Allied Medical Research Journal

Vol:4 No.3: 65-72

Clinical Utility

Hybrid Ensemble Learning for Multi-Class Chest X-Ray Classification Using Deep CNN

Abdul Rehman Khan Tareen[1], Muhammad Laiq Ur Rahman Shahid[1], Furqan Shaukat[1], Muhammad Hamza Zafar[2], Soban Abu Khifs[3]

[1]*Dept of Elecrical Engineering, University of Engineering and Technology Taxila, Pakistan*
[2]*Dept of Engineering Sciences, University of Agder, Grimstad, Norway*
[3]*Dept of Medical Surgery, Abbas Institute of Medical Sciences, Muzaffarabad, AJK, Pakistan*

ABSTRACT

Background: COVID-19, pneumonia, and TB (tuberculosis) are still the big killers of people suffering from chest disease and continue to be a serious health challenge globally. A timely diagnosis leads to timely treatment and improved patient outcomes. Chest X-ray (CXR) imaging is widely used for diagnostic purposes due to its speed, low cost, and availability in most healthcare facilities. Manual reading of CXR images is, however, challenging because of the similarity in the presentation of radiographic features across various chest diseases.

Methods: This research introduces a hybrid ensemble learning approach to classify chest X-ray images into four classes—Normal, COVID-19, Pneumonia, and Tuberculosis. Three Deep CNN network models, namely Xception, AlexNet, and EfficientNet-B0, were used for deep feature extraction. Additionally, texture features of the images were extracted using Gabor filters. The deep and texture features were combined and classified using logistic regression and a stacking ensemble learning approach. A publicly available chest X-ray image database containing 7,135 X-rays was used, with six-fold stratified cross-validation to assess the proposed approach.

Results: The ensemble models outperformed the individual CNN models. The Average Ensemble produced the best results with an accuracy of 91.18%, an Average Precision (AP) of 96.75%, and an Area Under the ROC Curve (AUC) of 98.87%. The proposed model performs well across all four disease classes. It exhibited high sensitivity in detecting Tuberculosis with considerable stability in classifying Normal, COVID-19, and Pneumonia.

Conclusion: The proposed framework demonstrates the effectiveness of integrating deep learning features, Gabor texture features, and ensemble learning for improved chest X-ray image classification. This can help computer-assisted diagnostics systems and aid medical workers in identifying chest diseases early.

Keywords: Chest X-ray, Deep Learning, Convolutional Neural Networks, Ensemble Learning, Medical Image Classification, Tuberculosis, Pneumonia.

Received: June 25th, 2026**; Revised:** July 23rd, 2026**; Accepted:** July 28th, 2026

Corresponding Email: khan0229508@gmail.com

DOI: https://doi.org/10.59564/amrj/04.03/000

INTRODUCTION

Chest diseases, including pneumonia, COVID-19, and tuberculosis (TB), remain among the leading causes of illness and death worldwide. The burden of these diseases is particularly high in developing countries, where access to advanced healthcare facilities may be limited. Early and accurate diagnosis is crucial for effective treatment and disease control. Chest X-ray (CXR) imaging is one of the most commonly used diagnostic tools for chest disease assessment because of its low cost, rapid acquisition, and widespread availability in clinical practice.

The medical field has been greatly helped by many artificial intelligence (AI) technologies for medical image interpretation. One type of deep learning network widely used in chest disease recognition is a convolutional neural network, as it can learn discriminative features from CXRs[2-5]. A large body of research has employed deep learning & ensemble learning principles for tuberculosis detection, achieving strong classification performance[1-4].

Transfer learning streamlines convolutional neural network training by leveraging disease-specific features from large-scale medical datasets, such as Xception and EfficientNet[5-10]. Furthermore, there have been attempts to improve the model's performance in terms of accuracy and generalization across various projects in the context of image improvement and explainable artificial intelligence[6,11,12].

Recent research indicates that using multiple deep learning models to classify chest X-rays can improve reliability. The purpose of ensemble learning is to overcome the drawbacks of single learning models by combining model predictions, thereby achieving better generalization and classification performance. Besides, hybrid models that combine deep learning with other feature extraction methods have also been shown to enhance the detection of chest diseases, particularly those with similar radiographic features[25,26].

However, a few studies on this topic continue to focus on classifying just two classes, most prominently the distinction between healthy and TB patients' CXRs[1-8]. On the other hand, several methods in the literature rely on a single CNN without considering the advantages of using hand-made textures for the classification problem. Since each of the three diseases, COVID-19, Pneumonia, and TB, has different radiological characteristics, deep + texture features can increase the ability to recognize the disease.

In this paper, a framework for classifying 4 types of chest diseases is proposed that combines texture and deep features using deep learning and ensemble learning techniques. Deep features such as Xception, AlexNet, and Efficient Net will be processed with Gabor filtering, and classification will be performed using logistic regression and a stacked ensemble for four classes (Normal, COVID-19, Pneumonia, and Tuberculosis).

Building on the above innovations, this research presents a hybrid model for automated chest disease classification that integrates transfer learning, Gabor texture analysis, feature fusion, and ensemble learning. The proposed method combines discriminative, complementary information from a variety of pre-trained CNN classifiers with hand-crafted texture features to improve classification performance. The framework was developed for computer-aided diagnosis and also helps health workers in promoting early detection of chest diseases[25,28].

CNN models are commonly used for chest X-ray classification, and some studies have applied ensemble learning methods for the same task; however, most studies have focused on binary classification or on a specific deep learning model. Some approaches use a combination of CNN models without paying much attention to texture information, thereby enhancing the representation of chest X-ray images. To achieve this, a hybrid model is proposed in this study that merges three pre-trained CNN deep features with Gabor deep texture features through deep feature fusion and ensemble. The proposed approach differs from many others in the literature in a few respects: it incorporates a single set of design rules for the classification problem of Normal, COVID-19, Pneumonia, and Tuberculosis. With this combination, the model obtains complementary deep and texture information, achieving good classification performance while preserving its simplicity and applicability.

Below are some chest images from 4 classes in this experiment. The Normal class represents clear lung images, and the COVID-19 and Pneumonia classes represent pulmonary opacities or infiltrates. However, tuberculosis pictures consist of localized and cavity appearance.

The key points about the contributions are highlighted below:

- Explores the possibility of a hybrid system in which Deep CNN features, Gabor texture features, feature fusion, and ensemble learning are incorporated into a four-class chest X-ray classification. The proposed system differs from previous studies, which focused on a few classes or used a single CNN model, by supporting multi-class disease classification within a single system.
- For feature extraction, we use deep learning models; in this case, we use Xception, AlexNet, and EfficientNet. For the texture feature, we use a Gabor filter. By using a combined algorithm that fuses deep and texture features, the classifier's discriminative ability can be enhanced.
- A total of 7,135 chest X-ray images are used to test the performance of the proposed framework. Evaluated metrics are Accuracy, Precision, Recall, F1-score, ROC-AUC, and Precision-Recall curves.

METHODOLOGY

The proposed framework is intended to categorize chest X-ray images into the Normal, COVID-19, Pneumonia, and Tuberculosis classes. It employs deep feature extraction using pre-trained CNN models, Gabor filter-based texture feature extraction, and stacked ensemble learning techniques.

The process flow, which indicates the steps of the proposed system, is shown in Figure 2 below. Preprocessed images will undergo two types of feature extraction techniques. Pre-trained CNN models will be used for deep feature extraction, and a Gabor filter will be used for texture feature extraction. Then, a Logistic regression Classifier will be trained using the obtained feature vectors. Finally, a stacked classifier based on ensemble learning will be used to obtain the classification result.

All chest X-ray images were preprocessed to ensure uniformity of the input to deep learning models before feature extraction. First, all images were resized to 224×224, the input size required by the selected CNN models. Normalization of Pixel values for improving training stability. Several data augmentation methods, such as random rotations, translations, and flips, were applied only to images in the training set to diversify the data and limit overfitting. This involved random rotation, horizontal reflections, magnification, and horizontal and vertical shifts. No efforts were made to enhance the testing images to provide a fair test of the proposed framework.

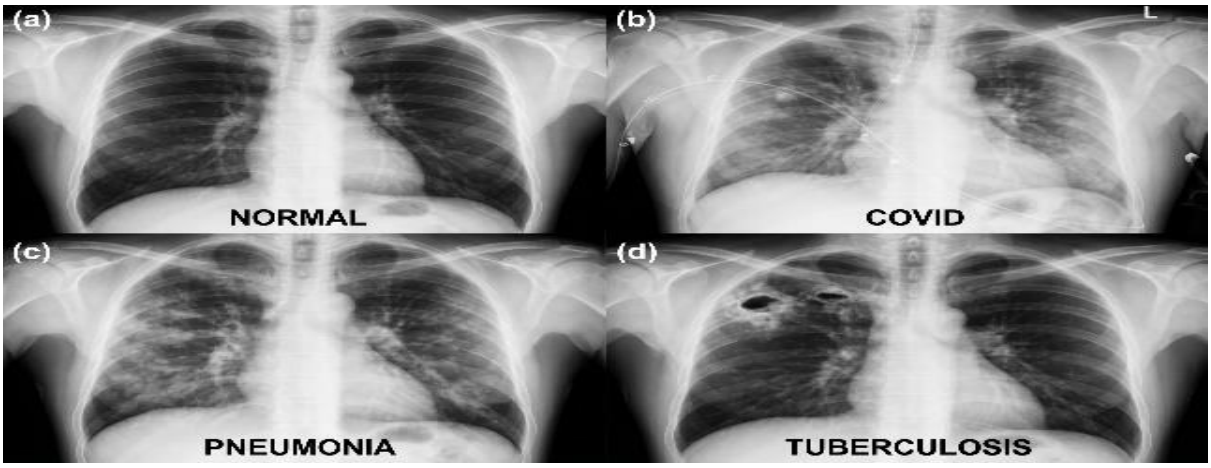

Figure 1: Representative chest X-ray images used in this study

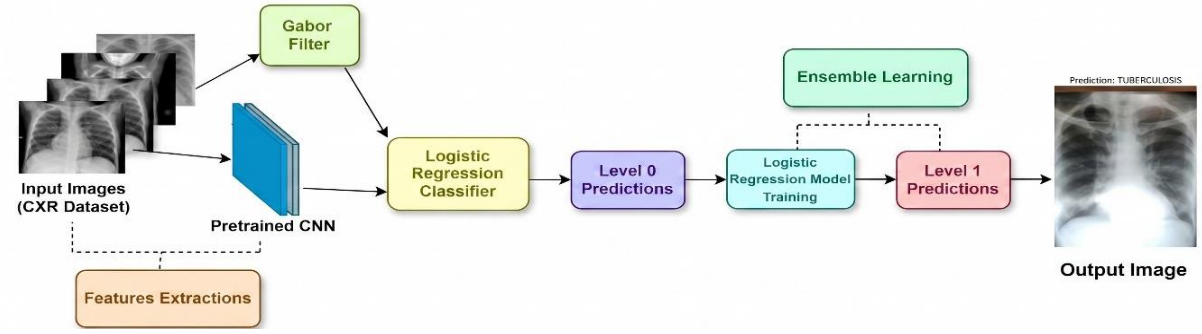

Figure 2: Block Diagram of the proposed method

The training set is augmented to improve the model's generalizability. The data augmentation techniques used are: Random rotation, Horizontal flipping, Zooming, Width shifting, and Height shifting. Those processes are applied to the training images to achieve a critical mass of diversity without losing the signal in the training data.

All deep learning models were initialized from an ImageNet-pretrained model and fine-tuned on the chest X-ray dataset. All images were scaled to 224x224 during training. The models were trained with the Adam optimizer at a learning rate of 0.0001 and a batch size of 32, using categorical cross-entropy loss and a Softmax activation for multi-class classification. They underwent training for 100 epochs, with the model achieving the highest validation score; this model was selected for further evaluation.

The models chosen were Xception, AlexNet, and EfficientNet-B0, each with a different network design and trained to learn different types of image features. Efficiently capturing complex patterns in chest X-ray images, Xception is effective at understanding the deep spatial features through depth wise separable convolutions. The AlexNet architecture is simple and well-known for efficiently learning low-level image features. Compound scaling of network depth, width, and resolution yields EfficientNet-B0, which offers a good trade-off between classification performance and computational cost. By combining three models, the proposed model learns complementary features from different network architectures, thereby enhancing the overall feature representation before feature fusion and ensemble learning.

For the three CNN models, different training approaches were used to improve training. Transfer learning was used to initialize Xception and EfficientNet-B0 with weights from a large-scale pre-trained model (ImageNet). In the first phase, the additional classification layers were trained, while the backbone networks were frozen. During fine-tuning, only the last 30 layers of both Xception and EfficientNet-B0 were unfrozen and updated, while the other layers, such as the Batch Normalization layers, were kept frozen. In contrast, AlexNet was trained from scratch, and all 10 trainable layers (5 convolutional layers, 2 Batch Normalization layers, and 3 fully connected layers) were optimized during training. This training procedure enabled the pre-trained models to fine-tune on the chest X-ray image dataset and to learn task-specific image features without relying on the existing knowledge or experience of AlexNet.

For each disease, class weights have been provided to ensure it has the same proportion of instances as the other classes in the training set. The heart and soul of the proposed framework is feature extraction. Both deep learning and texture-based feature extraction techniques are employed to provide a comprehensive feature set for chest X-rays.

Although deep learning models perform very well at learning semantic features of an image, texture characteristics are also of great importance, particularly in medical image applications such as the detection of pulmonary diseases.
In this study, texture features are learned from chest X-ray images using a two-dimensional Gabor filter. The mathematical equation of the Gabor filter can be expressed as:

$$G(x,y) = exp\left(-\frac{x'^2 + \Upsilon^2 y'^2}{2\sigma^2}\right) cos\left(2\pi\frac{x'}{\lambda} + \emptyset\right)$$

Where λ represents wavelength, $\emptyset$ ε represents orientation, π represents phase offset, σ represents Gaussian spread, and Υ represents spatial aspect ratio.

The final feature representation will be composed of CNN and Gabor features:

$$F_{Final} = [F_X, F_A, F_E, F_G]$$

Where F_G represents the Gabor texture features.
The features from the Xception, AlexNet, and EfficientNet-B0 models were stacked with Gabor texture features; this stacking was referred to as feature concatenation. This involves combining the feature vectors to create a representation of the whole without modifying the constituent feature values. The feature vector is a fusion of the deep learning features and the texture features. The fused features were then fed into the Logistic Regression classifier for disease diagnosis.

The feature vector extracted from each model had a different size. Before going into classification, the Global Average Pooling layer yielded a 2048-dimensional feature vector, and AlexNet yielded a 1024-dimensional feature vector from its last fully connected layer. Global Average Pooling outputs a 1280-dimensional feature vector from the input of EfficientNet-B0. The Gabor feature extraction stage generated a 48-dimensional vector of texture features based on statistical values of filter responses. These were all combined into a fused 4400-dimensional feature vector, which was used as input to the Logistic Regression classifier.

The set of features is combined into a single vector and used as input to a Logistic Regression classifier that produces class-label probabilities (i.e., the probability of chest disease).

$$L_o = [P_1, P_2, P_3, P_4]$$

Where P_1, P_2, P_3, and P_4 represent the probabilities of normal, Covid, Pneumonia, and Tuberculosis, respectively.

In this study, two ensemble learning strategies were tested: The Stacking Logistic Regression Ensemble and the Average Ensemble. For the stacking approach, all the prediction probabilities produced by the individual CNN models were considered as input to a Logistic Regression meta-classifier to learn how to combine predictions from the base learners. For prediction in the Average Ensemble, each CNN model's predicted probability was averaged, and the class with the highest average probability was selected. The performance of multiple classifiers in combination was evaluated to determine whether combining them would yield a more robust and overall better classification.

The Kaggle Chest X-ray dataset, a publicly available dataset, was used for all experiments and comprises chest X-ray images from four classes: Normal, COVID-19, Pneumonia, and Tuberculosis. In total, there are 7,135 images. For this, 1,341 Normal, 460 COVID-19, 3,875 Pneumonia, and 650 Tuberculosis images were used to train the model. Testing Set includes 234 images of Normal persons, 106 images of COVID-19 persons, 390 images of Pneumonia persons, and 41 images of TB persons. Furthermore, a validation set was used to develop the model, comprising 8 Normal, 10 COVID-19, 8 Pneumonia, and 12 Tuberculosis images. Class weights were used during training to mitigate the effects of class imbalance, as the Pneumonia class is much larger than the others.

TABLE 1: Distribution of the chest X-ray dataset used for training, validation, and testing.

Class	Training	Testing	Validation	Total
Normal	1341	234	8	1583
COVID-19	460	106	10	576
Pneumonia	3875	390	8	4273
Tuberculosis	650	41	12	703
Total	6320	771	38	7135

The distribution of images across the four classes is imbalanced, with the largest number of images in the Pneumonia class. For these reasons, when all deep learning models were trained, class weights were applied to reduce this imbalance during training. This enables the model to be successfully trained by both majority and minority classes. To make the training data more suitable for the model training and validation, a 6-fold stratified cross-validation was used to ensure the data in each fold had a similar class distribution.

This study does not include any patient identifiers in the publicly available Kaggle chest X-ray dataset. Hence, it was not possible to confirm patient-level separation between the training and validation sets. The experiments were conducted according to the specified data set structure given with the data set. Sixfold stratified cross-validation was used to train the models, creating folds with an equal distribution of the four disease classes, thereby minimizing the risk of biased performance estimates.

and AUC. All results from individual and ensemble models are summarized in Table 2 below.

TABLE 2: Performance comparison of the individual CNN models and ensemble methods.

Model	Accuracy (%)	AP (%)	AUC (%)
Xception	90.27	95.29	98.22
AlexNet	90.66	95.71	98.5
EfficientNet-B0	83.53	92.01	96.87
Stacking Logistic Regression Ensemble	91.05	95.88	98.19
Average Ensemble	91.18	96.75	98.87

Among individual CNN models, AlexNet achieved the highest accuracy, 90.66%, followed by Xception at 90.27%. As for the Ensemble models, the best accuracy was observed with the Average Ensemble Model at 91.18%, followed by the Stacking Logistic Regression Ensembles at 91.05%.

Based on the results, the ensembles have outperformed the CNNs. In particular, the Average Ensemble exhibited the best performance with an AP of 0.9675 and an AUC of 0.9887, thereby highlighting its outstanding classification capability.

Compared with CNN models, AlexNet performed best, achieving 90.66% accuracy, followed by Xception at 90.27%. Both models achieved a high level of accuracy (100%) when it came to diagnosing Tuberculosis (TB) cases. The EfficientNet-B0 model achieved the lowest accuracy, at 83.53%, but performed excellently in classifying COVID-19 cases.

The Average Ensemble, in terms of performance, had the highest accuracy of 91.18%, sensitivity of 93.20%, and specificity of 95.84%. Similarly, the Stacking Logistic Regression Ensemble achieved an accuracy of 91.05% and performed well across all diseases, particularly in the diagnosis of COVID-19 and TB.

To evaluate the individual effect of each component, an ablation study was conducted, in which different configurations of the suggested framework were investigated. Four experimental conditions were tested, including the effects of Gabor features, individual CNN models, feature fusion, and ensemble learning. The performance of each configuration with the identical evaluation settings is summarized in Table 3.

***TABLE 3: Ablation study showing the contribution of Gabor features, feature fusion, and ensemble learning to the proposed framework*.**

Model	Accuracy	Precision	Recall	F1-Score
Gabour Feature Only	0.7030	07103	0.7263	0.7038
EfficientNet-B0	0.8444	0.8616	0.8467	0.8489
AlexNet	0.8988	0.8872	0.9132	0.8950
Xception	0.9040	0.9120	0.9213	0.9130
CNN Stacking	0.9053	0.9174	0.9215	0.9151
Feature Fusion + Gabor + CNN	0.8923	0.8740	0.9032	0.8800
Average Ensemble (Proposed)	**0.9118**	**0.9301**	**0.92286**	**0.9263**

RESULTS

The results of all individual and ensemble CNN models were evaluated using accuracy, sensitivity, specificity, AP,

Results reveal that every element of the proposed framework contributed to the overall performance. The classification accuracy of the Gabor features was the lowest, indicating that the texture data alone could not enable effective disease classification. Of all the CNN models, Xception had the highest scores, followed by AlexNet and EfficientNet-B0. The classification results showed that the ensemble method further enhanced classification accuracy; the Average Ensemble achieved the best result at 91.18%. The results show that fusing complementary deep features with ensemble learning yields a stronger, more accurate model than either explicitly using a single feature or using ensemble learning alone.

The confusion matrix was analyzed to further assess the classification performance of the proposed model across all four disease types. A confusion matrix provides insight into correctly and misclassified samples and shows class-wise precision, recall, and F1-score to assess model performance on a per-class basis.

Confusion Matrix of the Proposed Average Ensemble Model

		Predicted Class			
		COVID19	NORMAL	PNEUMONIA	TUBERCULOSIS
Actual Class	COVID19	103	0	1	2
	NORMAL	1	181	50	2
	PNEUMONIA	0	12	378	0
	TUBERCULOSIS	0	0	0	41

Figure 3:Confusion matrix of the proposed Average Ensemble model.

It also illustrates that the proposed model achieved excellent accuracy in recognizing COVID-19 and tuberculosis COVID-19 cases, as indicated by the confusion matrix. The majority of misclassifications were misidentifying the Normal vs. Pneumonia classes, as they can appear very similar radiographically. The class-wise precision, recall, and F1 Scores also indicate that the proposed framework achieved consistent results across all four disease groups.

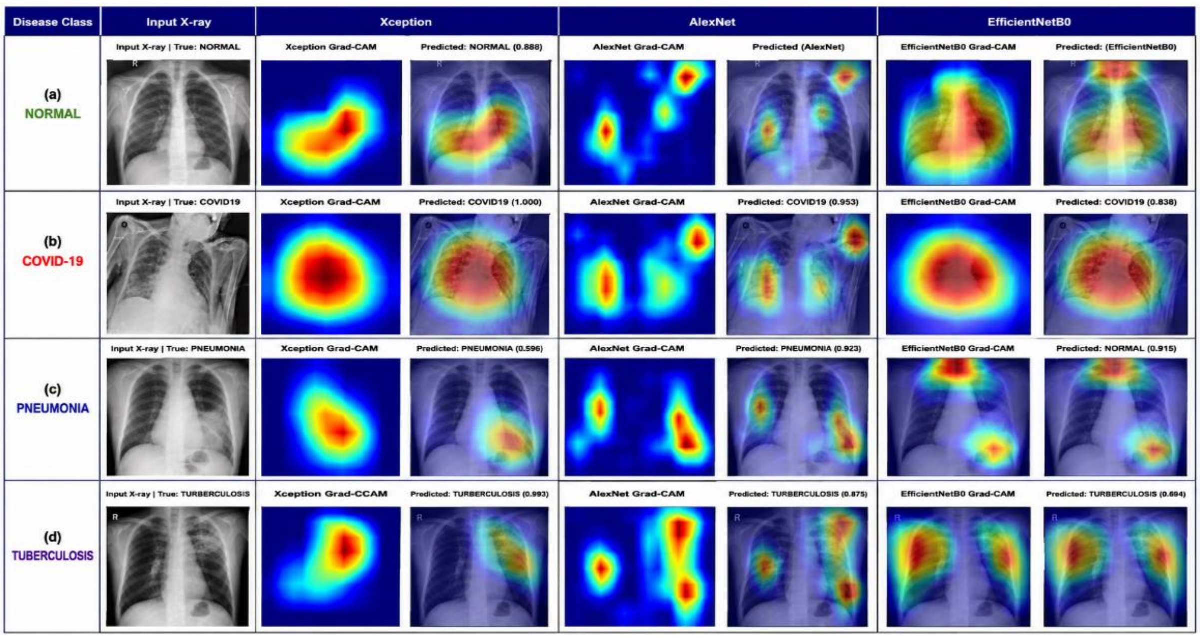

Figure 4: Grad-CAM visualizations for representative chest X-ray images from the four disease classes: (a) Normal, (b) COVID-19, (c) Pneumonia, and (d) Tuberculosis.

The Grad-CAM visualizations for a single chest X-ray image from each of the three disease classes are shown in Figure 4.

The key areas of focus for the CNN models during prediction are highlighted. The models primarily focused on the atypical areas of the lungs for each COVID-19 and Pneumonia case, where infection occurred. The top part of the lungs in the image of Tuberculosis was the focus, highlighting places where the lesions are visible. The Normal chest X-ray had the models allocated to areas of healthy lung tissue rather than any areas of abnormality. These visualizations indicate that the models had acquired useful features from these images and made their predictions on relevant parts of the lungs in a manner suitable for clinical use. As the proposed model is an ensemble of three CNN backbones (Xception, AlexNet, and EfficientNet-B0), Grad-CAM results from each backbone are shown before the ensemble prediction to illustrate the attention regions learned by each backbone.

The Average Ensemble performed well in classification across all four disease classes, as shown in Table 4. If we consider the results, the highest recall was achieved in the Tuberculosis class, with a 98% success rate, meaning that everything marked in this class was correctly identified. Additionally, it achieved good classification results, with a precision of 91.11%, an accuracy of 95.35%, and an F1 score of 95.35%. The model also achieved good accuracy for the COVID-19 and pneumonia classes, reflecting the meaningful features it learned across various chest diseases. Further, the macro-average F1-score of 92.63% indicates that the proposed system performed consistently across all four classes.

TABLE 4: Class-wise precision, recall, and F1-score of the proposed Average Ensemble model.

Model	Class	Precision	Recall	F1-Score
Average Ensemble	Normal	0.937824	0.773504	0.847775
Average Ensemble	COVID-19	0.990385	0.971698	0.980952
Average Ensemble	Pneumonia	0.881119	0.969231	0.923077
Average Ensemble	Tuberculosis	0.91111	1.00	0.953488
Average Ensemble	Macro Average	0.930110	0.928608	0.926323
Average Ensemble	Weighted Average	0.914946	0.911803	0.909797

To compare the proposed Average Ensemble model with the individual CNN models, a statistical significance test was conducted to determine whether the proposed model's performance improvement is statistically meaningful. The same test samples were used, and a significance level of 0.05 was employed in the analysis. The p-values here indicate whether these performance differences are statistically significant.

The results. The statistical significance analysis is presented in Table 5. The average Ensemble model performs significantly better than Gabor Features Only and EfficientNet-B0, with p-values < 0.05. Parts of these differences were not statistically significant between Average Ensemble and the models Xception and AlexNet. These results demonstrate that the proposed model achieves the highest overall accuracy; however, the performance gain over the strongest individual CNN models is also rather modest.

TABLE 5: Statistical significance analysis comparing the proposed Average Ensemble model with the baseline models.

Baseline Model	Proposed Model	P Value	Significant (P<0.05)
Gapour Feature Only	Average Ensemble	$2.0422e^{-29}$	True
Efficient Net-B0	Average Ensemble	$5.9735e^{-05}$	True
AlexNet	Average Ensemble	$5.9664e^{-01}$	False
Xception	Average Ensemble	$2.62437e^{-01}$	False

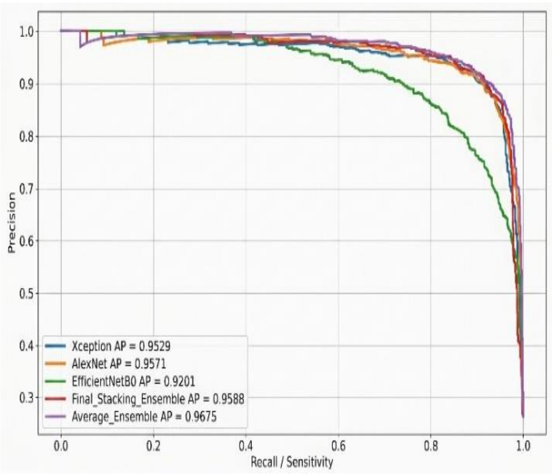

Figure 5: Precision–Recall curves of the evaluated models averaged over six-fold stratified cross-validation.

The precision and recall curve, as shown in Figure 5, also confirmed previous results. Here, Average Ensemble achieved the best AP of 0.9675. The Stacking Logistic Regression Ensemble was the second-highest, with an AP of 0.9588. Blended CNN models achieved the best AP value of 0.9571 for individual CNN models.

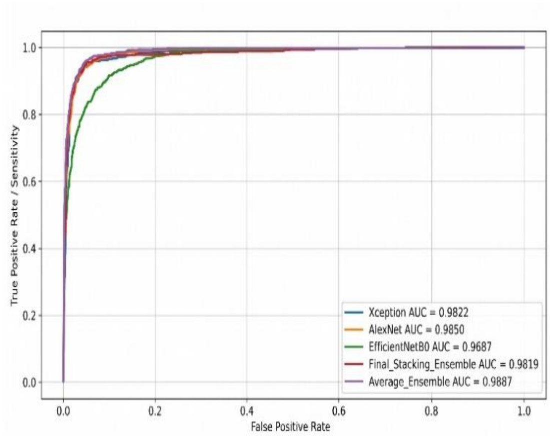

Figure 6: Receiver Operating Characteristic (ROC) curves of the evaluated models.

Micro-average ROC curves of all tested algorithms are given in Fig. 6. The Average Ensemble algorithm achieved an AUC value of 0.9887, while AlexNet achieved an AUC value of 0.9850. The Xception model achieved an AUC of 0.9822, while the Stacking Logistic Regression Ensemble algorithm and EfficientNet-B0 achieved AUCs of 0.9819 and 0.9687, respectively. Therefore, the proposed ensemble model can classify chest X-ray images with satisfactory results in four classes: Normal, COVID-19, Pneumonia, and Tuberculosis by means of the ROC and PR analysis.

Table 6 compares the proposed model with the other related studies presented by various authors. Most of the previously published research was based on binary tuberculosis classification, whereas the present study uses four classes, making the problem statement more complex. In any case, the proposed algorithm could have obtained 91.18% accuracy.

TABLE 6: Comparison of the Proposed Method With Existing Studies.

Ref.	Method	Accuracy (%)
Rajaraman et al. (2018) [1]	Stacked Generalization Ensemble	84
Khatibi et al. (2021) [2]	CNN + Complex Network + Stacked Ensemble	89.36
Ayaz et al. (2021) [4]	Hybrid Feature Descriptor + Ensemble Learning	90.42
Rajaraman & Antani (2020) [8]	Modality-Specific CNN Ensemble	90.13
Proposed Method	-	91.18

DISCUSSION

We compared the performance of both ensembles with that of the individual CNN models, and both ensembles outperformed the CNN models. Still, the Average Ensemble had slightly higher accuracy than the Stacking Logistic Regression Ensemble. This may be because Average Ensemble is a simple, balanced method that equally fuses the prediction probabilities of all base models. This will help minimize prediction variance and stabilize the final prediction output. Instead, the predictions of the Stacking Ensemble are combined using another meta-classifier. The overall performance of the meta-classifier might deteriorate slightly, as the data set is not very large and the classes are far from balanced. Consequently, the Average Ensemble achieved higher consistency in classification results in the four disease classes.

Based on the results, the framework classified chest X-ray images into the Normal, COVID-19, Pneumonia, and Tuberculosis groups. The individual CNN models were compared for accuracy, with AlexNet ranking first, followed by Xception, and EfficientNet-B0 ranking last. Future studies have also reported that learning features using transfer learning with CNN models can be achieved on chest X-ray images[5,7,8].

All three CNN models have distinct learning capacities, which makes them suitable for feature fusion. All these technologies can learn rich spatial information, while AlexNet delivers reliable low-level image features, and EfficientNet-B0 provides efficient feature extraction with fewer model parameters. The framework can extract complementary information that might not be extracted by a single CNN model by combining these models.

The ensemble models also showed further improvement in classification accuracy over the individual CNN models, consistent with the results presented[1,2]. In addition, using Gabor texture features provided a further increase in classification accuracy. The result is consistent with earlier work in which the use of Gabor texture features led to improved results[4,12].

The models' discriminatory power in differentiating Tuberculosis cases was validated using Precision–Recall and ROC curves, as well as the high AP and AUC for the model ensembles. Furthermore, some classification difficulties arose because the X-ray appearances of Normal and Pneumonia were very similar, as noted earlier in the references[7,9].

Compared to previous works that focus only on classifying Tuberculosis detection into two classes [1-6], this work considers a 4-class classification (Normal, COVID-19, Pneumonia, and Tuberculosis). The methods discussed above can be integrated to develop a robust approach to classifying chest diseases.

There are, however, some limitations to the proposed work. It is worth mentioning that the dataset used here is imbalanced with respect to the PNEUMA (Pneumonia) class. Furthermore, the experiments were conducted using a single dataset. An evaluation on multi-center datasets not included in the above work should be conducted to assess the efficiency of the developed method.

One constraint is that the public patient information is not stored with identifiers. However, it was not possible to confirm the independence of the training and validation data sets at the patient level. Further tests of the proposed framework are to be conducted using patient-level datasets to gain deeper insights into its generalization performance.

CONCLUSION

This study aimed to develop a hybrid ensemble learning method to classify chest X-rays into four categories: Normal, COVID-19, Pneumonia, and Tuberculosis. The proposed approach uses deep features from Xception, AlexNet, and EfficientNet, along with Gabor texture features and ensemble learning.

It showed that the ensemble model outperforms individual CNN models in accuracy, AP, and AUC, achieving 91.18% accuracy, 0.9675 AP, and 0.9887 AUC. Results showed that deep learning, texture features, and ensemble learning methods are useful in improving chest disease classification. Larger datasets and more sophisticated explainable AI techniques will be the subject of future studies.

Author Contributions

ARKT: Conceived and designed the study, supervised the research, and contributed to manuscript drafting.

MLURS: Assisted in study design, data collection, and data analysis.

FS: Contributed to the literature review, data interpretation, and manuscript preparation.

MHZ: Assisted in data collection, data entry, and manuscript editing.

SAK: Critically revised the manuscript, provided overall guidance, and approved the final version.

Grant Support and Funding Disclosure
None.

Conflict of Interests
None.

REFERENCES

1. Rajaraman S, Antani S, Candemir S, Xue Z, Alderson P, Kohli M, et al. A novel stacked generalization of models for improved TB detection in chest radiographs. 2018. doi: https://doi.org/10.1201/9780429029417
2. Khatibi T, Shahsavari A, Farahani A. Proposing a novel multi-instance learning model for tuberculosis recognition from chest X-ray images based on CNNs, complex networks, and stacked ensemble. Phys Eng

Sci Med. 2021 Mar 1;44(1):291–311. https://doi.org/10.1007/s13246-021-00980-w
3. Hwa SKT, Hijazi MHA, Bade A, Yaakob R, Jeffree MS. Ensemble deep learning for tuberculosis detection using chest X-ray and Canny edge detected images. IAES International Journal of Artificial Intelligence. 2019 Dec 1;8(4):429–35. doi: https://doi.org/10.11591/ijai.v8.i4.pp429-435
4. Ayaz M, Shaukat F, Raja G. Ensemble learning-based automatic detection of tuberculosis in chest X-ray images using hybrid feature descriptors. Phys Eng Sci Med. 2021 Mar 1;44(1):183–94. https://doi.org/10.1007/s13246-020-00966-0
5. Wong A, Lee JRH, Rahmat-Khah H, Sabri A, Alaref A, Liu H. TB-Net: A Tailored, Self-Attention Deep Convolutional Neural Network Design for Detection of Tuberculosis Cases From Chest X-Ray Images. Front Artif Intell. 2022 Apr 7;5. doi: https://doi.org/10.3389/frai.2022.827299
6. Munadi K, Muchtar K, Maulina N, Pradhan B. Image Enhancement for Tuberculosis Detection Using Deep Learning. IEEE Access. 2020;8:217897–907. doi: https://doi.org/10.1109/ACCESS.2020.3041867
7. Mirugwe A, Lillian Tamale M, Nyirenda J. Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures [Internet]. https://doi.org/10.2196/66029
8. Rajaraman S, Antani SK. Modality-Specific Deep Learning Model Ensembles Toward Improving TB Detection in Chest Radiographs. IEEE Access. 2020;8:27318–26. doi: https://doi.org/10.1109/access.2020.2971257
9. Rahman T, Khandakar A, Kadir MA, Islam KR, Islam KF, Mazhar R, et al. Reliable tuberculosis detection using chest X-ray with deep learning, segmentation, and visualization. IEEE Access. 2020;8:191586–601. doi: https://doi.org/10.48550/arXiv.2007.14895
10. Rajaraman S, Kim I, Antani SK. Detection and visualization of abnormality in chest radiographs using modality-specific convolutional neural network ensembles. PeerJ. 2020;2020(3). https://doi.org/10.7717/peerj.8693
11. Sahlol AT, Elaziz MA, Jamal AT, Damaševičius R, Hassan OF. A novel method for the detection of tuberculosis in chest radiographs using artificial ecosystem-based optimization of deep neural network features—symmetry (Basel). 2020 Jul 1;12(7). https://doi.org/10.3390/sym12071146
12. Win KY, Maneerat N, Hamamoto K, Sreng S. Hybrid learning of hand-crafted and deep-activated features using particle swarm optimization and optimized support vector machine for tuberculosis screening. Applied Sciences (Switzerland). 2020 Sep 1;10(17). https://doi.org/10.3390/app10175749
13. Hwa SKT, Bade A, Hijazi MHA, Jeffree MS. Tuberculosis detection using deep learning and contrast-enhanced canny edge-detected X-Ray images. IAES International Journal of Artificial Intelligence. 2020;9(4):713–20. http://doi.org/10.11591/ijai.v9.i4.pp713-720
14. Msonda P, Uymaz SA, Karaağaç SS. Spatial pyramid pooling in deep convolutional networks for automatic tuberculosis diagnosis. Traitement du signal. 2020 Dec 1;37(6):1075–84. https://doi.org/10.18280/ts.370620
15. Pasa F, Golkov V, Pfeiffer F, Cremers D, Pfeiffer D. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. Sci Rep. 2019 Dec 1;9(1). https://doi.org/10.1038/s41598-019-42557-4
16. Devnath L, Luo S, Summons P, Wang D. Tuberculosis (TB) classification in chest radiographs using deep convolutional neural networks. https://doi.org/10.3389/fmed.2024.1290729
17. Santosh K, Allu S, Rajaraman S, Antani S. Advances in Deep Learning for Tuberculosis Screening using Chest X-rays: The Last 5 Years Review. J Med Syst. 2022 Nov 1;46(11). https://doi.org/10.1007/s10916-022-01870-8
18. Nafisah SI, Muhammad G. An Explainable Graph Neural Network for Harnessing Long-Range Dependencies in Tuberculosis Classifications in Chest X-Ray Images. Neural Comput Appl. 2024 Jan 1;36(1):111–31. https://doi.org/10.3390/diagnostics15243236
19. Venkatachalam C, Shah P. A Dual-Branch Deep Learning Framework Combining Xception and ResNet for Accurate Lung and Colon Cancer Detection. IEEE Access. 2025;13:131483–97. doi: https://ieeexplore.ieee.org/document/11081468/
20. Singh Bisht A, Ajay A, Karthik R. DeepCRC-Net: An Attention-Driven Deep Learning Network for Colorectal Cancer Classification Using Xception and Efficient Lightweight Local Feature Fusion Networks. IEEE Access. 2025;13:49362–74. https://doi:10.1109/ACCESS.2025.3550004
21. Liu L, Xia K. BTIS-Net: Efficient 3D U-Net for Brain Tumor Image Segmentation. IEEE Access. 2024;12:133392–405. doi: https://doi.org/10.1109/ACCESS.2024.3460797
22. Jain M, Shah A. Anomaly Detection Using Convolutional Neural Networks (CNN). International Journal of Advancements in Computational Technology. 2024;2:12–22. doi: https://doi.org/10.56472/25838628/IJACT-V2I3P102
23. Brima Y, Atemkeng M, Djiokap ST, Ebiele J, Tchakounté F. Transfer learning for the detection and diagnosis of types of pneumonia, including pneumonia induced by COVID-19, from chest X-ray images. doi. 2021;11(8):1480. https://doi.org/10.3390/diagnostics11081480
24. Ahmed MS, Rahman A, AlGhamdi F, AlDakheel S, Hakami H, AlJumah A, et al. Joint diagnosis of pneumonia, COVID-19, and tuberculosis from chest X-ray images: A deep learning approach. doi. 2023;13(15):2562. https://doi.org/10.3390/diagnostics13152562
25. Mwendo I, Gikunda P, Maina A. Deep transfer learning for detecting COVID-19, pneumonia, and tuberculosis using CXR images: A review. arXiv. 2023. Doi : https://arxiv.org/abs/2303.16754
26. Mabrouk A, Díaz Redondo RP, Dahou A, Elaziz MA, Kayed M. Pneumonia detection on chest X-ray images using an ensemble of deep convolutional neural networks. arXiv. 2023. Doi. https://arxiv.org/abs/2312.07965
27. Shahzad S, Sajeela, Shah N, Khan SA. Ensemble deep learning for multi-class chest X-ray classification: Robust detection of pneumonia and tuberculosis. Med Res Arch. 2025;13(11). Doi: https://doi.org/10.18103/mra.v13i11.7065