PYTHON BASED PREDICTION OF PULMONARY EMPHYSEMA

[1]Turubati Saikiran, [2]K.Tulasi krishna kumar,
[1]MCA 2nd Year, [2]Associate Professor,
[1]Master of Computer Applications,
[1]Sanketika Vidya Parishad Engineering College, Visakhapatnam, India

***Abstract*—A respiratory infection brought on by bacteria or viruses, pulmonary emphysema affects a large number of people, especially in developing and underdeveloped countries where high levels of pollution, unhygienic living conditions, and overcrowding are relatively common, along with insufficient medical infrastructure. Pleural effusion is a disorder in which fluids fill the lung and make breathing difficult. It is brought on by pulmonary emphysema. To achieve effective treatment and boost survival chances, pulmonary emphysema must be identified early. The most common way to diagnose pulmonary emphysema is by chest X-ray imaging. The analysis of chest X-rays, however, is a difficult task that is vulnerable to subjectivity. We created a method in this work that uses chest X-ray pictures to automatically computer-aided diagnosis detect pulmonary emphysema. To address the lack of available data, we used deep transfer learning and created an ensemble of three convolutional neural network models, GoogLeNet, ResNet-18, and DenseNet-121. A weighted average ensemble strategy was used, and a unique method was used to decide the weights given to the base learners. Precision, recall, f1-score, and area under the curve scores from four common evaluation metrics are used to create the weight vector, which in studies published in the literature was typically set experimentally, an inefficient procedure. Using a five-fold cross-validation method, the suggested method was tested using two pulmonary emphysema X-ray datasets that were made accessible to the public by Kermany et al. and the Radiological Society of North America (RSNA), respectively. Using the Kermany and RSNA datasets, the suggested technique attained accuracy rates of 98.81% and 86.85%, as well as sensitivity values of 98.80% and 87.02%. The outcomes outperformed those of cutting-edge approaches, and our approach outperformed the widely used ensemble techniques. The datasets statistical evaluations using McNemar's and ANOVA tests demonstrated the robustness of the method.**

***Index Terms*—Pleural effusion, computer-aided diagnosis, convolutional neural network models, evaluation metrics, robustness, weighted average, precision, robustness, the Radiological Society of North America (RSNA).**

I. INTRODUCTION

Acute pulmonary infection known as pulmonary emphysema can be brought on by bacteria, viruses, or fungi and affects the lungs, inflaming the air sacs and leading to pleural effusion, a condition where the lung is flooded with fluid. More than 15% of deaths in children under the age of five are attributed to it. The majority of cases of pulmonary emphysema occur in impoverished and emerging nations, where there is a shortage of medical resources, excessive population, pollution, and unclean environmental conditions. Consequently, preventing the disease from turning fatal can be greatly aided by early diagnosis and care. The diagnosis of lung disorders typically involves radiological evaluation of the lungs using computed tomography (CT)[1],magnetic resonance imaging (MRI)[2], or radiography (X-rays)[3],An affordable, non-invasive way to examine the lungs is via X-ray images. An illustration of a pneumonic and a healthy lung can be seen in Fig. 1. The white infiltrates in the pneumonic X-ray (depicted with red arrows) separate a pneumonic from a healthy state. Yet, subjective variability might occur during chest X-ray scans to detect pulmonary emphysema. As a result, the detection of pulmonary emphysema must be automated. In this study, we created a computer-aided diagnosis (CAD) [4]system that accurately classifies chest X-ray images using an ensemble of deep transfer learning models.

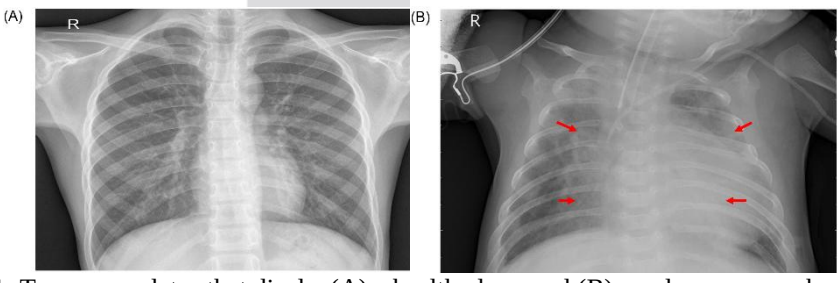

Figure 1: Two x-ray plates that display(A) a healthy lung and (B) a pulmonary emphysema

Deep learning is a powerful artificial intelligence tool that is essential in resolving a variety of challenging computer vision issues. Convolutional neural networks (CNNs)[5]in particular, are widely employed for a variety of image categorization issues. Nevertheless, these models only operate at their best when they are given a lot of data. Such a large volume of labelled data is challenging to obtain for biomedical image classification problems since each image must be classified by qualified medical professionals, which is an expensive and time-consuming process. To get around this problem, use transfer learning. In this method, a model trained on a large dataset is reused and the network weights determined in this model are employed to address

problems involving small datasets. For biomedical image classification applications, CNN models trained on huge datasets like Imagenet[6] which contains more than 14 million image are widely employed.

A common technique is ensemble learning, which combines the conclusions of various classifiers to provide the final prediction for a test sample. It is carried out to extract the discriminative information from each base classifier, producing more precise predictions as a result. Average probability, weighted average probability, and majority voting are some of the ensemble techniques that have been utilised in research in the literature most frequently. Each constituent base learner is given equal priority by the average probability-based ensemble. However for a specific issue, one basic classifier might be better equipped to gather data than another. Hence, weighting all of the base classifiers is a better technique. However for a specific issue, one basic classifier might be better equipped to gather data than another. Hence, weighting all of the base classifiers is a better technique. Yet, the importance of the weights given to each classifier is the most important component in ensuring the ensemble's improved performance. The majority of methods base this number on the outcomes of experiments. In this study, we developed a novel weight allocation technique in which the best weights for three base CNN models— GoogLeNet[7], ResNet-18[8], and DenseNet-121[9],were assigned using four assessment metrics: precision, recall, f1-score, and area under the receiver operating characteristics (ROC) curve (AUC). Studies in the literature often only took classification accuracy into account when deciding how much to weight base learners, which may not be a sufficient criterion, especially when the datasets are class-imbalanced. Better information for prioritising the base learners may come from other indicators. Fig. 2 displays the suggested ensemble framework's overall workflow.

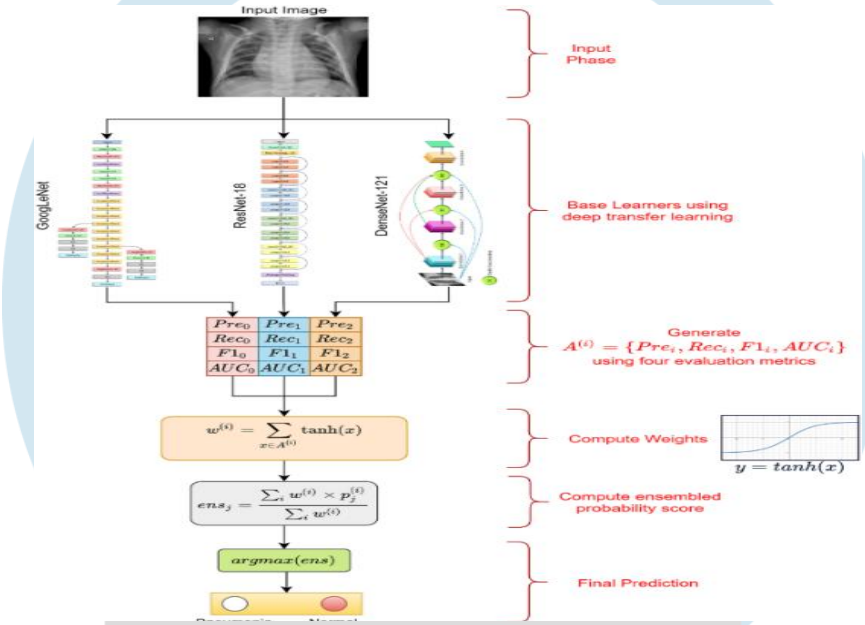

Figure 2: Proposed system of pulmonary emphysema framework

Chest X-ray-based pneumonia identification has long been a challenge, with the main drawback being the dearth of data that is readily accessible to the general public. There has been a lot of research into conventional machine learning techniques[10] .In order to categorise the lung regions from the chest X-ray pictures, Chandra et al. segmented the lung areas and retrieved eight statistical characteristics from them. They used the Multi-Layer Perceptron (MLP)[11], Random Forest[12], Sequential Minimal Optimization (SMO)[13], Classification through Regression, and Logistic Regression, five conventional classifiers. They tested their technique on 412 photos and used the MLP classifier to reach an accuracy rate of 95.39 percent. Kuo et al. identified pneumonia in 185 schizophrenia patients using 11 characteristics. They used these features in numerous regression and classification models, including logistic regression, decision trees, and support vector machines, and compared the outcomes of the models. Using a decision tree classifier, they were able to get the greatest accuracy rate of 94.5%; the other models were far less successful. Similar to this, Yue et al. used 6 characteristics to identify pneumonia in 52 patients' chest CT scan pictures; their best AUC value was 97%. Unfortunately, because they were tested on tiny datasets, these strategies cannot be applied generally.

Deep learning-based methods perform end-to-end classification, where the pertinent and informative features are automatically extracted from the input data and classified, in contrast to machine learning algorithms, for which handcrafted features need to be extracted and selected for classification or segmentation. Because they automatically extract translationally invariant features through the convolution of the input picture and filters, CNNs are recommended for the categorization of image data. Since CNNs outperform machine learning and conventional image processing techniques in image classification tasks and are translationally invariant, researchers frequently use them.

Simple CNN architectures were developed by Sharma et al. and Stephen et al. for the classification of pneumonic chest X-ray pictures. The lack of data was made up for via data augmentation. On the dataset provided by Kermany et al., referred to as the Kermany dataset, Sharma et al. acquired a 90.68% accuracy rate and Stephen et al. a 93.73% accuracy rate. However, data augmentation only offers a little amount of fresh data from which the CNNs can learn, and may not considerably improve their performance. Rajpukar et al .'s categorization of pneumonia using the DenseNet-121 CNN model only yielded a 76.8% f1-score. They hypothesised that a significant factor contributing to the poor performance of both their deep learning model and the radiologists with whom they compared the performance of their approach was the lack of patient history.

In order to reduce the dependence of models on the source of the datasets and create reliable predictions, Janizek et al. suggested a framework based on adversarial optimization. They achieved a source domain AUC score of 74.7% and a target domain AUC score of 73.9%. As a one-class problem, Zhang et al. suggested a confidence-aware module for anomaly identification in lung X-ray pictures (determining only the anomalies). For their dataset, they obtained an AUC value of 83.61%. The fuzzy tree modification was applied to the photos by Tuncer et al. using a machine learning-based technique, which was followed by an exemplar division. The data were then categorised using conventional classifiers after features were retrieved using a multikernel local binary pattern[14], Using a tiny dataset made up of COVID-19 and pulmonary emphysema sample, they tested the approach and found that it had a 97.01% accuracy rate.

Transfer learning, in which knowledge from a big dataset is utilised to fine-tune the model on a current small dataset, is now a widely used strategy to address the problem of data scarcity in biomedical image classification tasks. Rahman et al., Liang et al., Ibrahim et al., and Zubair et al. recently adopted solely transfer learning methodologies in which various CNN models are used for pneumonia classification and pre-trained using ImageNet data. The state of the art for the problem of pneumonia detection is tabulated in. Modern deep learning techniques for pneumonia detection typically concentrate on using a single CNN model. When judgements from numerous CNN models are combined, the ensemble model efficiently incorporates the key characteristics of all its base models, gathers supplementary data from other classifiers, and makes a more reliable decision. The exploration of this paradigm in respect to the pulmonary emphysema detection task is rare. In order to detect pulmonary emphysema traces by segmentation, Jaiswal et al. used a mask region-based CNN. For image thresholding[15],they used an ensemble model made up of ResNet-50 and ResNet-101. A deep learning system for the localisation of pulmonary opacity was proposed by Gabruseva et al., and it was built on a single-shot detector RetinaNet with Se-ResNext101 encoders. One of the best outcomes in the Radiological Society of North America (RSNA)[16] Pneumonia Detection Challenge was achieved by them when they carried out an ensemble of numerous checkpoints during the training phase (snapshot ensembling). They attained a mean average precision (mAP) of 0.26 over several intersection over union thresholds. Using an ensemble of the pulmonary emphysema detection models from Inception-ResNet v2, XceptionNet, and DenseNet-169 on the same challenge, Pan et al. achieved the best performance in the test, a mAP value of 0.33. To the best of our knowledge, ensemble models have not been applied to classification tasks in the pulmonary emphysema detection problem; therefore, in this study, we adopted ensemble learning for the first time in this field to categorise lung X-rays into " pulmonary emphysema " and "Normal" classes. GoogLeNet, ResNet-18, and DenseNet-121, three cutting-edge CNN models with transfer learning, were utilised to create the ensemble using a weighted average probability procedure, in which the weights are distributed using a novel method.

II.MOTIVATION AND CONTRIBUTION

As previously mentioned, pulmonary emphysema affects a large number of individuals, especially children, mostly in developing and underdeveloped countries characterised by risk factors such as overcrowding, poor hygienic conditions, and malnutrition, coupled with the unavailability of appropriate medical facilities. Early diagnosis of pneumonia is crucial to cure the disease completely. Examination of X-ray scans is the most common means of diagnosis, but it depends on the interpretative ability of the radiologist and frequently is not agreed upon by the radiologists. Thus, an automatic CAD system with generalising capability is required to diagnose the disease. To the best of our knowledge, the ensemble learning paradigm has not been investigated in this classification job, and the majority of prior methods in the literature concentrated on creating a single CNN model for the categorization of pulmonary emphysema cases. However, the ensemble learning model was used in this work because it integrates the discriminative data from all of the constituent base learners, enabling it to generate better predictions. Transfer learning models were utilised as base learners, and the decision scores of these models were ensembled, to deal with the limited amount of biological data that was accessible.

The following are the study's significant contributions.

1. To improve the performance of the base CNN learners in the categorization of pneumonia, an ensemble architecture was devised. A weighted average ensemble approach was used for this.
2. Precision, recall, f1-score, and AUC, four evaluation measures, were combined to generate the classifier weights. We employed a hyperbolic tangent function instead of choosing the weights purely based on the performance of classifiers or in accordance with the findings of tests.
3. Using the five-fold cross-validation configuration, the proposed model was tested on two publically accessible chest X-ray datasets, the Kermany dataset and the RSNA pulmonary emphysema Detection Challenge dataset[17]. The effectiveness of the procedure is demonstrated by findings that outperform those of state-of-the-art techniques.

III.PROPOSED METHOD

In this study, using a weighted average ensemble approach, we created an ensemble framework of three classifiers, GoogLeNet, ResNet-18, and DenseNet-121 (Fig 2). The weights assigned to the classifiers are constructed using a new scheme, which is covered in more detail in the following sections.

IV.GOOGLENET

Szegedy et al GoogLeNet's design is a 22-layer deep network made up of "inception modules" rather than layers that progress uniformly. An inception block hosts parallel convolution and pooling layers to support a large number of units at each level, leading to an unmanageable computational complexity due to the growing number of parameters. The GoogLeNet model uses inception blocks with dimension reduction, as illustrated in rather than the naive inception block to control the computational complexity. An ideal sparse architecture constructed from the available dense building blocks increases the performance of

artificial neural networks for computer vision applications, as demonstrated by the performance of GoogLeNet, in which the inception block was introduced.

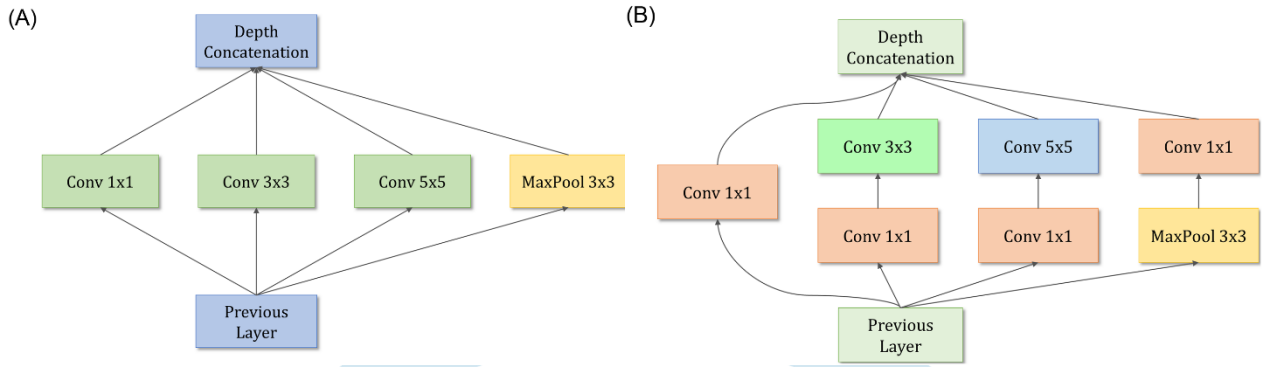

Figure 3: Inception modules in the GoogleNet architecture.

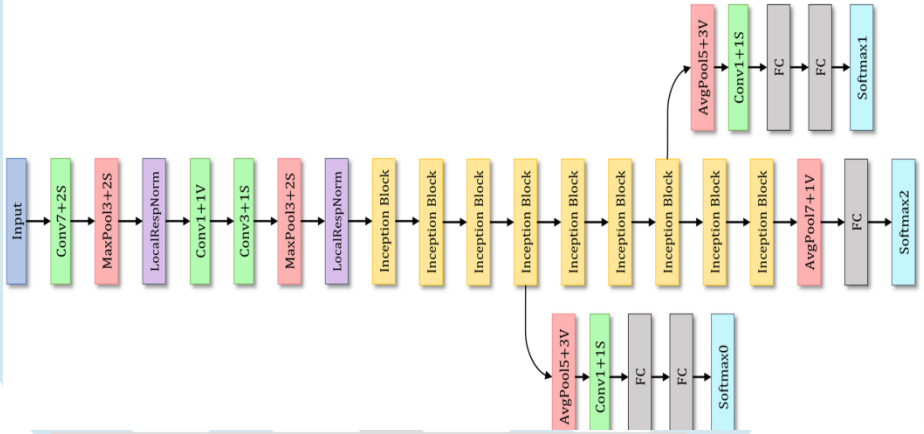

Figure 4: Architecture of the GoogleNet model used in this study.

V.ResNet-18

The ResNet-18 model put out by He et al. is built on a residual learning framework, which improves deep network training effectiveness. Contrary to the original unreferenced mapping in monotonically progressive convolutions[18], the residual blocks in ResNet models make it easier to optimise the overall network, which in turn increases model accuracy. Identity mapping is carried out via these residuals or "skip connections," which neither adds additional parameters nor makes the computation more complex presents the ResNet-18 model's design.

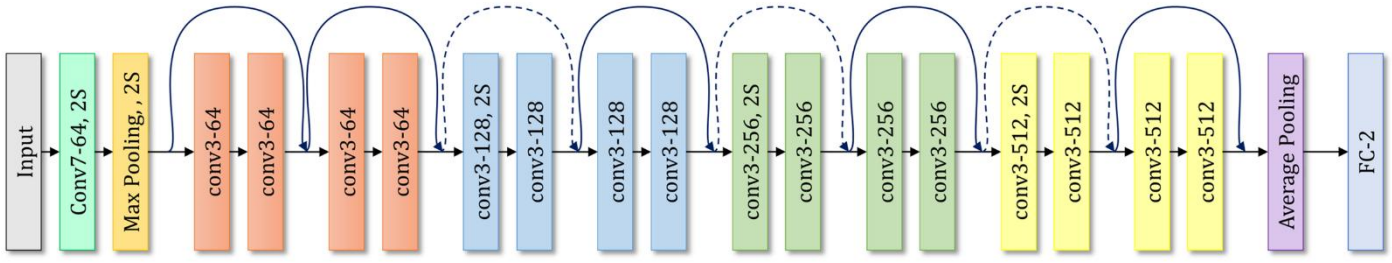

Figure 5: Architecture of the ResNet-18 model used in this study.

VI.DenseNet-121

Huang et al 's DenseNet designs offer a rich feature representation while being computationally effective. The fundamental explanation is that, as shown in Fig 6, the feature maps in each layer of the DenseNet model are concatenated with those from all the preceding levels. The convolutional layers can contain fewer channels, which reduces the number of trainable parameters and makes the model more computationally efficient. Moreover, the feature representation is improved by concatenating the feature maps from earlier layers with the current layer.

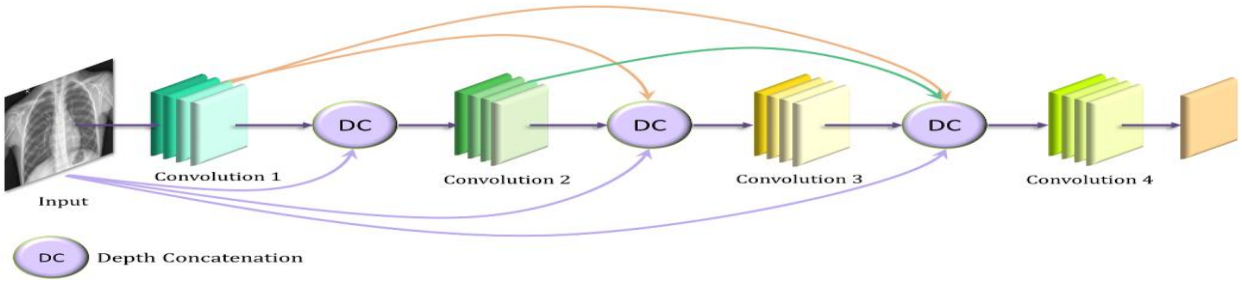

Figure 6: DenseNet's convolutional neural network model's basic design.

Table displays the empirically determined values of the hyperparameters that were utilised to train the learning algorithms (base learners).

Table 1: Hyperparameters that are used to train the basic learners in convolutional neural networks.

Hyperparameter	Value
Optimizer	Adam
Loss Function	Cross Entropy
Initial Learning Rate	0.0001
Learning Rate Scheduler	ReduceLROnPlateau
No. of Epochs	30

VII.A SUGGESTED ENSEMBLE STRATEGY

The ensemble learning model's predictions are superior to those of any of its constituent base learners because it helps to include the discriminative information of all of its constituent models. An effective approach for classifier fusion is weighted average assembly. Yet, one of the most important factors in assuring the ensemble's success is the selection of the weights to be given to the various foundation learners. The weights are typically set experimentally or solely on the classifier's accuracy in the literature. When there is a class imbalance in the dataset, this measurement may not be the best one. Other evaluation metrics including precision, recall (sensitivity), f1-score, and AUC may offer quite reliable data for figuring out how important the base learners are. In order to achieve this, we developed a novel weight allocation approach for this investigation, which is described below.

The weights allocated to each basis learner using the suggested technique are first calculated using the probability scores that the base learners acquired during the training phase. An ensemble that has been trained on the test set is created using these produced weights. In order to maintain the test set's independence from predictions, this method is used. The precision score (pre(i)), recall score (rec(i)), f1-score (f1(i)), and AUC score (AUC(i)) for the ith model () are created and compared with the true labels (y). Suppose that these variables combine to produce the array A(i) = "pre(i), rec(i), f1(i), AUC(i)". The hyperbolic tangent function is then used to calculate the weight (w(i)) given to each classifier, as illustrated in Eq 1. Because x represents an evaluation metric with a value that falls between [0, 1], the hyperbolic tangent function's range is [0, 0.762]. It monotonically rises in this range, therefore the tanh function rewards metrics with high values by giving them a high priority; otherwise, it penalises them.

$$w^{(i)} = \sum_{x \in A^{(i)}} \tanh(x)$$

$$= \sum_{x \in A^{(i)}} \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \quad \text{Eq. (1)}$$

For a binary class dataset, the probability array for the jth test sample by the ith base classifier is, where a 1 and the ensemble probability for the sample is ensemble probj = b, 1 b. These weights (w(i)) computed by Eq 1 are multiplied by the decision scores of the corresponding base learners to compute the weighted average probability ensemble, as shown in Eq 2.

$$ensemble\_prob_j = \frac{\sum_i w^{(i)} \times p_j^{(i)}}{\sum_i w^{(i)}} \quad \text{Eq. (2)}$$

Finally, the class predicted by the ensemble is computed by where $prediction_j$ denotes the predicted class of the sample.

$$prediction_j = argmax(ensemble\_prob_j) \quad \text{Eq.(3)}$$

VIII.RESULTS AND DISCUSSION

We summarize the findings of the proposed methods examination in this section. We used two chest X-ray datasets for pneumonia that were openly accessible. The first dataset, the Kermany dataset , consists of 5856 chest X-ray pictures that were randomly distributed between the classes "Pneumonia" and "Normal" from a sizable population of both adults and children. The RSNA donated the second dataset, which was a Kaggle challenge for pneumonia detection. Table 3 gives the distribution of photos across the two datasets. The table also includes a description of each image in the training and test sets for each fold of the 5-fold cross-validation method used in this study. Also, a discussion of the results' ramifications is included. To prove the suggested method's superiority over other models and popular ensemble techniques described in the literature, a comparative analysis was done.

Table 2: Images from the training and test sets in every iteration of the five-fold cross-validation on the two datasets utilised in this study are described.

Dataset	Division	Class	No. of Images	Size of Images (Range)	Size of Resized Images
Kermany	Train	Normal	1267	(127×384×3)—(2713×2517×3)	224×224×3
		Pneumonia	3419		
	Test	Normal	316	(189×490×3)—(2458×2720×3)	224×224×3
		Pneumonia	854		
	Total Images		**5856**		
RSNA	Train	Lung Opacity	16488	(1024×1024×3)	224×224×3
		No Lung Opacity	4801		
	Test	Lung Opacity	4111	(1024×1024×3)	224×224×3
		No Lung Opacity	1201		
	Total Images		**26601**		

IX.EVALUATION METRICS

Four common assessment criteria were employed to assess the proposed ensemble approach on the two pneumonia datasets: accuracy (Acc), precision (Pre), recall (Rec), and f1-score (F1). First, we define the phrases "True Positive," "False Positive," "True Negative," and "False Negative" in order to define these evaluation measures.Assume the two classes in the dataset are referred to as the "positive" and the "negative" class for a binary classification task. The following definition applies to the aforementioned concepts.

- A sample that is accurately identified by a model as being in the positive class is said to be True Positive (TP).
- A sample that is accurately identified as belonging to the positive class but actually belongs to the negative class is referred to as a false positive (FP).
- A sample that the model accurately identified as being in the negative class is referred to as True Negative (TN).
- False Negative (FN) describes a sample that is accurately identified as belonging to the positive class but is actually a member of the negative class.

Now, the four evaluation metrics can be defined as

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{Eq. (4)}$$

$$Pre = \frac{TP}{TP + FP} \quad \text{Eq. (5)}$$

$$Rec\ (or\ Sensitivity) = \frac{TP}{TP + FN} \quad \text{Eq. (6)}$$

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \quad \text{Eq. (7)}$$

The accuracy rate gives a broad indication of how many of the model's predictions were accurate. However, if the dataset is unbalanced, a model's high accuracy rate does not guarantee that it can evenly distinguish between different classes. In particular, a model that can be applied to all classes is needed for the classification of medical images. In these circumstances, the "accuracy" and "recall" numbers shed light on the model's functionality. The accuracy of the model's predicted positive label is displayed in "precision". This gives the proportion of accurate forecasts to all of the model's predictions. The percentage of ground truth positives that the model accurately predicted is measured by "recall," on the other hand. The model's ability to decrease the quantity of FP and FN predictions is measured by these two assessment criteria. By taking into account both FPs and FNs, "F1-Score" strikes a compromise between "accuracy" and "recall." Extreme "precision" and "recall" levels that are attained at the expense of one another are punished. To accurately identify both a healthy and a diseased person, it is useful in medical picture classification to take evaluation metrics into account in addition to the accuracy rate.

X.IMPLEMENTATION

This study employed a five-fold cross-validation method to thoroughly assess how well the suggested ensemble model performed. For the Kermany dataset and the RSNA challenge dataset , the findings for each fold as well as the average and standard deviation values over the five folds are given in Tables respectively. The high results for sensitivity (recall) and accuracy (accuracy) show how reliable the suggested technique is. Also, figure display the confusion matrices for the Kermany and RSNA datasets, respectively, while Figure displays the ROC curves for all five cross-validation folds using the proposed technique on the two datasets

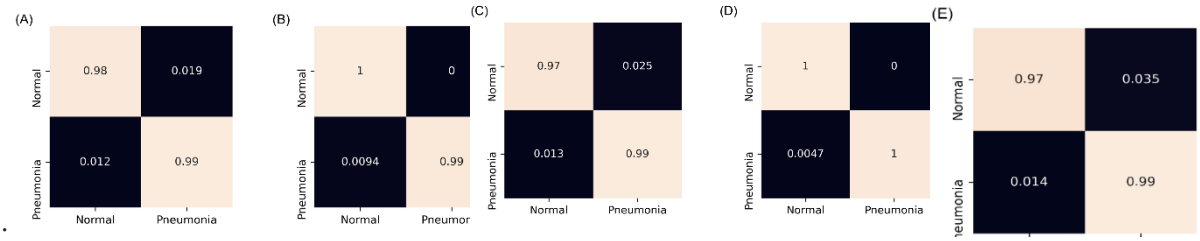

Figure 7: Confusion matrices were created using the suggested method by 5-fold cross validation on the Kermany pneumonia chest X-ray dataset.

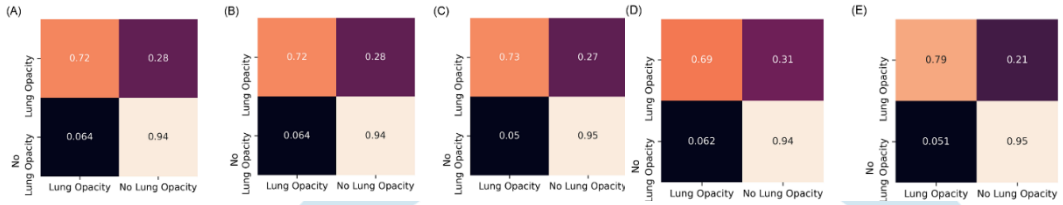

Figure 8: The suggested method by five-fold cross validation produced confusion matrices on the chest X-ray dataset for the pneumonia challenge from the Radiological Society of North America.

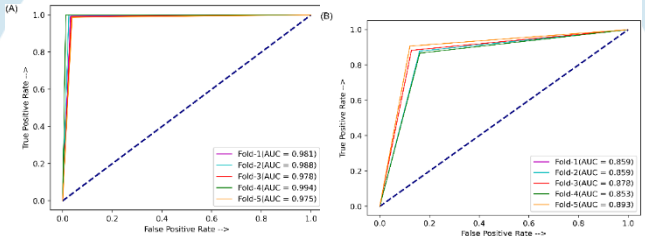

Figure 9: Receiver operating characteristic curves on the two pneumonia chest X-ray datasets acquired using the proposed ensemble technique in this study.

Table 3:- Outcomes of the proposed ensemble method's five-fold cross-validation on the pulmonary emphysema Kermany dataset

Fold	Acc(%)	Pre(%)	Rec(%)	F1(%)	AUC(%)
1	98.63	98.64	98.63	98.63	98.12
2	99.31	99.33	99.32	99.32	98.82
3	98.38	98.46	98.38	98.29	97.86
4	99.68	99.66	99.66	99.66	99.43
5	98.03	98.03	98.03	98.03	97.54
Avg±Std. Dev.	**98.81±0.61**	**98.82±0.59**	**98.80±0.60**	**98.79±0.61**	**98.35±0.68**

Table 4: Results of the suggested ensemble method's five-fold cross-validation on the pneumonia challenge dataset from the Radiological Society of North America

Fold	Acc(%)	Pre(%)	Rec(%)	F1(%)	AUC (%)
1	86.63	86.78	86.63	86.70	86.63
2	86.78	87.05	87.05	87.05	86.78
3	87.97	88.00	87.80	87.90	87.97
4	85.98	86.00	86.63	86.31	85.98
5	86.89	86.63	86.98	86.80	86.89
Avg±Std. Dev.	**86.85±0.72**	**86.89±0.73**	**87.02±0.48**	**86.95±0.59**	**86.85±0.72**

Demonstrates the accuracy rates attained by the transfer learning base learners using various optimizers on the Kermany dataset. The Adam optimizer[19] was chosen to train the basis learners for the ensemble framework because it produced the best results for all three base learners.

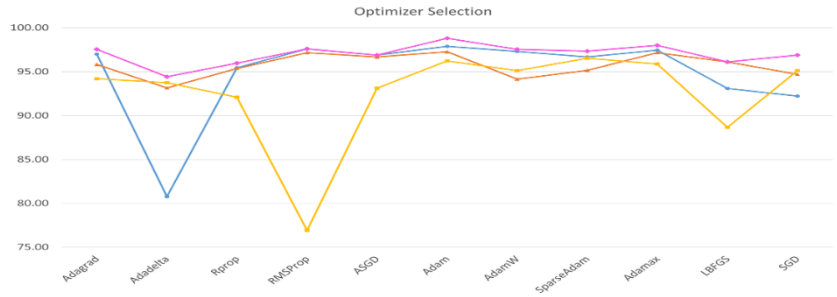

Figure 10: The three base learners GoogLeNet, ResNet-18, and DenseNet-121 and their ensemble achieved different accuracy rates on the Kermany dataset depending on the optimizers employed for fine tuning.

The three base learners GoogLeNet, ResNet-18, and DenseNet-121 and their ensemble achieved different accuracy rates on the Kermany dataset depending on the optimizers employed for fine tuning. results from several ensembles made up of GoogLeNet, ResNet-18, ResNet-50, ResNet-152, DenseNet-121, DenseNet-169, DenseNet-201, MobileNet v2, and NasMobileNet on the Kermany dataset (including recently proposed architectures). The GoogLeNet, ResNet-18, and DenseNet-121 base learner combinations that were used in this investigation were chosen based on the results. The accuracy rating for the ensemble combination was 98.81%. The combination of Google Net, ResNet-18, and MobileNet v2 produced the second-best result, with an accuracy rate of 98.54%. Also, we adjusted some of the layers in the ensemble's execution for the chosen set of base learners, GoogLeNet, ResNet-18, and DenseNet-121, and trained the models to determine the best configuration. The outcomes are displayed in Figure On both datasets, the ensemble produced its best results when all of the layers were trainable (no layers were frozen). Thus, we decided to use this configuration for the ensemble framework.

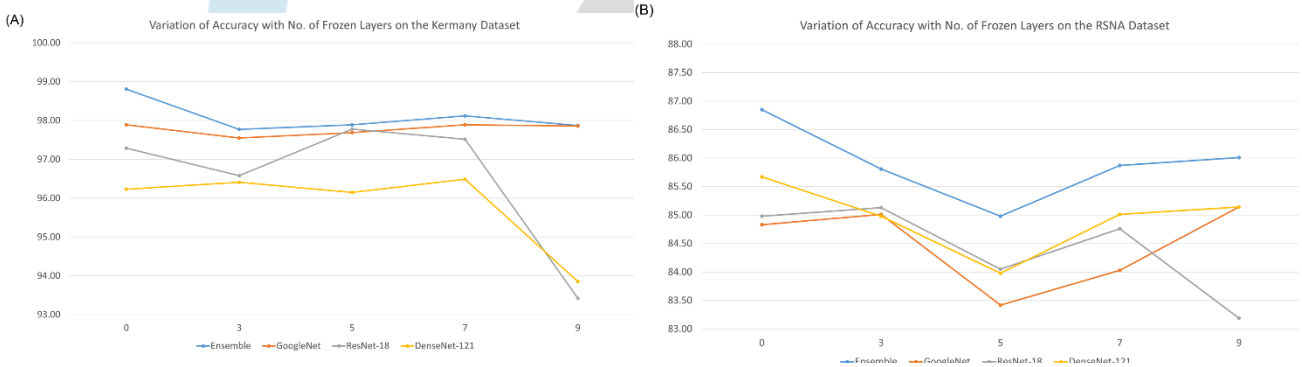

Figure 11: On the two datasets utilised in this work, there were differences in the performance (accuracy rates) of the ensemble depending on the quantity of fixed, untrainable layers in the base learners.

Table 5: Extensive tests were conducted to identify the base learners who would make up the ensemble in this study.

Model-1	Model-2	Model-3	Acc(%)	Pre(%)	Rec(%)	F1(%)
NasNetMobile	MobileNet v2	ResNet-152	96.67	96.70	96.67	96.68
NasNetMobile	MobileNet v2	ResNet-50	97.00	97.02	97.01	97.01
NasNetMobile	MobileNet v2	DenseNet-169	96.41	96.40	96.41	96.41
NasNetMobile	MobileNet v2	DenseNet-201	96.06	96.21	96.07	96.11
MobileNet v2	ResNet-152	DenseNet-169	96.92	97.03	96.92	96.95
MobileNet v2	ResNet-50	DenseNet-169	97.77	97.82	97.78	97.79
MobileNet v2	ResNet-50	DenseNet-201	95.98	96.40	95.98	96.06
MobileNet v2	ResNet-152	DenseNet-201	94.87	95.57	94.87	94.99
NasNetMobile	ResNet-152	DenseNet-169	95.21	95.71	95.21	95.31
NasNetMobile	ResNet-152	DenseNet-201	92.56	94.06	92.56	92.81
NasNetMobile	ResNet-50	DenseNet-169	96.41	96.66	96.41	96.46
NasNetMobile	ResNet-50	DenseNet-201	92.99	94.28	92.99	93.20
GoogLeNet	ResNet-152	DenseNet-121	97.17	97.37	97.18	97.21
GoogLeNet	ResNet-152	DenseNet-201	95.04	95.65	95.04	95.15
GoogLeNet	ResNet-18	DenseNet-201	98.20	98.23	98.21	98.21
GoogLeNet	MobileNet v2	DenseNet-121	98.29	98.29	98.29	98.29
GoogLeNet	ResNet-18	MobileNet v2	98.54	98.54	98.55	98.54
GoogLeNet	MobileNet v2	NasNetMobile	98.12	98.13	98.12	98.11
GoogLeNet	**ResNet-18**	**DenseNet-121**	**98.81**	**98.82**	**98.80**	**98.35**

XI.Analysis of gradient-weighted class activation maps

Gradient-weighted class activation maps[20] (GradCAM) were employed in this study to present a visual representation of the distinguishing regions in the chest X-ray images, that is, the regions on which the classifier focuses to make a prediction. CAM calculates the number of weights of each feature map (FM) based on the last convolution layer to compute the contribution of the FM to the prediction $\hat{y}$ at location (*i*, *j*), where our objective is a computed value of L_{ij}^{g} that satisfies $y^{g} = \sum_{i,j} L_{ij}^{g}$. The final FM ($C_{ij}^{k}$) and the prediction $\hat{y}$ are represented through a linear relationship in which the linear layers contain global average pooling[21] (GAP) layers and fully connected layers (FCLs). (1) GAP outputs $A_k = C_{ij}^{k}$ and (2) the FCLs, which hold weight w_k^{g}, generate an output as in,where C_k represents the visualization of the k^{th} FM: $L_{ij}^{g} = \sum_k w_k^{g} C_{ij}^{k}$

$$y^{g} = \sum_{k} w_k^{g} A_k = \sum_{k} w_k^{g} \sum_{i,j} C_{ij}^{k} = \sum_{i,j} \sum_{k} w_k^{g} C_{ij}^{k}$$

. Eq.(8)

CAM is an unsuitable method because of the problem of the vanishing nonlinearity of classifiers. Thus, instead of pooling them, we use GradCAM for globally averaging the gradients of the FM as weights. While the heat maps are plotted, class-specific weights are collected from the last convolution layer through globally averaged gradients (GAG) of the FM instead of pooling, as in Eq9, where *P* is the number of pixels in an FM, *g* is the gradient of the class, and C_{ij}^{k} is the value of the k^{th} FM

$$\alpha_k^{g} = \frac{1}{P} \sum_{i} \sum_{j} \frac{\partial y^{g}}{\partial C_{ij}^{k}}$$ Eq. (9)

After the relative weights have been gathered, the coarse saliency map (L^{c}) is calculated as the weighted sum, $\alpha_k^{c} \times C_{ij}^{k}$, of the ReLU activation where α_k^{c} represents the neuron importance weights. It introduces a linear combination to the FM because only the features that have a positive influence on the respective class are of interest; the negative pixels in the image that belong to other categories are discarded .

$$L^{g} = ReLU\left(\sum_{i} \alpha_k^{g} C^{k}\right)$$ Eq.(10)

shows the results of the GradCAM analysis of a pneumonic and a healthy lung X-ray, where all three models were used to form the ensemble. Evidently, the different models focused on different regions of the lung X-rays, indicating that the base learners capture complementary information. This led to the success of the ensemble approach. The confidence scores for the pneumonic lung X-ray shown in are GoogLeNet: 99.99%, ResNet-18: 75.21%, and DenseNet-121: 98.90%; all predicted correctly. For the healthy lung case shown in , the confidence scores are GoogLeNet: 99.47%, ResNet-18: 97.61%, and DenseNet-121: 98.93%; all predicted correctly.

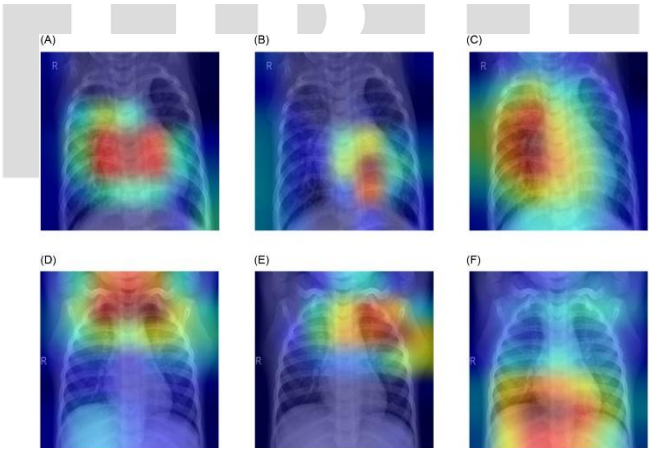

Figure 12: When the three selected base learners were used to create the ensemble, chest X-ray images were visualised using a gradient-weighted class activation map (GradCAM).

XII.Comparison with cutting-edge techniques

For the Kermany pneumonia dataset[22], Table 7 compares the effectiveness of the suggested ensemble framework and that of the presently used approaches in the literature. The proposed method performed better than every other method, it should be highlighted. The fact that the proposed ensemble framework outperformed all of these earlier approaches (Mahmud et al. , Zubair et al. , Stephen et al. , Sharma et al. , and Liang et al.) for the classification of pneumonic lung X-ray images is also noteworthy. This shows that the ensemble technique developed in this study is a reliable approach for the image classification task under consideration. As far as we are aware, no research has been done on the classification of images in the RSNA pneumonia dataset. We therefore contrasted the suggested model's performance with that of a number of baseline CNN algorithms for this dataset.

compares the assessment results for the proposed technique on both of the study's datasets to those obtained using the base CNN models used to create the ensemble and a number of other conventional CNN transfer learning models. In both datasets, it can be seen that the suggested ensemble method performed fairly well compared to the base learners and other transfer learning models. Also, the outcomes are summarised in Table 9 to demonstrate the suggested ensemble scheme's superiority to well-known traditional ensemble procedures.

The ensembles used the same three basic CNN learners—GoogLeNet, ResNet-18, and DenseNet-121—and the average performance throughout the five cross-validation folds is displayed for both the Kermany and RSNA challenge datasets. The suggested ensemble approacn performed better than widely used ensemble systems. It is clear from both datasets that the performance of the weighted average ensemble, which simply takes the accuracy metric into account when determining the weights, came the closest to matching the performance of the proposed ensemble technique. The class that received the most votes from the base learners is expected to be the class of the sample in the majority voting-based ensemble. The predicted class of the sample is determined by adding the probability scores for each class across all base learners in the maximum probability ensemble, whereas equal weight is given to each contributing classifier in the average probability ensemble.

Table 6: Using the Kermany pneumonia dataset and the Radiological Society of North America challenge dataset, the suggested method is compared to other methods that have been described in the literature.

Dataset	Model	Acc(%)	Pre(%)	Rec(%)	F1(%)	AUC(%)
Kermany	GoogLeNet	97.89	98.12	98.12	98.12	97.89
	AlexNet	97.17	97.22	97.18	97.19	97.17
	VGG-16	97.09	97.12	97.09	97.1	97.09
	DenseNet-121	96.23	96.63	96.24	96.31	96.23
	ResNet-18	97.29	98.31	98.29	98.3	97.29
	Proposed Method	**98.81**	**98.82**	**98.80**	**98.79**	**98.35**
RSNA	GoogLeNet	84.83	84.98	84.83	84.90	84.83
	AlexNet	85.86	85.15	84.86	85.00	85.86
	VGG-16	81.05	80.08	81.17	80.62	81.05
	DenseNet-121	85.67	84.98	85.55	85.26	85.67
	ResNet-18	84.98	85.55	85.37	85.46	84.98
	Proposed Method	**86.85**	**86.89**	**87.02**	**86.95**	**86.85**

Table 7: On the Kermany and Radiological Society of North America challenge datasets, comparison of the suggested ensemble framework with a number of well used convolution neural network models in the literature.

Dataset	Model	Acc(%)	Pre(%)	Rec(%)	F1(%)	AUC(%)
Kermany	GoogLeNet	97.89	98.12	98.12	98.12	97.89
	AlexNet	97.17	97.22	97.18	97.19	97.17
	VGG-16	97.09	97.12	97.09	97.1	97.09
	DenseNet-121	96.23	96.63	96.24	96.31	96.23
	ResNet-18	97.29	98.31	98.29	98.3	97.29
	Proposed Method	**98.81**	**98.82**	**98.80**	**98.79**	**98.35**
RSNA	GoogLeNet	84.83	84.98	84.83	84.90	84.83
	AlexNet	85.86	85.15	84.86	85.00	85.86
	VGG-16	81.05	80.08	81.17	80.62	81.05
	DenseNet-121	85.67	84.98	85.55	85.26	85.67
	ResNet-18	84.98	85.55	85.37	85.46	84.98
	Proposed Method	**86.85**	**86.89**	**87.02**	**86.95**	**86.85**

XIII.Analysis of errors

In Fig. 13, two test samples from the Kermany dataset are shown. Two base learners gave inaccurate predictions with low confidence rates, whereas a third base learner gave the correct forecast with a very high confidence rate. This resulted in the ensemble framework accurately predicting the sample. Figure 13(a) displays a sample where GoogLeNet, ResNet-18, and DenseNet-121 each correctly predicted "Pneumonia" with a confidence score of 52.1%, 73.8%, and 89.4%, respectively. With a confidence score of 68.1%, the ensemble framework's final prediction that the sample belonged to the "Normal" class was accurate. Similar to how GoogLeNet predicted "Normal" in the instance of Fig. 13(a), ResNet-18 predicted "Pneumonia" with a confidence score of 58.3%, while DenseNet-121 predicted "Pneumonia" with a confidence score of 69.3%. With a confidence score of 66.3%, the suggested ensemble framework correctly identified the sample as "Normal." This demonstrates how well the ensemble framework performed.

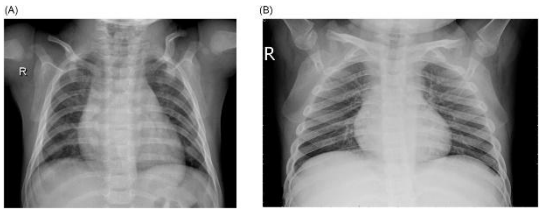

Figure 13: Examples of samples from the Kermany dataset where the ensemble produced the correct prediction despite two out of three base learners producing wrong predictions.

demonstrates a number of test samples from the Kermany dataset where the ensemble framework was unable to appropriately categorise the samples. A sample from the class "Normal" that was mistakenly labelled as " pulmonary emphysema " is presented in Fig. 14(a), and the matching GradCAM analysis images are shown in sections (c), (d), and (e) (e). This might be the result of low image quality, wherein the contrast of the image is insufficient, causing the base learners to mistakenly categorise the sample. According to the GradCAM analysis, ResNet-18 concentrated on the white area of the retracted lungs while GoogLeNet and DenseNet-121 focused on the spinal cord in the X-ray, making inaccurate predictions. Figure 14(b) illustrates a situation in which the model incorrectly identified a " pulmonary emphysema "-class image as "Normal." The images from the GradCAM analysis are displayed in (f), (g), and (h). The GoogLeNet and DenseNet-121 models, like in the prior instance, concentrated on the spinal cord, whereas the ResNet-18 model concentrated on a portion of the spinal cord and the retracted left lung. As seen in Fig 1, abscesses or pleural effusion[23], or fluid in the alveoli, which shows up as white patches on a lung X-ray, are characteristics of a pneumonic lung condition. It is conceivable that the CNN models failed to detect pneumonia at such an early stage, when the white infiltrates have just recently begun to emerge sporadically in the lungs. In these situations, air bronchogram signals are used by clinicians to identify pneumonia. Lung cancer, TB, and pulmonary emphysema were differentiated based on the form and lumen of the bronchi with air bronchogram[24] indications.

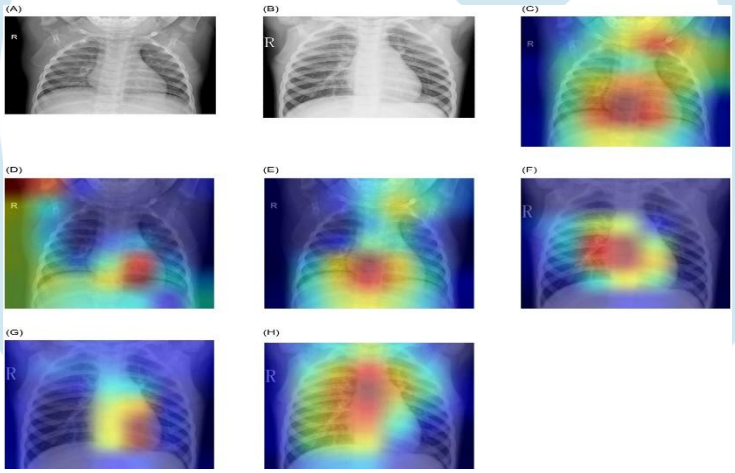

Figure 14: Examples of samples from the Kermany dataset that the suggested ensemble framework erroneously categorised.

XIV.Analytical statistics

The proposed ensemble model was compared to the base classifiers, the probability scores of which were used in this study to determine the formation of the ensemble: GoogLeNet, ResNet-18, and DenseNet-121, in two non-parametric tests, McNemar's statistical test[25] and the analysis of variance (ANOVA) test[26] , to statistically analyse the viability of the proposed ensemble framework. The results of the McNemar's test and the ANOVA test on the two pulmonary emphysema chest X-ray datasets utilised in this investigation are tabulated in Tables 10 and 11, respectively. According to Tables 10 and 11, every example in both datasets has a p-value that is less than 0.05, which is what is required for both the McNemar's test and the ANOVA test to reject the null hypothesis (5%). The outcomes of both statistical tests thus disproved the null hypothesis. In order to guarantee that the ensemble model is statistically distinct from any of the contributing models, it is established that the proposed ensemble framework captures complementary information from the base classifiers and that its predictions are superior.

Table 7: Results of McNemar's statistical test of the ensemble model and the base learners on both datasets.

McNemar's Test	p-value	
	Kermany dataset	RSNA dataset
GoogLeNet	0.0000	0.0017
ResNet-18	0.0430	0.0214
DenseNet-121	0.0002	0.0006

Table 8: Results of the statistical analysis of variance (ANOVA) test conducted on both datasets using the ensemble model and the base learners.

ANOVA Test	p-value	
	Kermany Dataset	RSNA Dataset
GoogLeNet	0.0435	0.0131
ResNet-18	0.0021	0.0001
DenseNet-121	0.0017	0.0056

XV.CONCLUSION AND FUTURE WORK

For the condition to be properly treated and to avoid endangering the patient's life, early diagnosis of pulmonary emphysema is essential. The most popular method for diagnosing pneumonia is a chest radiograph[27], although there is inter-class variability, and the diagnosis rests on the physicians' skill at spotting early pulmonary emphysema signs. An automated CAD system[28] was created in this work to help doctors by categorising chest X-ray pictures into the " pulmonary emphysema " and "Normal" classes using deep transfer learning-based categorization. The decision scores from three CNN models—GoogLeNet, ResNet-18, and DenseNet-121—were taken into account in an ensemble framework to create a weighted average ensemble. Four assessment metrics—precision, recall, f1-score, and AUC—were combined using the hyperbolic tangent function in a novel way to determine the classifier weights. Using a five-fold cross-validation scheme, the framework was tested on two publicly available datasets of pneumonia chest X-rays and achieved accuracy rates of 98.81%, sensitivity rates of 98.80%, precision rates of 98.82%, and a f1-score of 98.79% on the Kermany dataset and accuracy rates of 86.86%, sensitivity rates of 87.02%, precision rates of 86.89%, and a f1- In these two datasets, it fared better than cutting-edge techniques. The strategy is viable, according to statistical studies of the suggested model utilising McNemar's[29] and ANOVA tests. The suggested ensemble approach is also domain-independent, making it applicable to a wide range of computer vision tasks.As was already established, the ensemble framework occasionally failed to provide the right predictions. We may look into methods like image contrast enhancement or other pre-processing processes in the future to increase the image quality. When classifying the lung image, we might also think about segmenting it to help the CNN models extract features more effectively. Also, the computation cost is larger than that of the CNN baselines generated in research in the literature since three CNN models are needed to train the proposed ensemble. In the future, we might try to use techniques like snapshot ensembling to lessen the processing requirements.

REFERENCES

[1] An article reference of computed tomography https://www.sciencedirect.com/science/article/abs/pii/S0720048X16300754
[2]An article reference of magnetic resonance imaging https://www.sciencedirect.com/science/article/abs/pii/S0749806322006806
[3] An article reference of radiography https://pubs.rsna.org/doi/full/10.1148/radiol.221926
[4] A book reference of computer-aided diagnosis https://link.springer.com/chapter/10.1007/978-981-19-2057-8_15
[5] An article reference of Convolutional neural networks https://www.sciencedirect.com/science/article/pii/S1746809422008990
[6] A web reference of ImageNet https://arxiv.org/abs/2201.05119
[7] A web reference of GoogLeNet https://openaccess.thecvf.com/content_cvpr_2017/html/Wu_A_Compact_DNN_CVPR_2017_paper.html
[8] An article reference of ResNet-18 https://www.mdpi.com/2072-4292/14/19/4883
[9] An article reference of DenseNet https://www.tandfonline.com/doi/abs/10.1080/2150704X.2019.1633486
[10] A web reference of conventional machine learning techniques https://dl.acm.org/doi/abs/10.1145/3495162
[11] An article reference of Multi-Layer Perceptron https://www.sciencedirect.com/science/article/abs/pii/S0960148117309485
[12] An article reference of Random Forest https://www.nature.com/articles/s41598-020-62133-5
[13] An article reference of Sequential Minimal Optimization https://www.sciencedirect.com/science/article/abs/pii/S0141933118303879
[14] An article reference of multikernel local binary pattern https://www.sciencedirect.com/science /article/pii/S0169743920301970
[15]A web reference of image thresholding https://onlinelibrary.wiley.com/doi/abs/10.1002/ima.22589
[16] An article reference of Radiological Society of North America (RSNA) https://journals.sagepub.com/doi/pdf/10.1177/0846537120968919
[17] A web reference of RSNA Pneumonia Detection Challenge dataset https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.15185
[18] An article reference of monotonically progressive convolutions https://www.nature.com/articles/nrn3275
[19] An article reference of Adam optimizer https://link.springer.com/chapter/10.1007/978-3-031-10548-7_43
[20] An article reference of Gradient-weighted class activation maps https://www.sciencedirect.com/science/article/pii/S0933365722001075
[21] An article reference of global average pooling https://ojs.aaai.org/index.php/AAAI/article/view/12154
[22]A web reference of Kermany pneumonia dataset https://pdfs.semanticscholar.org/9cc2/8b51948039bf2d4c7a26f340343e77a342cf.pdf
[23] A web reference of pleural effusion https://thorax.bmj.com/content/78/1/32.abstract
[24] An article reference of air bronchogram https://academic.oup.com/ejcts/article/45/4/699/360890
[25] An article reference of McNemar's statistical test https://www.sciencedirect.com/science/article/abs/pii/S0003267002014228

[26] A book reference of analysis of variance (ANOVA) test https://www.taylorfrancis.com /books/mono/10.4324/9780203763933/anova-behavioral-sciences-researcher-rudolf-cardinal-michael-aitken
[27] A web reference of chest radiograph https://pubs.rsna.org/doi/full/10.1148/radiol.2020201874
[28]A web reference of An automated CAD system https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.2214177
[29] An article reference of McNemar's https://www.tandfonline.com/doi/abs/10.1080/00365510701666031
[30] An article reference of the CNN models https://www.mdpi.com/1424-8220/21/3/951=

BIBILIOGRAPHY

Turubati Saikiran is studying his 2nd year Master of Computer Applications in Sanketika Vidya Parishad Engineering College, Visakhapatnam, A.P

With his interest in Python,Deep Learning and as a part of academic project he chose Image processing using Python. The article have been evolved from an idea to understand the flaws in conventional reporting and keeping time consistency, quality report generation in pulmonary emphysema. A full fledged project along with code has been submitted for Andhra University as an Academic Project.

Kandhati Tulasi Krishna Kumar: Project Guide & Training & Placement Officer with decade plus experience in training & placing the students into IT, ITES & Core profiles & trained more than9,000 UG, PG candidates & trained more than 350 faculty through FDPs. Authored various books for the benefit of the diploma, pharmacy, engineering & pure science graduating students. He is a Certified Campus Recruitment Trainer from JNTUA, did his Master of Technology degree in CSE from VTA and in process of his Doctoral research. He is a professional in Pro-E, CNC certified by CITD He is recognized as an editorial member of IJIT (International Journal for Information Technology & member in IAAC, IEEE, MISTE, IAENG, ISOC, ISQEM, and SDIWC. He published articles in various international journals on Databases, Software Engineering, Human Resource Management and Campus Recruitment & Training.