different ensemble learning strategies to reduce variance among base learner predictions (Bokefode & Rao, 2022; Jin et al., 2021). Kim et al. (2024) used an adaptive stacked ensemble model for covid19 severity detection on clinical patient data under three categories. The authors presented a greedy search algorithm to automate the selection of optimal base and meta classifiers. An enhanced kernel convolution for CNN stacking is proposed for multi classification of lung infections (Bhimavarapu et al., 2023).

Sampangi Rama Reddy et al., conducted experiments with a stacked neural network architecture for lung cancer prediction from segmented lung regions using CT scan data (Sampangi Rama Reddy et al., 2024). A stacked deep ensemble model is proposed for covid-19 detection from symptoms dataset and chest x ray images (AlMohimeed et al., 2023). Here the authors demonstrated the model effectiveness by using machine learning base models for symptoms dataset and deep CNN models for chest x ray dataset. Both methods utilized support vector machine as the meta learner.

An optimal ensemble model with weighted local gabour binary pattern for feature pattern retrieval, is proposed by authors for lung disease prediction from chest x-ray images (Ishwerlal et al., 2024). Farhan & Yang, (2023) demonstrated a 2D convolutional neural network model incorporating minmax scaling for feature extraction and optimization, and feature classification by machine learning classifiers (Farhan & Yang, 2023). A hybrid deep learning framework exemplified the superiority of ensemble deep learning over transfer learning models for both augmented and non-augmented datasets (Dubey et al., 2023). A custom CNN network with quantum classifiers is presented in (Rao et al., 2024). The quantum classifiers in this work are formed with encoding, measurement and entanglement properties.

The discussed works demonstrate the efficacy of ensemble models in disease classification. Ensemble approach improves generalization and is less susceptible to overfitting, thus ensures robustness across disparate datasets. Random selection of base learners may lead to weak models which limits the overall performance of ensemble model. Proposed stacked ensemble method utilizes transfer learning for base learner selection, thus possessing highly accurate and stable pre-trained base learners. Thus, the final meta learner model is more efficient in fixing individual model errors when compared to ensembles of randomly initiated base models. The significance of the proposed method is highlighted by comparing it with state-of-the-art deep learning models.

3 Methodology

Dataset

QaTa COV-19 dataset, devised by a team of researchers of Qatar University, includes 21165 chest x-ray images categorised as covid-19, lung opacity, healthy and viral pneumonia cases (Degerli et al., 2021).

The distribution of dataset is represented in table 1, as four categories namely covid 19 (3616 images), lung opacity (6012 images), normal (10192 images) and viral pneumonia (1345 images). The dataset is splitted into train, validation and test data in the ratio 8:1:1. Training data includes 2893 Covid-19,

1076 viral pneumonia, 8154 normal and 4810 lung opacity images. As the dataset is imbalanced with minimum positive cases, the images are undergone data augmentation to distribute the classes equally.

Table 1: Dataset distribution for the proposed model

Dataset	Disease	Train	Validation	Test
QaTa COV-19	Covid-19	2893	362	362
	Lung opacity	4810	601	601
	Normal	8154	1019	1019
	Viral Pneumonia	1076	134	134

Base Learner Selection from Deep Network Models

Deep learning models are usually incorporated to attain high performance on very challenging tasks. As the depth of hidden neurons increase so as the number of trainable parameters. The knowledge gained through a prior assignment by a model can be utilized to solve another target problem of same kind. Such pretrained models like VGG16, VGG19, AlexNet, Resnet50, InceptionV3, Mobilenet, XCeption are trained using Imagenet database containing millions of images categorized as 1000 classes. As deep learning algorithms demand huge dataset for training, and collection of such a labelled dataset seems to be a tedious task, pretrained models can be used to moderate that burden. In the technique of transfer learning, pretrained models are refined according to the input and output pair data for the new specified task.

Table 2: Characteristics of deep learning models

Network	Conv layer depth	Input size	Parameters (million)	Network size (in MB)
Resnet50	50	224x224	25.6	98
Densenet201	201	224x224	20.0	80
Resnet101	50	224x224	44.6	171
InceptionV3	48	299x299	23.9	92
Incep.ResV2	164	299x299	55.9	215
MobilenetV2	28	224x224	3.4	14

The characteristics of deep learning models such as convolutional layer depth, input image size, number of parameters and network size are represented in table 2. Residual Network or Resnet represents a convolutional neural network architecture (CNN) which works by the principle of stacking different residual blocks. It uses skip connections to remediate vanishing gradient issue. ResNet 101 depicts a 101-layered CNN model (He et al., 2016). The key featured bottleneck design, lessens the number of trainable parameters and computations required, in turn makes training faster. Normally, large size filters demand computationally expensive convolutions. Inception V3 represents a competent classifier model that factorizes large convolutions into small and in turn to asymmetric ones (Szegedy et al., 2015). For alleviating issues on vanishing gradients, this network has auxiliary classifiers for regularization.

MobilenetV2 is a light weight deep learning model designed to remove non-linearity in narrow layers through an inverted residual structure design (Sandler et al., 2019). Densenet 201 utilizes densely packed blocks which interconnect adjacent layers each other. Convolution and pooling layers are succeeded by dense and transition layers (Huang et al., 2017). Each of dense block comprises of batch normalisation, Rectified Linear Unit (ReLU) activation and Convolution layers. Density of looping in between different layers defines the efficiency of the model. The model has a strong feature propagation strategy as learned features in each layer is shared through succeeding layers.

Deep learning techniques demand the need of large amount of data to ensure accurate model performance. Gathering and labelling different sets of data for any particular application seems to be

very hectic. Data augmentation is a method used to artificially increase the size of existing datasets, thus improve the efficacy of deep learning models by reducing over fitting. Geometric transformations like random cropping of images into patches, zooming, horizontal flipping, rescaling, translation, addition of noise, zooming etc. are applied in research works to synthesize more images from the available ones.

The workflow of proposed ensemble approach for multiclass lung disease classification is represented in figure 2. Initially the augmented images in the dataset should be subjected to necessary preprocessing steps to mould the data suitable for building efficient deep learning architectures for feature extraction and classification. Now the processed dataset is splitted into training and testing ones. Different pretrained models utilized in this work include ResNet 50, Resnet 101, DenseNet201, InceptionV3, and Mobilenet V2.

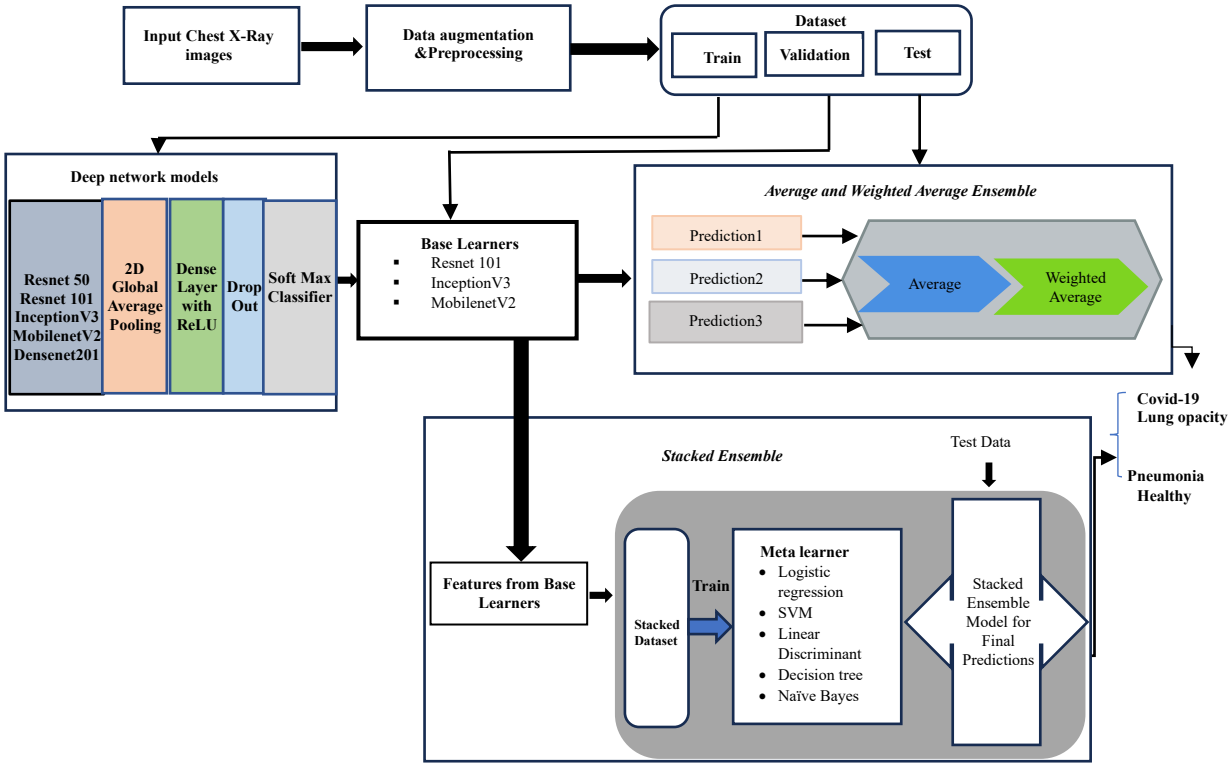


Figure 2: Proposed ensemble model for lung disease classification

During training phase, pretrained model's primary layer features are kept same and final three-layer characteristics are reformed to effectively analyse distinct features for a differently assigned task and dataset. The network hyperparameters are fine-tuned according to the problem domain before being trained. The weights for the initial layers of the pretrained models are kept frozen and are not updated during training. A 2D global average pooling layer, followed by a dense layer with ReLU activation function and a dropout layer are incorporated as the final layers to reduce the problem of overfitting and to minimize associated system loss, thus enhance the overall generalization capability of the model. Global average pooling layer offers excellent feature localization. The model hyper parameters like learning rate, loss function, optimizer and batch size are suitably selected. During testing, performance of the deep CNN models is evaluated using validation data and test data.

Generalization ability of different deep learning models is different based on the underlying features of input data. Thus, there might be huge variance in the prediction results from multiple models. The technique of ensemble learning reduces the prediction errors and variance and thereby improving the

performance of final model. The best classification results on validation samples are obtained for Resnet101, MobilenetV2 and InceptionV3 out of five evaluated models and are identified as the base learners for the ensemble models.

Average and Weighted Average Ensemble Approach

Each time when a model is fitted on same or different training data, the model parameters may be different which leads to different predictions. Thus, fine-tuned pre-trained models result in high variance predictions when used as a single model. The accuracy of individual classifiers always relies on the characteristic feature of input data. Ensemble learning helps in reducing the variance and thereby narrowing the prediction error in the final model.

In this work, out of five deep learning models evaluated, the best performed models such as MobilenetV2, Resnet101 and InceptionV3 are identified as the base learners for the ensemble model. The prediction probability score of softmax classifier of each base learner on test data is a 4-column vector, as it is a four-class problem. In average ensemble classifier, the prediction probability score of base models under each class for each test data are averaged and categorized based on the highest probability class. Let x be the pre-processed input image, $\mathrm{B} = \{\mathrm{BL}_1, \mathrm{BL}_2, \dots, \mathrm{BL}_J\}$ represents the trained base classifier model set and C denotes the number of classes.

The average ensemble prediction is computed as,

$$\widetilde{\mathrm{A}}(x) = \frac{1}{J}\sum_{i=1}^{J} p_i\,(x) \qquad (1)$$

The final predicted class $\hat{y}$ is represented as,

$$\hat{y} = \arg \underset{c\in\{1,2,\dots\mathrm{C}\}}{max} \widetilde{\mathrm{A}}_{\mathrm{c}}(x) \qquad (2)$$

In equation (1), $\mathrm{p}_i(\mathrm{x}) = \{pi1, pi2 \dots piC\}$ represents the classwise probability output for model BL_i.The final predicted class label $\hat{y}$, as described in equation (2), categorize the test image as one among the four classes based on highest average classification probability. In weighted average ensemble approach, weights or priorities are assigned to ensemble members according to model's performance on test data. The weighted predictions are then averaged and classified according to the resultant probability, as one among the four classes. More weightage for the top performing model and less priority for the worst ones. Rather than choosing the weights randomly, a grid search algorithm is utilised to assign weights for the base ensemble members. Let $\mathrm{w} = \{w_1, w_2, \dots, w_J\}$ represents the weights assigned for each base learner model such that $\sum_{i=1}^{J} w_i = 1$.

The weighted average ensemble prediction is calculated as,

$$\widetilde{W_A}(x) = \frac{1}{J}\sum_{i=1}^{J} w_i \cdot p_i(x) \qquad (3)$$

The final predicted class $\hat{y}$ is represented as,

$$\hat{y} = \arg \underset{c\in\{1,2,\dots\mathrm{C}\}}{max} \widetilde{W_A}_{\mathrm{c}}(x) \qquad (4)$$

$\widetilde{W_A}(x)$ in equation (3) represents the average of weighted base learner classification probabilities. The final predicted class label in equation (4), classifies the test image based on highest weighted average classification probability.

Stacking Base Learner Predictions and Meta Learner Training

Stacking is an efficient heterogeneous ensemble learning method used to enhance the generalization ability of disease classification algorithms. Here heterogeneous base learners are trained and tested, and the base learner feature vectors are utilized by a meta learner estimator to boost the overall efficiency of the model. Initially, base learners are chosen from deep network models via transfer learning, based on performance. The base learners might perform differently on same data. A relevant meta leaner model trained on the base learner feature vectors, improves the accuracy of the stacked model, than individual base model predictions.

Algorithm 1: Stacked Ensemble Model Prediction

Input: Chest x-ray images-train (DS_{train}), validation ($DS_{validation}$) and test (DS_{test}) data

Output: Classified labels- Covid-19, Lung opacity, Viral pneumonia, Healthy

Step 1:

Initialisation: Input Train Image (x_i), True label (y_i)

Step 2:

Train pre trained deep CNN models using training dataset $\{x_i, y_i\}$. Evaluate the performance on validation data x_j and choose base learners B for the stacked model.

Step 3:

Extract the feature vectors from the global average pooling layers of base learner models B.

Step 4:

Concatenate the feature vectors fv , to form stacked feature vector v and formulate the meta model training dataset $\{v_j, y_j\}$.

Step 5:

Train the meta model classifier $\overline{ML}$, incorporating 5-fold cross-validation scheme.

Step 6:

Use the stacked ensemble model on test data x_k and evaluate the final predicted label $\hat{y}_k$.

The necessary steps involved in implementing proposed stacked ensemble approach is described in Algorithm 1. Let $\mathrm{DS} = \{(x_i, y_i) \in \mathrm{DS}_{\mathrm{train}}, (x_j, y_j) \in \mathrm{DS}_{\mathrm{validation}}, (x_k, y_k) \in \mathrm{DS}_{\mathrm{test}})\}$ be the dataset containing pre processed chest x ray images and associated labels. Let $\mathrm{B} = \{\mathrm{BL}_1, \mathrm{BL}_2, \ldots, \mathrm{BL}_J\}$ be the set of base learners, $\overline{ML}$ be the meta learner classifier and $\hat{y}$ be the final predicted class label. Each base learner is trained using the images in the training dataset such that the loss function is minimised. For each base learner $BL_k \in \mathrm{B}$,

$$BL_k \leftarrow \arg\min_{\emptyset_k} \sum_{(x_i, y_i) \in DS_{\mathrm{train}}} \mathrm{l}(BL_k(x_i; \emptyset_k), y_i) \qquad (5)$$

In equation (5), $\emptyset_k$ denotes parameters of base leaner, and l is the loss function to be minimised. Now for each sample in the validation dataset $x_j \in DS_{\mathrm{validation}}$, feature vectors are extracted using the selected base learners. Next, the stacked feature set, $v_j = [fv_1(x_j), fv_2(x_j), \ldots, fv_J(x_j)]$ is generated for training the meta learner $\overline{ML}$ as,

$$\overline{ML} \leftarrow \arg\min_{\theta} \sum_{(v_{i'}, y_i)} l(\overline{ML}(v_i; \theta), y_i) \quad (6)$$

θ in equation (6), denotes parameters of the meta-learner. For each test sample $x_k \in DS_{\text{test}}$, the base learner feature vectors are stacked to form, $v_k = [fv_1(x_k), fv_2(x_k), \ldots, fv_J(x_k)]$ and the trained meta learner estimator, $\overline{ML}$ is utilised for final prediction as,

$$\hat{y}_k = \overline{ML}(v_k) \quad (7)$$

where $\hat{y}_k$ in equation (7) represents the final predicted class label for the test image. The popular machine learning models such as logistic regression, support vector machine (SVM), decision tree, linear discriminant analysis and Naive Bayes are chosen as the meta models in this work. Logistic regression represents a supervised machine learning algorithm that predicts a categorical dependent variable from a set of independent variables. It works by one versus all method in multi class classification. Light weight computations and efficient training help to make it a reliable algorithm for classification tasks. Support vector machine (SVM) has exhibited its efficacy in complex classification problems. Support vector machine works by determining an optimal hyperplane that discriminates data points into different classes, ensuring maximum marginal separation among categories. SVM supports multi class classification by prediction through one versus all scheme. It has a powerful regularization method to reduce overfitting and improving generalization ability of the model to unseen data.

Linear discriminant analysis (LDA) is a versatile classification and dimensionality reduction technique which chooses linear combination of features for categorization of input data. It helps to reduce the feature space dimension by preserving inter class separability. Decision tree represents a supervised machine learning algorithm that is used for regression and classification tasks. Decision tree classifies the samples by sorting them down from the root node to leaf node.

Base learner models are selected based on their performance on validation dataset. The selection of meta model is crucial in ensemble technique, so as to reduce the generalization errors associated with base learner predictions. Once the deep CNN base learners in ensemble are finalized, the meta learning model has to be trained relative to base learner predictions. The training dataset for meta learner is constructed from features extracted by the base models on validation data. For this, the feature vectors collected from the 2D global average pooling layer of the base models are stacked together. Once the training dataset is set, the meta learners along with true class labels of validation dataset, are trained using 5-fold cross validation scheme. The trained meta learner along with the base learner members are used for making predictions on the test data.

4 Results

The efficiency of proposed ensemble models depends on the selection of base learner models and meta learner. For stacked ensemble model, we have experimented the performance of five machine learning meta models for the optimized base learners. The training and testing was performed on Intel i7 core processor supported by Nvidia Geforce GTX 1660Ti CUDA enabled GPU using PyTorch backend (python 3.10), executed in the Spyder IDE. Qata Cov-19 dataset with 21165 chest x-ray images splitted among four classes is used in this work. After dividing the dataset into train, validation and test with 8:1:1 ratio, necessary preprocessing is done to ensure suitability of data for deep learning models, thus enhancing the generalization prediction ability of stacked model.

Initially, image denoising is performed to reduce unwanted noise while preserving clinically significant lung regions. The denoised images are then resized to a fixed resolution as demanded by the

deep learning models. Normalization is then applied to standardize the pixel intensity values. Now, data augmentation is performed on the train dataset to ensure equal samples among the classes. For base learner selection, the pre trained models are initialized with Imagenet data weights, fine-tuned and trained for 20 epochs using the training dataset samples. The training time varied according to the associated network parameters. Adam optimizer with L2 regularization, learning rate-0.0001, categorical cross entropy for loss function and batch size of 16 is chosen.

Table 3: Performance metrics of deep learning network models on validation data for base learner selection

Model	Covid-19	Normal	Lung Opacity	Viral Pneumonia	True label predictions Accuracy (%)
Resnet 50	356	563	921	134	93.3
Resnet 101	357	573	945	134	94.9
Inception V3	356	584	946	134	95.5
Mobilenet V2	358	569	973	134	96.1
Densenet 201	351	585	902	134	93.2

Base learners for the proposed model are selected based on the performance of deep CNN models on validation data samples. The performance metrics for the CNN base models on validation dataset is shown in table 3. Out of five models trained, based on the prediction accuracy on validation dataset, three deep CNN models are selected as the base learners for the ensemble models. The best performing models such as Resnet 101, Inception V3, and MobilenetV2 are chosen as the base learners with accuracies 94.9%, 95.5% and 96.1% respectively and helps to extract the deep level features from the dataset. The models are efficient to classify 2009, 2020 and 2034 true labels among the validation data.

Table 4: Performance metrics of base learner models on test data

Network	Class	Precision	Recall	% F1 Score	Accuracy
ResNet101	Covid	98.1	100	99.0	94.6
	Normal	96.7	92.0	94.2	
	Lung opacity	89.2	94.5	91.8	
	Viral Pneumonia	95.0	100	97.4	
Inceptionv3	Covid	99.7	100	99.86	95.4
	Normal	98.8	91.7	95.10	
	Lung opacity	87.7	98.2	92.7	
	Viral Pneumonia	98.5	99.3	98.9	
MobilenetV2	Covid	98.1	100	99.0	95.7
	Normal	96.7	92.0	94.3	
	Lung opacity	89.2	94.5	91.8	
	Viral Pneumonia	95.0	100	97.4	

The performance metrics for the CNN base models on test dataset is represented in table 4. For the test data, the base learners gained 94.6%, 95.4% and 95.7% prediction accuracies respectively.

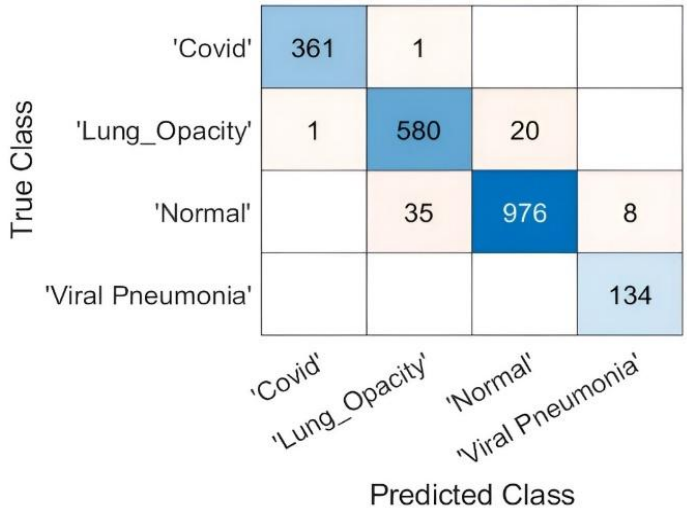
$$Accuracy = \frac{t_N + t_p}{t_p + f_p + t_N + f_N} \quad (8)$$

$$Recall, R = \frac{t_p}{t_p + f_N} \quad (9)$$

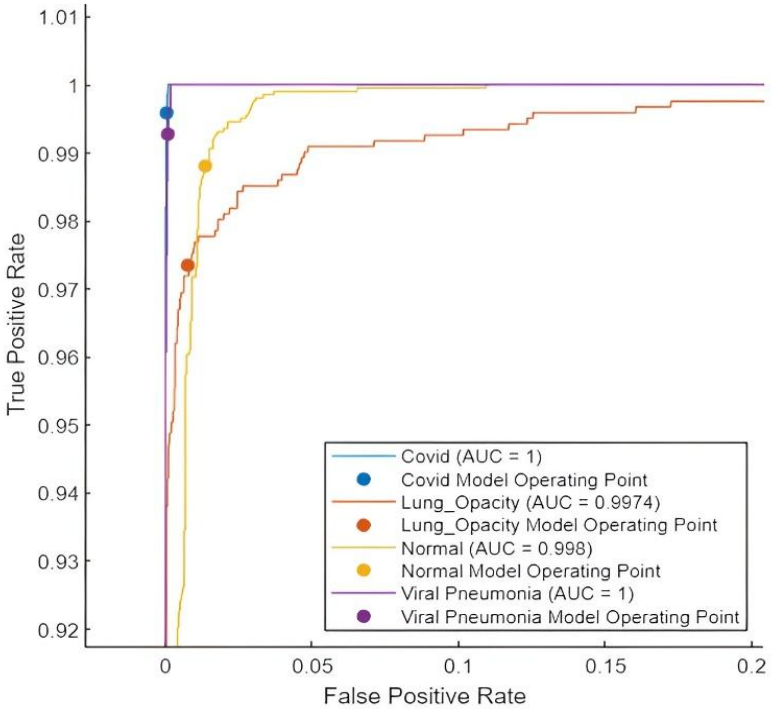
$$Precision, P = \frac{t_p}{f_p + t_p} \quad (10)$$

$$F1\ score = \frac{2PR}{P+R} \tag{11}$$

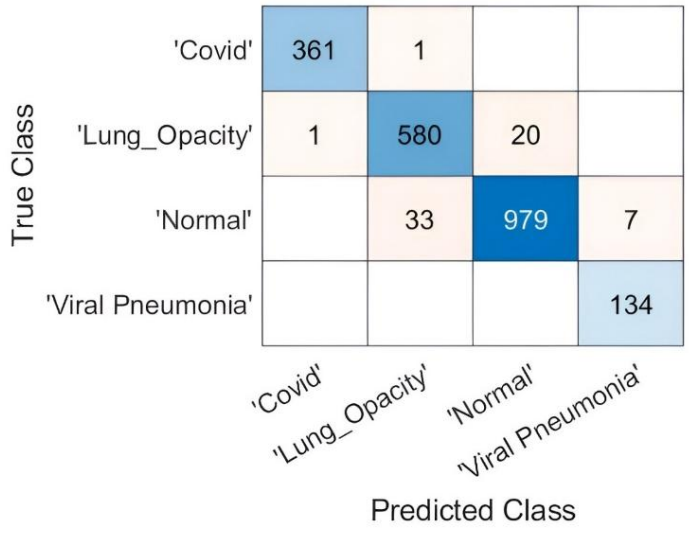
Accuracy, Recall, Precision and F1 score in equation (8), equation (9), equation (10) and equation (11) are the evaluation parameters used for measuring the efficiency of the proposed method, where t_p, t_N, f_p, f_N refer correctly and incorrectly labelled positive and negative predictions.



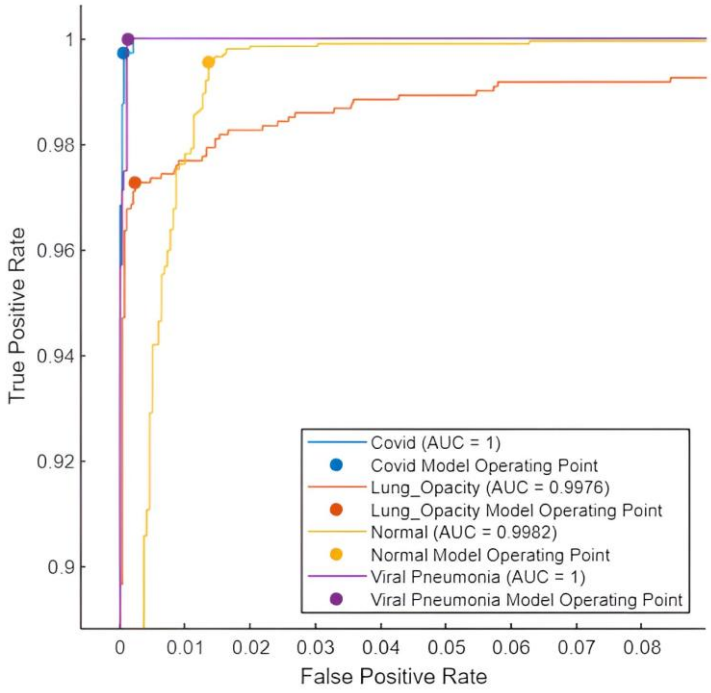
(a)



(b)



(c)



(d)

Figure 3: (a, b) Confusion matrix and ROC curve for average ensemble model (c, d) confusion matrix and ROC curve for weighted average ensemble model

The confusion matrix and the ROC curve for average ensemble model are presented in figures 3(a) and 3(b) respectively. For average ensemble model, the base learner predictions are averaged and 361 covid, 580 Lung opacity, 976 normal and 134 viral pneumonia chest x-rays are correctly classified among 2116 test images. ROC curve is depicted as one positive class versus the rest as negative. Class 0, 1, 2 and 3 represent Covid-19, Lung opacity, Healthy and Viral Pneumonia classes respectively. For Class 0, Covid 19 is taken as positive class with respect to other three classes and Area Under the Curve (AUC) is represented for each category. In weighted average ensemble model, a grid search algorithm

is used for allocating proportional weights for the selected base learners. Best accuracy is obtained for the combination of weights as 0.4,0.3, 0.3 respectively for MobileNetV2, InceptionV3 and ResNet101. The confusion matrix and ROC curve for weighted average ensemble model are presented in figures 3(c) and 3(d) respectively. Weighted average ensemble model is more efficient to classify 361 covid, 580 Lung opacity, 979 normal and 134 viral pneumonia chest x-rays. From the evaluation metrics and ROC curves, it is evident that the false negatives and false positives for viral pneumonia and covid-19 classes are relatively less as compared to other two classes. Also, efficacy of ensemble models in discriminating feature vectors is much improved than individual models.

Table 5: Performance metrics of proposed average ensemble model in multiclass classification

Metrics	**Class 0 (Covid-19)**	**Class 1 (Lung Opacity)**	**Class 2 (Normal)**	**Class 3 (Viral Pneumonia)**
TP	361	580	976	**134**
TN	1753	1479	1077	1974
Precision (%)	99.72	94.15	97.99	94.4
Recall (%)	99.72	96.5	95.78	100
F1 score (%)	99.72	95.31	96.87	97.11
Overall accuracy	96.92%			

Table 6: Performance metrics of proposed weighted average ensemble model in multiclass classification

Metrics	**Class 0 (Covid-19)**	**Class 1 (Lung Opacity)**	**Class 2 (Normal)**	**Class 3 (Viral Pneumonia)**
TP	361	580	979	134
TN	1753	1481	1077	1974
Precision (%)	99.72	94.15	97.99	94.4
Recall (%)	99.72	96.5	96.06	100
F1 score (%)	99.72	95.31	97.01	97.11
Overall accuracy	97.06%			

The performance metrics for the average and weighted average ensemble models on test dataset are presented in table 5 and table 6 respectively. For test data, the average and weighted average ensemble models showed a marginal level of mispredictions among various classes and gained an overall accuracy of 96.92% and 97.06% respectively. For stacked ensemble model, the validation feature vectors from the global average pooling layer of ensemble base models are concatenated to form the training dataset for meta learners. The global average pooling layers of Resnet 101 and Inception V3 output 2048 features while MobilenetV2 outputs 1280 features. The resultant 5376 features along with 2116 class labels of validation data samples are used to train five meta learners namely Logistic Regression, SVM, Decision Tree, Linear Discriminant Analysis and Naive Bayes. A 5-fold cross validation scheme is opted during meta learner training to avoid overfitting. Lasso regularization function is used in logistic regression ensemble model to deal with overfitting. Linear kernel function is employed in SVM as the feature count exceeds training samples. One versus all approach is utilized in logistic regression and SVM for multi class classification. The split criterion applied in decision tree is Gini's diversity index. The trained meta learner along with deep CNN base learners, efficiently classified the test data into four classes with relatively high accuracy than individual base model predictions.

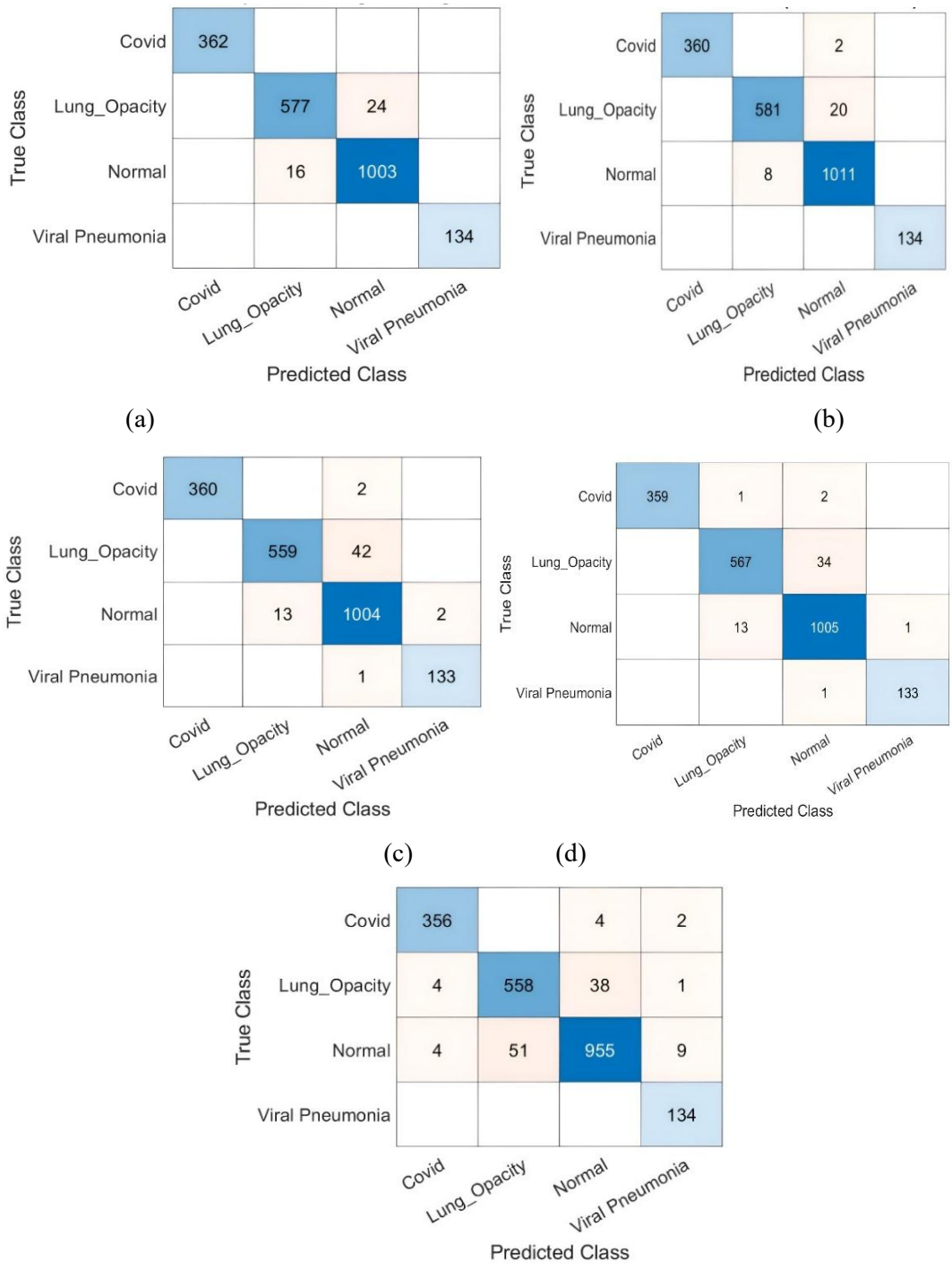


Figure 4: Confusion matrices of stacked ensemble model for a) Logistic regression b) SVM, c) linear discriminant analysis d) decision tree e) naive bayes as meta learners

The confusion matrices for the stacked ensemble models with meta learners such as Logistic Regression, SVM, Linear Discriminant Analysis, Decision Tree and Naive Bayes are presented in figures 4a, 4b, 4c, 4d and 4e respectively. SVM meta classifier correctly classified 360 covid, 581 lung opacity, 1011 normal and 134 viral pneumonia cases among 2116 test images.

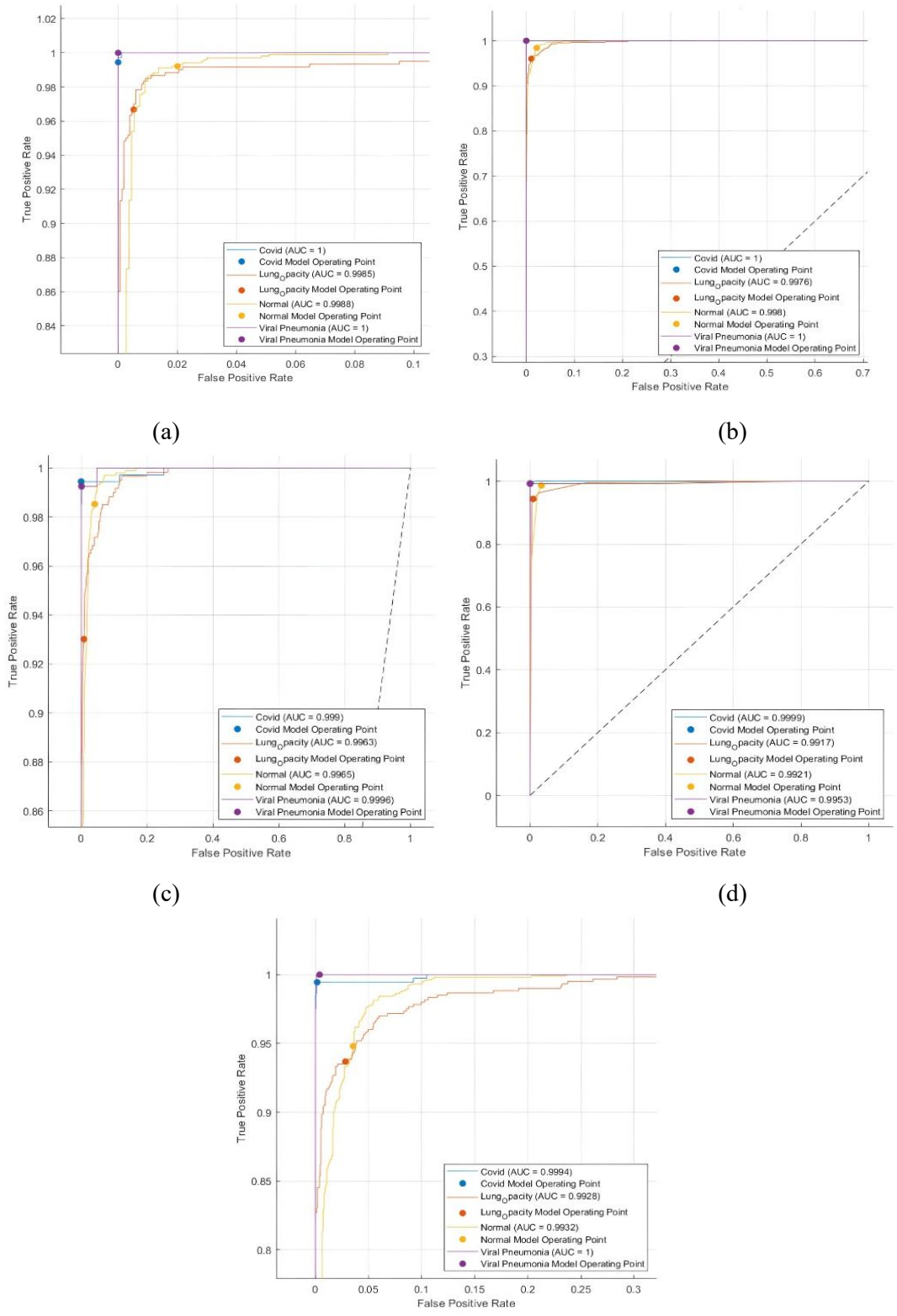


Figure 5: Receiver operating characteristic curves (ROC) of stacked ensemble model for a) Logistic regression b) SVM, c) Linear discriminant analysis d) Decision tree e) Naive bayes as meta learners

The ROC Curves for the stacked ensemble models with meta learners such as Logistic Regression, SVM, Linear Discriminant Analysis, Decision Tree and Naive Bayes are presented in figures 5a, 5b, 5c, 5d and 5e respectively.

In table 7 describe the performance metrics of stacked ensemble classifiers on test data. For the proposed deep stacked ensemble model, support vector machine as meta learner gained the best performance with 98.6% accuracy, while Logistic regression, Linear discriminant analysis, Decision tree and Naive Bayes marked 98.1%, 97.2%, 97.5% and 95.6% accuracies respectively. The comparison of proposed method with existing deep learning models is represented in table 8.

Table 7: Performance metrics of stacked ensemble model on test data

Meta learner	Class	Precision	Recall	% F1 Score	Accuracy
Logistic Regression	Covid	99.7	98.9	99.3	98.1
	Normal	95.9	98.2	97.0	
	Lung opacity	97.1	93.3	95.2	
	Viral Pneumonia	98.5	99.3	98.9	
SVM	Covid	99.7	99.4	99.5	98.6
	Normal	96.5	98.2	97.3	
	Lung opacity	98.6	96.4	97.5	
	Viral Pneumonia	100	99.3	99.6	
LDA	Covid	98.6	99.7	99.1 9	97.2
	Normal	94.4	99.1	6.7	
	Lung opacity	98.7	91.5	95.0	
	Viral Pneumonia	96.9	94.8	95.8	
Decision Tree	Covid	100	100	100	97.5
	Normal	97.8	97.1	97.4	
	Lung opacity	95.7	96.5	96.1	
	Viral Pneumonia	97.8	99.3	98.5	
Naive Bayes	Covid	97.1 96.8	100	100	95.6
	Normal	95.7	97.0	97.3	
	Lung opacity		96.5	96.1	
	Viral Pneumonia	96.8	98.3	96.5	

Table 8: Comparison of proposed method with existing deep CNN models

Network	Accuracy
Resnet 101 (B1) (He et al., 2016)	94.6
Mobilenet V2 (B2) (Sandler et al., 2019)	95.7
Inception V3(B3) (Szegedy et al., 2015).	95.4
Proposed Average Ensemble	96.92
Proposed Weighted Average Ensemble	97.06
Proposed stacked ensemble (B1+B2+B3+ Logistic regression)	98.1
Proposed stacked ensemble (B1+B2+B3+ Naive Bayes)	95.6
Proposed stacked ensemble (B1+B2+B3+ LDA)	97.2
Proposed stacked ensemble (B1+B2+B3+ Decision tree)	97.5
Proposed stacked ensemble (B1+B2+B3+ SVM)	98.6

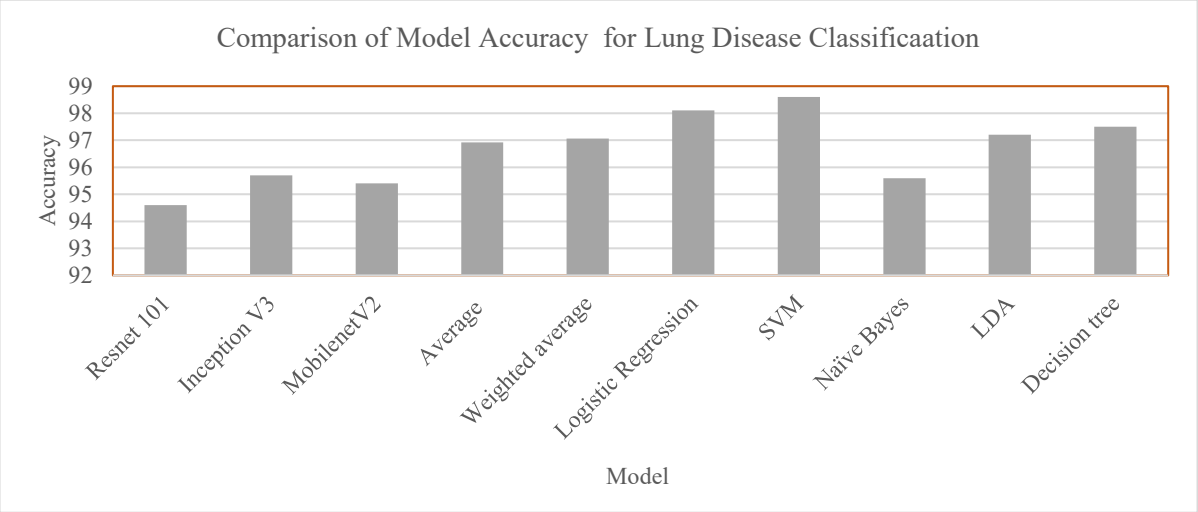


Figure 6: Analysis of model performance in terms of accuracy

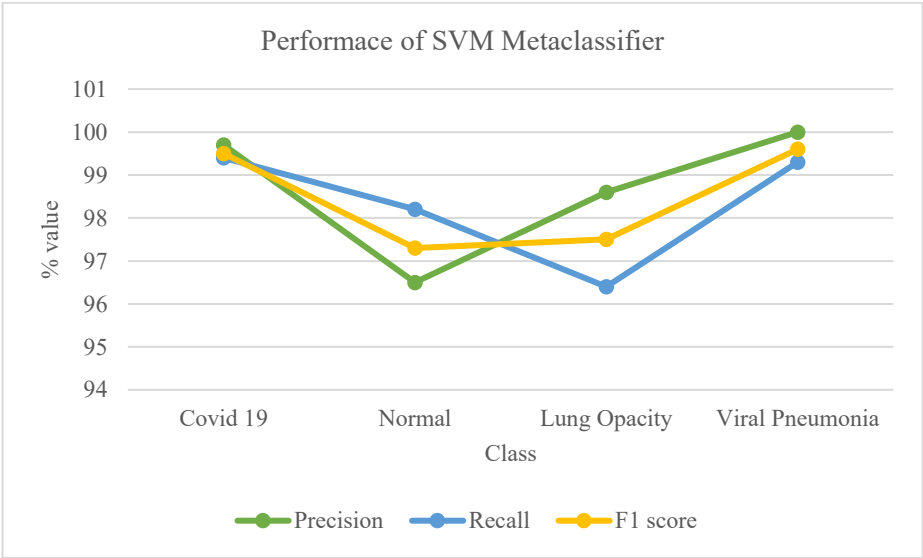


Figure 7: Precision, recall and F1 - score curves for the best performed meta learner SVM on different classes

In figure 6 shows the evaluation of deep network models in terms of accuracy metric. Proposed stacked ensemble model with SVM as meta classifier is superior to other deep CNN models in efficiently classifying chest x-ray images among four classes. Figure 7 shows the recall, precision and F1-score curves for the best performed SVM metaclassifier model.

5 Discussion

Medical experts utilized Computer based disease diagnosis system for better interpretation of lung diseases using Chest X-Rays. Such systems incorporated enhancement techniques to reduce noise and improve the quality of acquired images so as to increase the efficiency of lung disorder identification. Even though deep learning techniques boosted the system efficiency, analysis reveals that the outcome is highly dependent on several factors. Deep learning techniques need sufficient amount of data to be trained by the system to develop fully efficient models. Data disproportion in the available dataset turned

as a case of concern in establishing deep learning algorithms. The relative number of positive disease images are comparatively lower in most dataset which escalate overfitting issues and the developed model becomes biased.

The limited availability of dataset forced the researchers to work with the technique of data augmentation, which allows to create multiples data from the existing ones through rotation, flipping, scaling, translation and so on. As the constructed dataset resembles the original, question arises on the image diversity. Generative Adversarial Network based data augmentation is utilized in many works and proved efficient in gaining high classification accuracy.

Limitations and Future Scope

Ensemble learning that accumulates the performance of single models, proved useful in discriminating features of all different classes in a dataset. The choice of the ensemble members contributes towards the final prediction accuracy. Presupposing the hyper parameters and selecting ensemble members by trial will not be effective as they might not be necessarily trained using same training data as well of same type networks. The option to give percentage of weight to each ensemble member, makes the system reliable. The fusion strategy might not outperform always than individual classifiers that are trained and tested for specific applications. Thus, ensemble model can sometimes lead to lower testing accuracy than the best performing ensemble member. This is because of the technique adapted for combining models. The performance and robustness are the major factors that make ensemble learning a choice for researchers. Thus, efficient ensemble models explore the research opportunity for solving real world problems in medical industry.

Recently, attention networks and vision transformers have shown significant improvement in image classification tasks. Vision transformers utilize the technique of self-attention mechanism and are capable enough to extract long term dependencies among various features of input images. Attention networks give priority to the features from regions of interest and discard redundant information. Several hybrid models are being proposed in the field, thus producing effectual results in disease classification.

6 Conclusion

As the health care sector demands the automation of disease identification system, as with the increase in number of patients and sudden rise in pandemic situations such as COVID19, the research community is heading more towards blooming deep learning techniques. In this work, an efficient four class average, weighted average and stacked ensemble classifiers for lung disease classification is proposed by utilizing transfer learning approach for base learner selection. For stacked ensemble model, initially conventional machine learning meta models are trained using the high-level feature vectors extracted from base learner predictors. Fivefold cross validation is employed during meta learner training. On validating the performance of classifiers with unseen test data, it is evident from the quantitative analysis that, the proposed models outperform individual base predictors in multi class lung disease classification. The top performer is support vector machine (SVM) with 98.6% accuracy. Feature concatenation ensures that meta learners are trained using diversified feature vectors and the prediction error can be further minimised. Thus, the proposed models reduce the variance in heterogeneous base learner predictions and improved the generalization capability in disease prediction. Further to investigate, we plan to explore the potential of hybrid deep learning ensemble architecture employing vision transformer on multi class lung disease prediction.

References

[1] AlMohimeed, A., Saleh, H., El-Rashidy, N., Saad, R. M., El-Sappagh, S., & Mostafa, S. (2023). Diagnosis of COVID-19 using chest X-ray images and disease symptoms based on stacking ensemble deep learning. *Diagnostics*, *13*(11). https://doi.org/10.3390/diagnostics13111968

[2] Bhimavarapu, U., Chintalapudi, N., & Battineni, G. (2023). Multi-classification of lung infections using improved stacking convolution neural network. *Technologies, 11*(5), 128. https://doi.org/10.3390/technologies11050128

[3] Bokefode, J., & Rao, M. P. (2022). Ensemble deep learning models for lung cancer diagnosis in histopathological images. *Procedia Computer Science*, *215*, 471-482. https://doi.org/10.1016/j.procs.2022.12.049.

[4] Degerli, A., Ahishali, M., Kiranyaz, S., Chowdhury, M. E., & Gabbouj, M. (2021, September). Reliable covid-19 detection using chest x-ray images. In *2021 IEEE International Conference on Image Processing (ICIP)* (pp. 185-189). IEEE. https://doi.org/10.1109/ICIP42928.2021.9506442.

[5] Dubey, A. K., Chabert, G. L., Carriero, A., Pasche, A., Danna, P. S., Agarwal, S., ... & Suri, J. S. (2023). Ensemble deep learning derived from transfer learning for classification of COVID-19 patients on hybrid deep-learning-based lung segmentation: a data augmentation and balancing framework. *Diagnostics*, *13*(11), 1954. https://doi.org/10.3390/diagnostics13111954

[6] Emara, H. M., Shoaib, M. R., El-Shafai, W., Elwekeil, M., Hemdan, E. E. D., Fouda, M. M., ... & El-Samie, F. E. A. (2023). Simultaneous super-resolution and classification of lung disease scans. *Diagnostics*, *13*(7), 1319. https://doi.org/10.3390/diagnostics13071319.

[7] Rao, G. E., Rajitha, B., Srinivasu, P. N., Ijaz, M. F., & Woźniak, M. (2024). Hybrid framework for respiratory lung diseases detection based on classical CNN and quantum classifiers from chest X-rays. *Biomedical Signal Processing and Control*, *88*, 105567. https://doi.org/10.1016/j.bspc.2023.105567.

[8] Farhan, A. M. Q., & Yang, S. (2023). Automatic lung disease classification from the chest X-ray images using hybrid deep learning algorithm. *Multimedia Tools and applications*, *82*(25), 38561-38587. https://doi.org/10.1007/s11042-023-15047-z.

[9] Alshmrani, G. M. M., Ni, Q., Jiang, R., Pervaiz, H., & Elshennawy, N. M. (2023). A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images. Alexandria Engineering Journal, 64, 923-935. https://doi.org/10.1016/j.aej.2022.10.053

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90

[11] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp.2261-2269. https://doi.org/10.1109/CVPR.2017.243

[12] Ishwerlal, R. D., Agarwal, R., & Sujatha, K. S. (2024). Lung disease classification using chest X ray image: An optimal ensemble of classification with hybrid training. *Biomedical Signal Processing and Control*, *91*, 105941. https://doi.org/10.1016/j.bspc.2023.105941

[13] Jin, W., Dong, S., Dong, C., & Ye, X. (2021). Hybrid ensemble model for differential diagnosis between COVID-19 and common viral pneumonia by chest X-ray radiograph. *Computers in biology and medicine*, *131*, 104252. https://doi.org/10.1016/j.compbiomed.2021.104252

[14] Kablan, R., Miller, H. A., Suliman, S., & Frieboes, H. B. (2023). Evaluation of stacked ensemble model performance to predict clinical outcomes: A COVID-19 study. *International Journal of Medical Informatics*, *175*, 105090. https://doi.org/10.1016/j.ijmedinf.2023.105090

[15] Kim, G. W., Ju, C. Y., Seok, H., & Lee, D. H. (2024). Adaptive stacking ensemble techniques for early severity classification of COVID-19 patients. *Applied Sciences, 14*(7), 2715. https://doi.org/10.3390/app14072715

[16] Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences*, *35*(2), 757-774. https://doi.org/10.1016/j.jksuci.2023.01.014

[17] Prakash, J. A., Ravi, V., Sowmya, V., & Soman, K. P. (2023). Stacked ensemble learning based on deep convolutional neural networks for pediatric pneumonia diagnosis using chest X-ray images. *Neural Computing and Applications*, *35*(11), 8259-8279. https://doi.org/10.1007/s00521-022-08099-z

[18] Prinzi, F., Currieri, T., Gaglio, S., & Vitabile, S. (2024). Shallow and deep learning classifiers in medical image analysis. *European radiology experimental, 8*(1), 26. https://doi.org/10.1186/s41747-024-00428-2

[19] Ravi, V., Acharya, V., & Alazab, M. (2023). A multichannel EfficientNet deep learning-based stacking ensemble approach for lung disease detection using chest X-ray images. *Cluster Computing*, *26*(2), 1181-1203. https://doi.org/10.1007/s10586-022-03664-6

[20] Sampangi Rama Reddy, B. R., Sen, S., Bhatt, R., Dhanetwal, M. L., Sharma, M., & Naaz, R. (2024). Stacked neural nets for increased accuracy on classification on lung cancer. *Measurement: Sensors*, *32*, 101052. https://doi.org/10.1016/j.measen.2024.101052

[21] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520). https://doi.org/10.1109/CVPR.2018.00474

[22] 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). https://doi.org/10.1109/CVPR.2016.308

Authors Biography

V.G. Sreena received her Bachelor of Technology degree in Electronics and Communication Engineering from Cochin University of Science and Technology, Kerala, India in 2004 and her Post graduate degree in Applied Electronics and Instrumentation Engineering from Kerala University, Kerala, India in 2008. She is currently pursuing her Ph.D. degree in Electronics and Communication Engineering from Karunya University, Tamil Nadu, India. Her research interest includes medical imaging, deep learning and digital image processing. She is presently working as Assistant Professor in Marian Engineering College, Trivandrum, Kerala. She has publications in reputed conference proceedings and has presented research works in various international conferences.

Dr.D. Narain Ponraj received his Bachelor of Technology degree in Electronics and Communication Engineering from Pondicherry University, India in 2005 and his Master of Engineering degree in Applied Electronics from Karunya University, India in 2007. He received his doctoral degree from Karunya University, India in 2017. His research area is medical imaging. He has published several papers in reputed journals and conference proceedings. He has filed three patents. He is now working as Assistant Professor in the department of Electronics and Communication Engineering in Karunya University, India.

P.L. Deepa is currently an Assistant Professor in Mar Baselios College of Engineering and Technology, Trivandrum. She received her postgraduate degree with specialization in Signal Processing from the College of Engineering Trivandrum, Kerala. She is currently pursuing Ph.D. in the area of Medical Image Processing at Karunya Institute of Technology and Science, Coimbatore. She also presented various academic as well as research-based papers at several national and international conferences. She has participated in several competitions including NTIRE 2020 Image and Video Deblurring Challenge. Her areas of interest include Signal Processing, Image Processing, Computer

Vision, Deep Learning, etc. She has a total of 13 years of teaching experience. She is a mentor of the IBM campus programme in Artificial Intelligence and one of the NPTEL translators.

Dr. Xiao-Zhi Gao received his B.Sc. and M.Sc. degrees from the Harbin Institute of Technology, China in 1993 and 1996, respectively. He obtained his D.Sc. (Tech.) degree from the Helsinki University of Technology, Finland in 1999. He has been working as a professor at the University of Eastern Finland, Finland since 2018. He has published more than 450 technical papers in refereed journals and international conferences. His research interests are nature-inspired computing methods with their applications in optimization, data mining, machine learning, control, signal processing, and industrial electronics.

Dr. Tony Jose received his Bachelor of Technology degree in Electronics and Communication Engineering from Mahatma Gandhi University, India, in 2008 and his Master of Technology degree in Microwave and Television Engineering from Kerala University, India, in 2010. He obtained his doctoral degree from Kerala University, India, in 2018. His research interests include Optical Communication, Sensors, and Signal Processing. He is currently working as Assistant Professor in the Division of Electronics and Communication Engineering at Karunya Institute of Technology and Sciences, India.