DEEP DIVE INTO THE USE OF IMAGE PROCESSING AND OBJECT DETECTION TO IDENTIFY PNEUMONIA

Prithvi Sairaj Krishnan

Department of Computer Science, Westwood High School, Austin, United States of America

ABSTRACT

Pneumonia is a major respiratory infection causing significant global morbidity and mortality, especially in developing nations with inadequate medical infrastructure. Early diagnosis through chest X-ray imaging is crucial but challenging. This study developed an automated computer-aided diagnosis system using deep learning to detect pneumonia from chest X-rays. An ensemble of three pre-trained convolutional neural network models (GoogLeNet, ResNet-18, DenseNet-121) was employed, with a novel weighted average ensemble technique based on evaluation metric scores. Evaluated on two public pneumonia X-ray datasets using five-fold cross-validation, the approach achieved high accuracy (98.2%, 86.7%) and sensitivity (98.19%, 86.62%), outperforming state-of-the-art methods. With pneumonia-causing over 2.5 million annual deaths worldwide, this accurate automated model can assist radiologists in timely diagnosis, especially in resource-limited settings. Its integration into clinical decision support systems has the potential to improve pneumonia management and outcomes significantly.

KEYWORDS

Convolutional Neural Networks, Pneumonia, X-Rays, Model, Machine Learning

1. INTRODUCTION

A dangerous lung infection brought on by bacteria, viruses, or fungi is known as pneumonia. Pleural effusion, a condition marked by fluid accumulation and inflammation of the lungs' air sacs, results from pneumonia. Many deaths in children under five are caused by pneumonia, especially in emerging and undeveloped nations with high rates of pollution, overcrowding, poor hygiene, and insufficient access to medical care. It is imperative to receive early diagnosis and appropriate treatment for pneumonia to avoid fatality. Pneumonia is typically diagnosed by radiological exams such as computed tomography (CT), magnetic resonance imaging (MRI), or X-rays. A non-invasive, reasonably priced way to examine the lungs is with X-ray imaging. White patches known as infiltrates, seen by red arrows in the sample image, are seen in pneumonic X-rays; they distinguish a pneumonic condition from a healthy lung. However, chest X-ray examinations for pneumonia detection are subject to subjective variability. Therefore, an automated system for pneumonia detection is necessary. In this study, the researcher developed a computer-aided diagnosis (CAD) system that utilises an ensemble of deep transfer learning models for accurate classification of chest X-ray images to detect pneumonia.

David C. Wyld et al. (Eds): SIGI, CSTY, AIMLNET, AI – 2024
pp. 01-16, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.141301

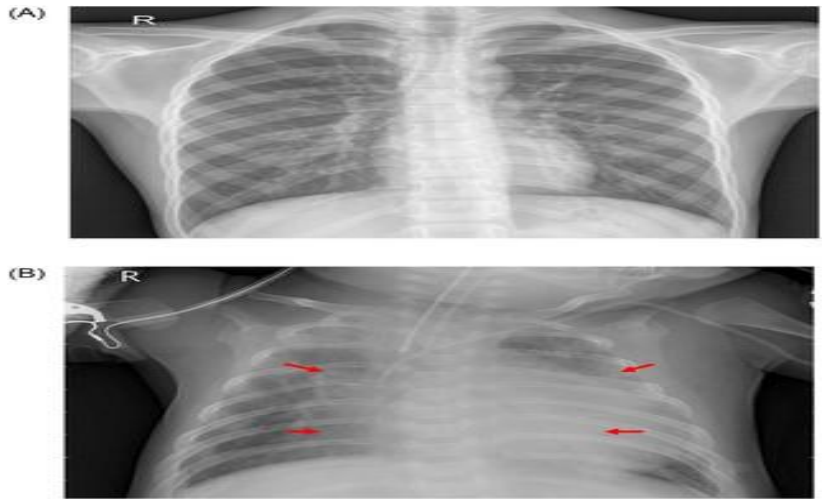

Fig 1. Examples of two X-ray plates that display (a) a healthy lung and (b) a pneumonic lung.

The red arrows in (b) indicate white infiltrates, a distinguishing feature of pneumonia. The images were taken from the Kermany dataset [2].

Convolutional neural networks (CNNs), in particular, are potent artificial intelligence tools frequently employed in deep learning to solve challenging computer vision problems. But for these models to function at their best, a lot of data is needed, and this can be difficult to come by for biomedical image classification tasks because each image must be classified by a team of highly qualified clinicians, which is costly and time-consuming. One method to overcome this challenge is transfer learning, which involves taking a model that was trained on a massive dataset—like ImageNet, which has over 14 million images—and using the learned network weights to solve a problem and make accurate predictions a final prediction for a test sample by combining the decisions of numerous classifiers is a popular approach known as ensemble learning. It seeks to extract the discriminative information from every base classifier to produce more accurate predictions. Average probability, weighted average probability, and majority voting are examples of common ensemble approaches. Although the average probability-based ensemble gives every basic learner equal weight, it is a better idea to give the base classifiers weights because some may be better at capturing information than others for a given task. Nonetheless, it guarantees improved performance, it is essential to ascertain the ideal weight values for every classifier.

Using four evaluation metrics—precision, recall, f1-score, and area under the receiver operating characteristic (ROC) curve (AUC)—the researcher developed a novel weight allocation strategy for this study. GoogLeNet, ResNet-18, and DenseNet-121 were the three base CNN models to which the optimal weights were assigned. Prior research has mostly focused on classification accuracy when determining base learner weights, which may not be sufficient, especially when dealing with datasets that are not evenly distributed in class. Also, other criteria might offer more useful data for ranking the base learners in order of importance. Figure 2 presents the suggested ensemble framework's overall workflow.

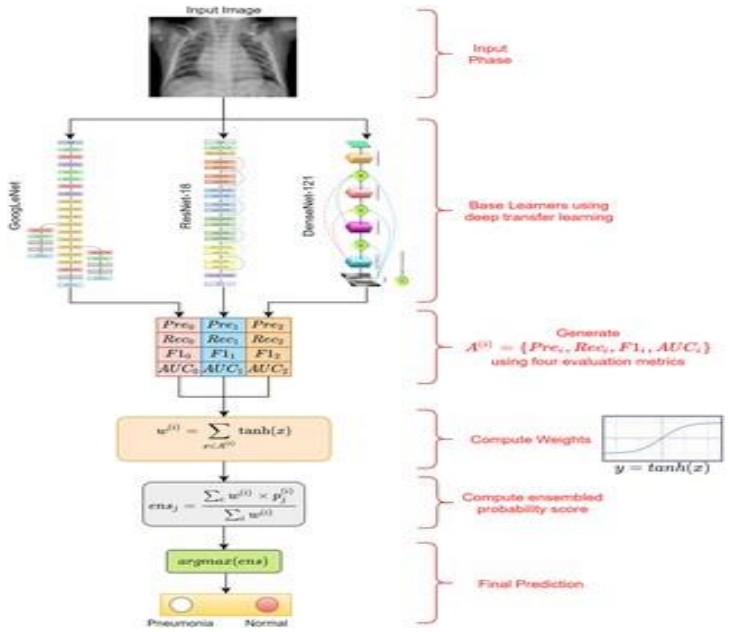

Fig 2. Representation of the proposed pneumonia detection framework.

Pre = Precision score, Rec = Recall score, F1 = F1-score, AUC = AUC score, and A(i) = {Prei, Reci, F1i, AUCi}; w(i) is the weight generated for the i^{th} base learner to compute the ensemble, is the probability score for the j^{th} sample by the i^{th} classifier, and enjoy is the fused probability score for the j^{th} sample; and the arg max function returns the position having the highest value in a 1D array, i.eIn this case, it generates the predicted class of the sample.

2. RELATED WORK

Table 1. Existing methods for pneumonia detection

Method	Approach	Merits	Demerits
Alhoshli et al. [9]	• Transfer Learning using InceptionResNet-V2	Reuse of models pretrained on a large dataset	Oversimplified for a complex pattern recognition task; Performance obtained is poor and not fit for practical use
Rahman et al. [10]	• Transfer Learning using DenseNet-201		
Liang et al. [11]	• Transfer learning using ResNet-50 pre-trained on ChestX-ray14 dataset		
Ibrahim et al. [12]	• Transfer learning using AlexNet		
Zubair et al. [13]	• Transfer learning using VGG-16		
Rajpurkar et al. [14]	• Transfer learning using DenseNet-121		
Alhoshli et al. [15]	• Used generative adversarial networks to generate synthetic data. • Classification using ResNet-152	Generation of synthetic data to balance the classes of the data because medical data are scarce	Classification results (41% accuracy rate) are not fit for practical use
Chandra et al. [16]	• Segmentation of lung X-rays using image processing • Extraction and classification of eight statistical features	Segmentation of lungs before classification allows localization of the disease	The use of handcrafted features limits its ability to perform in complex pattern recognition tasks; Evaluation on a small dataset (412 images) cannot be generalized
Kuo et al. [17]	• Used 11 features from patient data to fit traditional classifiers	Use of 10-fold cross validation with 3 repeats avoids overfitting	Patient data are often private and not publicly available to fit to classification models
Yue et al. [18]	• Segmented lung lobes using U-Net; • Extracted and classified radiomic features from CT-scan images	Segmentation before classification helps extract important features for radiologists and allows localization of the disease	Method evaluated on a small dataset (72 lesion segments) and thus difficult to generalize
Sharma et al. [19]	• Devised a CNN model for classification of X-ray images	Automatic feature learning for complex tasks	Simple linearly progressing CNN model increases computation cost without providing strong boost to performance
Stephen et al. [20]	• Developed a simple seven-layer CNN model for classification of X-ray images		
Ieracitano et al. [21]	• Developed a deep learning framework based on adversarial optimization	Adversarial optimization removed dependency on the source of the dataset and view of the X-rays for classification	Results (AUC 74.7%) are not fit for deployment in the field
Zhang et al. [22]	• Developed a confidence-aware module for anomaly detection in lung X-ray images	Posing the detection task as a one-class problem helped improve the model performance	The sensitivity obtained on the dataset was too low (71.70%) for practical use
Tuncer et al. [23]	• Applied fuzzy tree transformation to X-ray images • Extracted local features for classification using an ensemble of traditional classifiers	Generation of three different feature images improves the model performance	Handcrafted feature extraction limits performance in complex pattern recognition tasks; Evaluation on a small dataset cannot be generalized
Jaiswal et al. [24]	• Developed a mask region-based CNN for segmentation • Used an ensemble model for image thresholding	Use of threshold value in background boosts the performance	An irregular trend was observed, where results of the training set were lower than those of the testing set
Gabruseva et al. [25]	• Localized pulmonary opacity based on a single-shot detector • Used a snapshot ensemble model for segmentation	One-shot detector alleviates the problem of scarcity of data	Irregular trend of validation loss over epochs during model training
Pan et al. [26]	• Used an ensemble of Inception-ResNet-v2, XceptionNet, and DenseNet-169 for bounding box prediction	Ensemble learning allows the fusion of salient properties of all its base learners	Pan et al. [26] suspect that their model evaluated on only one dataset may not generalize over data acquired from a different source

3. MOTIVATIONS AND CONTRIBUTIONS

Many people, particularly children, suffer greatly from pneumonia. This condition is most common in developing and impoverished nations when risk factors such as overcrowding, poor hygiene, hunger, and a lack of proper medical facilities are present. It takes an early diagnosis to fully recover from pneumonia. The most popular diagnostic technique is X-ray examination, however, it depends on the radiology's interpretive skills, which frequently causes radiologists to disagree. For an accurate diagnosis, then, a generalisation-capable automated computer-aided diagnosis (CAD) system is required. The majority of earlier research ignored the possible advantages of ensemble learning in favour of creating a single convolutional neural network (CNN) model for the categorization of pneumonia. Better predictions are made possible by ensemble learning, which combines discriminative data from several base learners. Ensemble learning was used in this study to address a lack of medical data by using transfer learning models as base learners and ensembling their decision scores.

An ensemble framework was developed to boost the performance of base CNN learners in pneumonia classification by adopting a weighted average ensemble technique. The weights allocated to the classifiers were established by combining four assessment metrics: precision, recall, f1-score, and area under the curve (AUC), using a hyperbolic tangent function, as opposed to merely depending on classifier accuracy or experimental outcomes. Using five-fold cross-validation, the suggested model was assessed on two publicly accessible chest X-ray datasets: the RSNA Pneumonia Detection Challenge dataset and the Kermany dataset. The outcomes surpassed the most advanced techniques, indicating the method's feasibility for real-world implementations.

4. PROPOSED METHOD

In this study, I designed an ensemble framework consisting of three classifiers: GoogLeNet, ResNet-18, and DenseNet-121, using a weighted average ensemble scheme. The weights allocated to the classifiers were generated using a novel scheme, as explained in detail below.

4.1. GoogLeNet

Proposed by Szegedy et al., the GoogLeNet architecture is a 22-layer deep network that uses "inception modules" rather than uniformly progressive layers. By hosting parallel convolution and pooling layers, an inception block can support a large number of units at each level. However, the increased number of parameters results in an uncontrollable computational complexity. Instead of using the naive inception block (Fig. 3(a)) as utilised in earlier work, the GoogLeNet model uses inception blocks with dimension reduction, as illustrated in Fig. 3(b), to regulate the computational complexity. An ideal sparse architecture constructed from readily available dense building blocks enhances the performance of artificial neural networks for computer vision tasks, as demonstrated by the performance of GoogLeNet, which introduced the inception block. The GoogLeNet model's architecture is displayed in Fig 4.

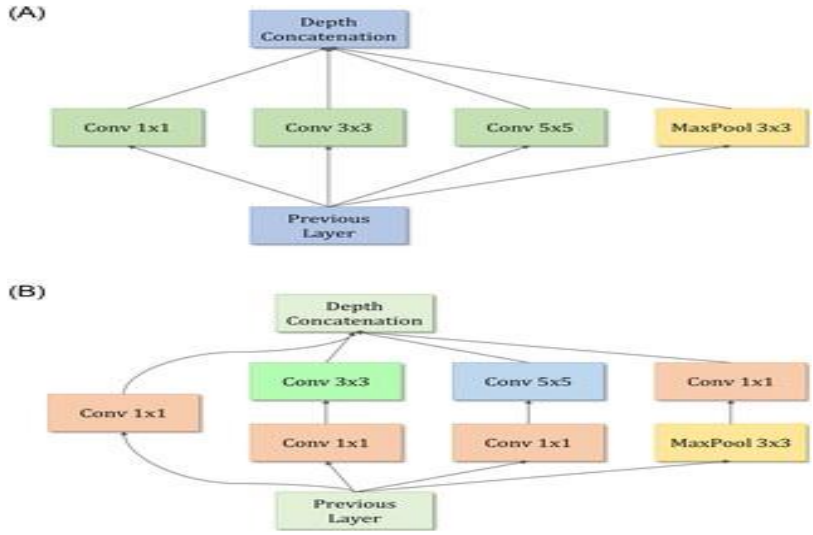

Fig 3. Inception modules in the GoogLeNet architecture. (a) The naive inception block is replaced by (b) the dimension reduction inception block in the GoogLeNet architecture to improve computational efficiency.

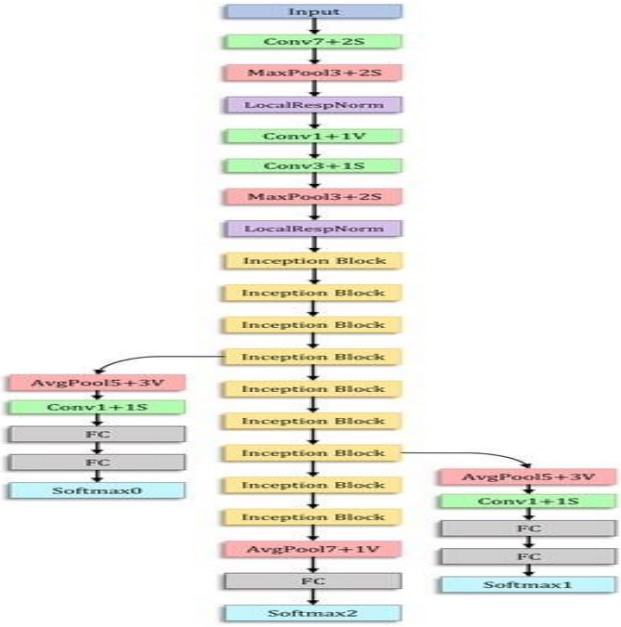

Fig 4. The architecture of the GoogLeNet model was used in the study

The inception block is shown in Fig 3(b).

4.2. ResNet-18

The residual learning approach that underpins Huang et al.'s ResNet-18 model boosts the effectiveness of deep network training. ResNet models' residual blocks help optimise the network as a whole, which increases model accuracy. This is not the same as the original unreferenced mapping found in convolutions that continue inversely. Identity mapping is carried out by these residuals or linkages, which neither adds parameters nor raises computational complexity. Figure 5 displays the ResNet-18 model's design.

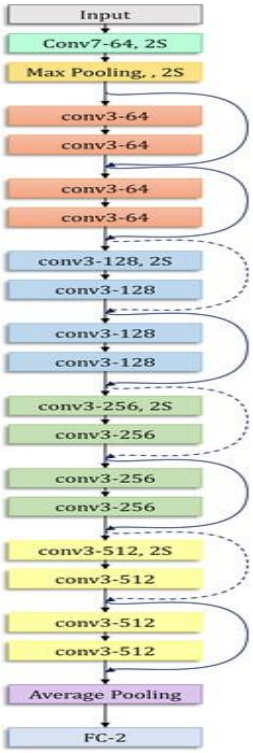

Fig 5. The architecture of the ResNet-18 model used in this study

4.3. DenseNet-121

Huang et al. suggested that DenseNet architectures are computationally efficient and offer a rich feature representation. The main explanation is that, as Fig. 6 illustrates, feature maps from each layer of the DenseNet model are concatenated with feature maps from all previous layers. The model becomes computationally efficient when the number of trainable parameters is decreased due to the convolutional layers' ability to handle fewer channels. Moreover, the feature representation capability is improved by concatenating the feature maps from earlier layers with the present layer.

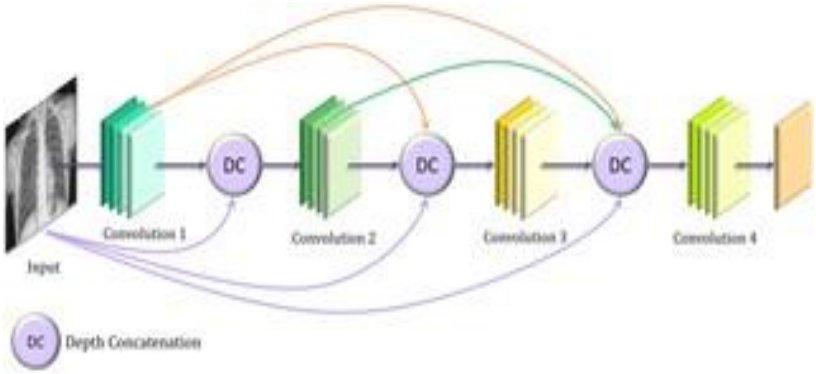

Fig 6. The basic architecture of the DenseNet convolutional neural network model.

The values of the hyperparameters used for training the learning algorithms (base learners) were set empirically and are shown in Table 2.

Table 2. Hyperparameters are used for training the convolutional neural network base learners.

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

5. PROPOSED ENSEMBLE SCHEME

The ensemble learning model helps incorporate the discriminative information from all its constituent models, resulting in superior predictions compared to any of its base learners. Weighted average ensemble is a powerful classifier fusion mechanism. However, the choice of weights allocated to the respective base learners plays a pivotal role in ensuring the success of the ensemble. Most approaches in the literature set the weights experimentally or solely based on the classifier's accuracy. However, this approach may not be suitable when the dataset exhibits class imbalance. The use of other evaluation measures, such as precision, recall (sensitivity), f1-score, and area under the curve (AUC), may provide more robust information for determining the priority of the base learners. To this end, this study devised a novel strategy for weight allocation, which is explained below.

First, the probability scores obtained during the training phase by the base learners are utilised to calculate the weights assigned to each base learner using the proposed strategy. These generated weights are used in the formation of an ensemble trained on the test set. This strategy is implemented to ensure that the test set remains independent for predictions. The predictions of the i^{th} model $\hat{y}^i$ are generated and compared with the true labels (y) to generate the corresponding precision score (pre^i), recall score (rec^i), f1-score ($f1^i$), and AUC score (AUC^i). Assume that this forms an array $A^i = \{pre^i, rec^i, f1^i, AUC^i\}$. The weight ($w^i$) assigned to each classifier is then computed using the hyperbolic tangent function, as shown in Eq 1. The range of the hyperbolic tangent function is [0, 0.762] because x represents an evaluation metric, the value of which is in the range [0, 1]. It monotonically increases in this range; thus, if the value of a metric x is high, the tanh function rewards it by assigning it a high priority; otherwise, the function penalises it.

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, where the probability array (for a binary class dataset) of the j^{th} test sample by the i^{th} base classifier is, where $a \leq 1$ and the ensemble probability for the sample is $ensemble\_prob_j = \{b, 1 - b\}$. Finally, the class predicted by the ensemble is computed by Eq 3, where $prediction_j$ denotes the predicted class of the sample.

6. RESULTS AND DISCUSSION

The evaluation findings of the suggested method are shown in this section. I used two datasets of chest X-rays for pneumonia that are publicly available. The first dataset is called the Kermany dataset, and it consists of 5856 chest X-ray images that are unequally distributed between the "Normal" and "Pneumonia" classes. The photos are from a broad population of adults and children. The second dataset was made available by the RSNA and presented as a pneumonia detection Kaggle challenge. The image distribution between the two datasets and the image

descriptions for the training and testing sets for each fold of the 5-fold cross-validation strategy used in this study are shown in Table 3. Moreover, the consequences of the acquired outcomes are examined. A comparative analysis was done to show how much better the proposed method is over other models and frequently used ensemble techniques published in the literature.

Table 3. Description of images in the training and testing sets in each fold of five-fold cross-validation in the two datasets used in this study.

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	(189×690×3)—(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	6001		
	Test	Lung Opacity	6111	(1024×1024×3)	224×224×3
		No Lung Opacity	1201		
	Total Images		26601		

7. EVALUATION METRICS

Four common evaluation measures were applied to the two pneumonia datasets to assess the suggested ensemble method: f1-score (F1), accuracy (Acc), precision (Pre), and recall (Rec). First, I define the phrases "True Positive," "False Positive," "True Negative," and "False Negative" to define these evaluation measures. Now, the four evaluation metrics can be defined as:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Pre = \frac{TP}{TP + FP} \quad Rec\ (or\ Sensitivity) = \frac{TP}{TP + FN}$$

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

Fig 7. The different evaluation metrics of the pneumonia detection ensemble model using components of the confusion matrix make up the evaluation metrics.

The accuracy rate gives a general indication of how many of the model's predictions came true. However, if the dataset is unbalanced, a model's high accuracy rate does not guarantee that it can distinguish between different classes equally. Specifically, a universally applicable model is needed for medical picture classification. In these situations, the model's performance can be understood by examining the "precision" and "recall" variables. The precision of the model's positive label prediction is shown. This gives the ratio of the right forecasts to all of the predictions the model produced. On the other hand, "recall" quantifies the proportion of positive ground truth data that the model accurately anticipated. These two assessment criteria determine whether the model can lower the quantity of FN and FP forecasts. The "F1-Score," which takes into account both FPs and FNs, offers a compromise between "precision" and "recall." Extreme values of "precision" and "recall," which are attained at the expense of one another, are penalised. For this reason, evaluation metrics—rather than only accuracy rate—should be taken into account in medical image classification to accurately identify both healthy and sick individuals.

8. IMPLEMENTATION

A 5-fold cross-validation strategy was used in this work to assess the performance of the suggested ensemble model in detail. The results for the Kermany dataset and the RSNA challenge dataset, respectively, are shown in Tables 4 and 5, together with the average and standard deviation values over all five folds. The suggested method's dependability is demonstrated by the excellent accuracy and sensitivity (recall) scores. Moreover, the confusion matrices on the RSNA and Kermany datasets are shown in Figures 7 and 8, and the ROC curves produced by the suggested method for each of the two datasets' five cross-validation folds are shown in Figure 8.

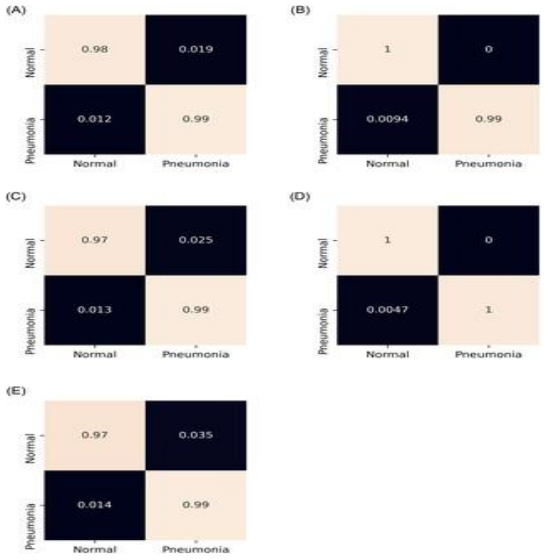

Fig 8. Confusion matrices were obtained on the Kermany pneumonia chest X-ray dataset by the proposed method by 5-fold cross-validation. a) Fold-1. (b) Fold-2. (c) Fold-3. (d) Fold-4. (e) Fold-5.

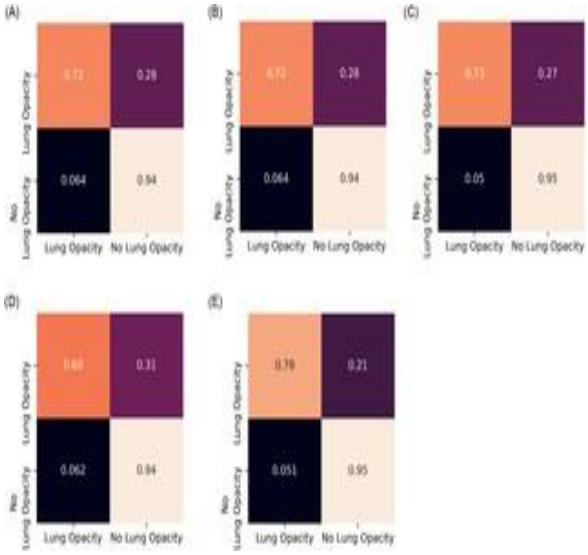

Fig 9. Confusion matrices obtained on the Radiological Society of North America pneumonia challenge chest X-ray dataset by the proposed method by five-fold cross-validation. a) Fold-1. (b) Fold-2. (c) Fold-3. (d) Fold-4. (e) Fold-5.

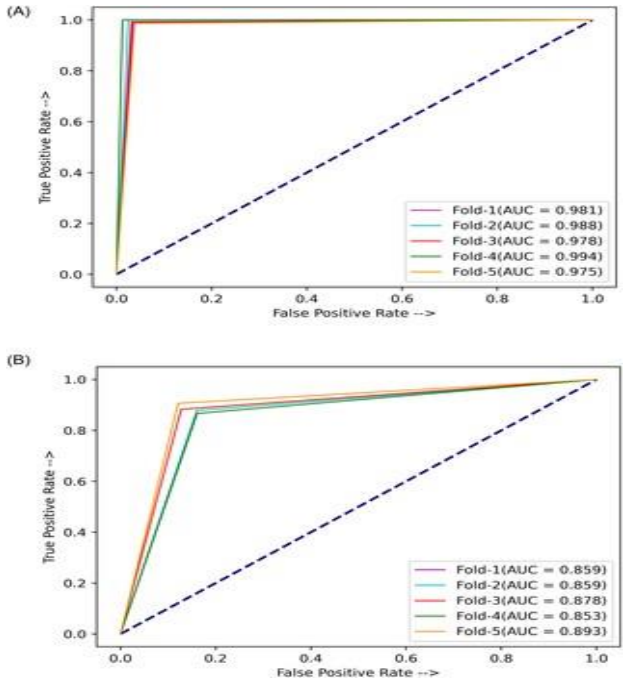

Fig 10. Receiver operating characteristic curves obtained by the proposed ensemble method on the two pneumonia chest X-ray datasets used in this research:(a) Kermany dataset [2]. (b) RSNA challenge dataset [16].

Table 4. Results of five-fold cross-validation of the proposed ensemble method on the pneumonia Kermany dataset [2].

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**

Avg: average *Std.Dev*: Standard Deviation.

Table 5. Results of five-fold cross-validation of the proposed ensemble method on the pneumonia Radiological Society of North America challenge dataset.

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**

Avg: average *Std.Dev*: Standard Deviation.

Figure 10 showcases the accuracy rates achieved by the base learners in transfer learning using different optimizers on the Kermany dataset. The Adam optimizer yielded the best results for all

three base learners and was consequently chosen as the optimizer for training the base learners in the ensemble framework.

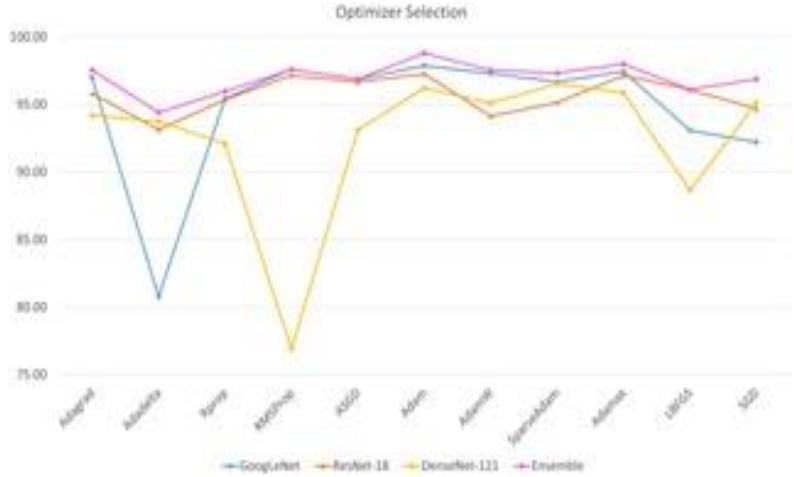

Fig 11. Variation of accuracy rates on the Kermany dataset [2]) was achieved by the three base learners, GoogLeNet, ResNet-18, and DenseNet-121, and their ensemble, according to the optimizers chosen for fine-tuning.

Table 6 shows the outcomes of several ensembles with three distinct base learners, including newly suggested architectures on the Kermany dataset: GoogLeNet, ResNet variants, DenseNet variations, MobileNet v2, and NASMobileNet. The outcomes support the selection of the GoogLeNet, ResNet-18, and DenseNet-121 base learner combinations utilised in this investigation. The accuracy rate of this ensemble combination was 98.2%. With an accuracy rate of 98.54%, the ensemble comprising GoogLeNet, ResNet-18, and MobileNet v2 achieved the second-best result. Furthermore, some layers were fixed for the selected set of base learners, and the models were trained to determine the best configuration. Figure 12 presents the results, which show that the ensemble performed best when every layer was trainable (0 layers frozen) on both datasets. Due to this, the exact setting was chosen for the ensemble framework.

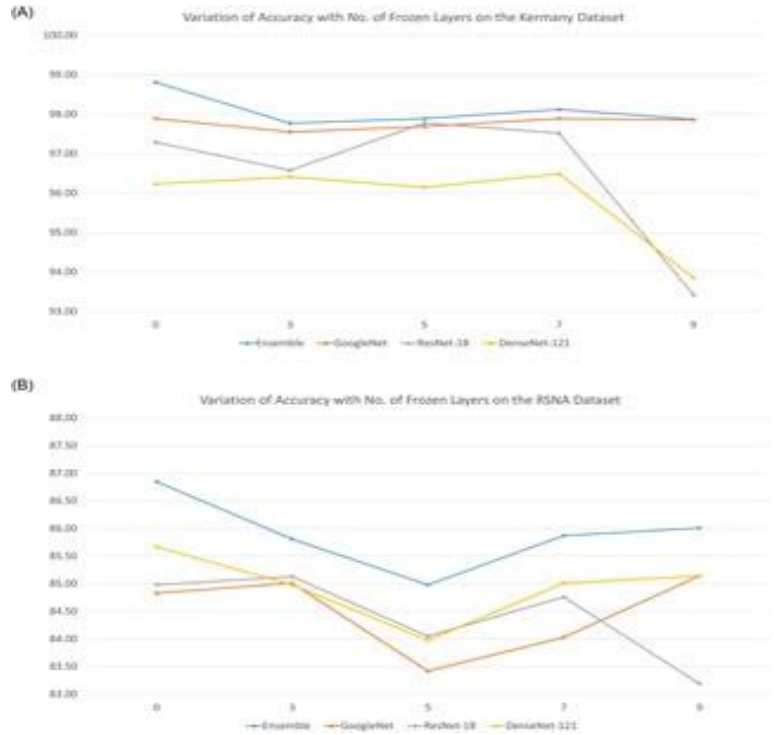

Fig 12. Variation in performance (accuracy rates) of the ensemble concerning the number of fixed non-trainable layers in the base learners on the two datasets used in this study:(a) Kermany dataset [2]. (b) RSNA challenge dataset [16].

Table 6. Results of extensive experiments performed to determine the base learners for forming 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**

8.1. Comparison with State-of-the- Art Methods

Table 7 compares the performance of the proposed ensemble framework and those of the existing methods in the literature on the Kermany pneumonia dataset. It should be noted that the proposed method outperformed all the other methods. It is also noteworthy that all these previous methods (Mahmud et al. [18], Zubair et al. [8], Stephen et al. [15], Sharma et al. [14], and Liang et al. [6]) revolved around using a single CNN model for the classification of pneumonic lung X-ray images and that the proposed ensemble framework outperformed them, indicating that the ensemble technique devised in this study is a reliable method for the image

classification task under consideration. To the best of our knowledge, no studies on the classification of images in the RSNA pneumonia dataset exist. Hence, for this dataset, I compared the performance of the proposed model to that of several baseline CNN models.

Table 7. Comparison of the proposed method with other methods in the literature on the Kermany pneumonia dataset [2] and the Radiological Society of North America challenge dataset [16].

Dataset	Method	Acc(%)	Pre(%)	Rec(%)	F1(%)	AUC(%)
Kermany	Mahmud et al. [39]	98.10	98.00	98.50	98.30	-
	Zubair et al. [13]	96.60	97.20	98.10	97.65	-
	Stephen et al. [20]	93.73	-	-	-	-
	Sharma et al. [19]	90.68	-	-	-	-
	Liang et al. [11]	90.50	89.10	96.70	92.70	-
	Proposed Method	**98.81**	**98.82**	**98.80**	**98.79**	**98.35**
RSNA	Antin et al. [40]	-	-	-	-	61.00
	Zhou et al. [41]	79.70	-	-	80.00	-
	Yao et al. [42]	-	-	-	-	71.30
	Rajpukar et al. [14]					76.80
	Proposed Method	**86.85**	**86.89**	**87.02**	**86.95**	**86.85**

Table 8 compares the assessment results of the proposed technique with those of the basic CNN models used to construct the ensemble and various other conventional CNN transfer learning models on both datasets utilised in this work. On both datasets, it is evident that the suggested ensemble approach did rather well compared to alternative transfer learning models and the base learners. In addition, Table 9 compiles the findings to demonstrate the superiority of the suggested ensemble scheme over conventional popular ensemble strategies. For both the Kermany and RSNA challenge datasets, the average results across the five folds of cross-validation are displayed. The ensembles employed the same three basic CNN learners, GoogLeNet, ResNet-18, and DenseNet-121. Popular ensemble techniques were outperformed by the suggested ensemble approach. It is evident from both datasets that the weighted average ensemble that uses the accuracy metric as the sole weighting factor produced the best results, nearly matching the suggested ensemble approach. The class that received the most votes from the base learners is expected to be the sample class in the majority voting-based ensemble. In the maximum probability ensemble, all base learners' probability scores are added up, and the class with the highest probability is designated as the sample's predicted class. In the average probability ensemble, on the other hand, each contributing classifier is given the same weight.

Table 8. Comparison of the proposed ensemble framework with several standard convolution neural network models in the literature on both the Kermany and the Radiological Society of North America challenge datasets.

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	85.13	84.86	85.00	85.86
	VGG-16	81.05	80.08	81.17	80.62	81.05
	DenseNet-121	85.67	84.98	85.35	85.26	85.67
	ResNet-18	84.98	85.35	85.37	85.46	84.98
	Proposed Method	**86.85**	**86.89**	**87.02**	**86.95**	**86.85**

In addition, Table 9 compiles the findings to demonstrate the superiority of the suggested ensemble scheme over conventional popular ensemble strategies. For both the Kermany and RSNA challenge datasets, the average results across the five folds of cross-validation are displayed. The ensembles employed the same three basic CNN learners, GoogLeNet, ResNet-18, and DenseNet-121. Popular ensemble techniques were outperformed by the suggested ensemble approach. It is evident from both datasets that the weighted average ensemble that uses the accuracy metric as the sole weighting factor produced the best results, nearly matching the suggested ensemble approach. The class that received the most votes from the base learners is expected to be the sample class in the majority voting-based ensemble.

Regarding the maximum probability ensemble, each class's probability scores are added up across all base learners, and the class with the highest probability is designated as the sample's predicted class. In contrast, each contributing classifier in the average probability ensemble is given the same weighting.

Table 9. Performance comparison of the proposed ensemble technique and popular ensemble schemes in the literature for the two datasets used.

Dataset	Ensemble technique	Acc(%)	Pre(%)	Rec(%)	F1(%)	AUC (%)
Kermany	Maximum Probability	97.77	97.84	97.79	97.76	97.79
	Average Probability	97.85	97.81	97.79	97.78	97.81
	Majority Voting	98.11	98.13	98.12	98.10	97.85
	Weighted Average with only accuracy	98.20	98.22	98.20	98.18	98.11
	Proposed Ensemble	**98.81**	**98.82**	**98.80**	**98.79**	**98.35**
RSNA	Maximum Probability	85.67	85.60	85.67	85.73	85.67
	Average Probability	86.10	85.98	86.11	86.04	86.11
	Majority Voting	85.98	85.67	85.98	85.82	85.98
	Weighted Average with only accuracy	86.63	86.54	86.63	86.58	86.54
	Proposed Ensemble	**86.85**	**86.89**	**87.02**	**86.95**	**86.85**

The same base learners were used in all the ensembles: GoogLeNet, ResNet-18, and DenseNet-121.

9. CONCLUSION AND FUTURE WORK

To treat pneumonia appropriately and keep the patient's life from being in danger, early recognition of the illness is essential. The most common method for diagnosing pneumonia is a chest radiograph; however, there can be inter-class variation in these images, and the diagnosis relies on the doctor's skill to identify early signs of pneumonia. In this work, an automated CAD system was created to aid medical professionals. It classifies chest X-ray pictures into two groups, "Normal" and "Pneumonia," using deep transfer learning-based classification. A weighted average ensemble is formed by an ensemble framework that takes into account the decision scores from three CNN models: GoogLeNet, ResNet-18, and DenseNet-121. An innovative technique was used to calculate the weights assigned to the classifiers, whereby five evaluation parameters, accuracy, precision, recall, f1-score, and AUC, were fused using the hyperbolic tangent function. The framework, evaluated on two publicly available pneumonia chest X-ray datasets, obtained an accuracy rate of 98.2%, a sensitivity rate of 98.19%, a precision rate of 98.22%, and an f1-score of 98.29% on the Kermany dataset and an accuracy rate of 86.7%, a sensitivity rate of 86.62%, a precision rate of 86.69%, and an f1-score of 86.65% on the RSNA challenge dataset, using a five-fold cross-validation scheme. It outperformed state-of-the-art methods on these two datasets. Statistical analyses of the proposed model using McNemar's and ANOVA tests indicate the viability of the approach. Furthermore, the proposed ensemble model is domain-independent and thus can be applied to a large variety of computer vision tasks.

However as was already indicated, there were times when the ensemble framework was unable to generate accurate forecasts. In the future, I might look into methods like image contrast enhancement or other pre-processing procedures to raise the quality of the photographs. To help the CNN models extract more features from the lung image, I suggest segmenting the image before classifying it. Moreover, the computing cost of the suggested ensemble is greater than that of the CNN baselines created in studies published in the literature since three CNN models are needed to train it. In the future, I might try using techniques like snapshot ensembling to try and lower the processing requirements.

ACKNOWLEDGEMENTS

I would like to acknowledge my parents for buying me the computer on which I did all of my research.

REFERENCES

[1] WHO Pneumonia. World Health Organization. (2019), https://www.who.int/news-room/fact-sheets/detail/pneumonia

[2] Kermany D., Zhang K. & Goldbaum M. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. (Mendeley,2018)

[3] Dalhoumi S., Dray G., Montmain J., Derosière, G. & Perrey S. An adaptive accuracy-weighted ensemble for inter-subjects classification in brain-computer interfacing. 2015 7th International IEEE/EMBS Conference On Neural Engineering (NER). pp. 126-129 (2015)

[4] Albahli S., Rauf H., Algosaibi A. & Balas V. AI-driven deep CNN approach for multi-label pathology classification using chest X-Rays. PeerJ Computer Science. 7 pp. e495 (2021) pmid:33977135

[5] Rahman T., Chowdhury M., Khandakar A., Islam K., Islam K., Mahbub Z., et al. Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. Applied Sciences. 10, 3233 (2020)

[6] Liang G. & Zheng L. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. Computer Methods And Programs In Biomedicine. 187 pp. 104964 (2020) pmid:31262537

[7] Ibrahim A., Ozsoz M., Serte S., Al-Turjman F. & Yakoi P. Pneumonia classification using deep learning from chest X-ray images during COVID-19. Cognitive Computation. pp. 1–13 (2021) pmid:33425044

[8] Zubair S. An Efficient Method to Predict Pneumonia from Chest X-Rays Using Deep Learning Approach. The Importance Of Health Informatics In Public Health During A Pandemic. 272 pp. 457 (2020)

[9] Rajpurkar P., Irvin J., Zhu K., Yang B., Mehta H., Duan T., et al. & Others Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. ArXiv Preprint ArXiv:1711.05225. (2017)

[10] Albahli S., Rauf H., Arif M., Nafis M. & Algosaibi A. Identification of thoracic diseases by exploiting deep neural networks. Neural Networks. 5 pp. 6 (2021)

[11] Chandra T. & Verma K. Pneumonia detection on chest X-Ray using machine learning paradigm. Proceedings Of 3rd International Conference On Computer Vision And Image Processing. pp. 21-33 (2020)

[12] Kuo K., Talley P., Huang C. & Cheng L. Predicting hospital-acquired pneumonia among schizophrenic patients: a machine learning approach. BMC Medical Informatics And Decision Making. 19, 1–8 (2019) pmid:30866913

[13] Yue H., Yu Q., Liu C., Huang Y., Jiang Z., Shao C., et al. & Others Machine learning-based CT radiomics method for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: a multicenter study. Annals Of Translational Medicine. 8 (2020) pmid:32793703

[14] Sharma H., Jain J., Bansal P. & Gupta S. Feature extraction and classification of chest x-ray images using cnn to detect pneumonia. 2020 10th International Conference On Cloud Computing, Data Science & Engineering (Confluence). pp. 227-231 (2020)

[15] Stephen O., Sain M., Maduh U. & Jeong D. An efficient deep learning approach to pneumonia classification in healthcare. Journal Of Healthcare Engineering. 2019 (2019) pmid:31049186

[16] Wang X., Peng Y., Lu L., Lu Z., Bagheri M. & Summers R. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localTocommon thorax diseases. Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition. pp. 2097-2106 (2017)

[17] Selvaraju R., Cogswell M., Das A., Vedantam R., Parikh D. & Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings Of The IEEE International Conference On Computer Vision. pp. 618-626 (2017)

[18] Mahmud T., Rahman M. & Fattah S. CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization. Computers In Biology And Medicine. 122 pp. 103869 (2020) pmid:32658740

[19] Dietterich T. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation. 10, 1895–1923 (1998) pmid:9744903

[20] Cuevas A., Febrero M. & Fraiman R. An anova test for functional data. Computational Statistics & Data Analysis. 47, 111–122 (2004)

AUTHOR

Prithvi is a driven high school student set to graduate in 2025 with an impressive academic record in computer science and STEM fields. His diverse pursuits, ranging from founding a global AI education platform to pioneering research in AI-powered medical imaging analysis, exemplify his innovative mindset and passion for developing socially impactful technology. With an unwavering determination to be at the forefront of ethical AI innovation, Prithvi aims to continue pushing boundaries and creating lasting positive impacts in the field.