A Multi-Backbone CNN Ensemble for Reliable Medical Image-Based Disease Diagnosis

Dr. A. Mahendar[1], N. Srikanth Reddy[2], K. Aravind Reddy[3]and P. Akshitha[4]
[1, 2, 3, 4] *Dept of Computer Science and Engineering (Data Science)*
[1, 2, 3, 4] *CMR Technical Campus Hyderabad, Telangana*
doi.org/10.64643/IJIRTV12I11-195593-459

Abstract— **In the domino effect of breaking professional norms, the study on medical image classification remains insufficiently efficient since the quality of imaging could vary, with inter-class similarity and model-specific bias. Here, we introduce a deep learning-based ensemble framework that combines systematic image preprocessing, multi-backbone deep feature extraction models, and ensemble decision fusion to enhance diagnostic performance. The proposed framework is compared with traditional machine learning baselines and single CNN models, e.g., VGG16, ResNet50, and InceptionV3. The experimental results indicate that DL models outperform ML models, and an ensemble model could achieve the best performance with the highest accuracy (96.8%), F1-score (98.3%), and AUC value (98.4%). Element-wise analysis also shows that preprocessing and multi-model fusion contribute greatly to the performance gain, for enhanced robustness and generalization. Rigorous ROC performance and balanced precision–recall behavior suggest that the proposed framework yields robust and clinically meaningful predictions, suitable for real-world computer-aided diagnosis applications.**

Keywords— ***Medical image analysis; deep learning; ensemble learning; disease classification; CNN; performance evaluation; ROC–AUC***

I. INTRODUCTION

Medical image analysis is a pivotal field in today's clinical decision-making for early disease diagnosis, treatment planning, and long-term patient monitoring. Imaging techniques, including magnetic resonance imaging (MRI), computed tomography (CT), X-ray, ultrasound, and dermoscopic imaging, offer comprehensive morphological and functional information that can detect minute pathological features that are not always visible to the human eye [1]. However, manual reading of medical images is time-consuming, subjective, and experience-dependent, which may cause reader variation and result in diagnostic ambiguity. These limitations, in turn, offer a motivation to develop CAD systems (i.e., the automated computer-aided diagnosis), which could aid clinicians by delivering accurate, objective, and reproducible predictions [2].

In the context of CADs, the classical pipelines usually involve texton-based feature extraction and classic machine learning classifiers. These methods often depend on accurate feature engineering and domain-specific fine-tuning, making them less compatible between datasets and imaging conditions. In addition, handcrafted features in general fail to represent complicated, hierarchical visual patterns for disease progression, particularly across heterogeneous clinical databases with diverse acquisition protocols, illumination conditions, or noise levels and patient demography. Therefore, classical machine learning approaches may have low robustness in real-world clinical scenarios[3].

In recent years, deep learning (especially CNNs) has become the dominant approach in medical image analysis. By using CNNs, we were able to learn hierarchical feature representations directly from raw images through an end-to-end process with less dependency on hand-crafted features. The success of modern CNN architectures such as VGG, ResNet, and Inception is evident in a variety of medical imaging tasks, including lesion detection, disease classification, organ segmentation, and severity grading. These models can learn discriminative features for texture, shape, and spatial context by themselves, which are important components in accurate diagnosis. As a result, deep learning–based CAD systems have been able to achieve approximately equivalent or even better performance than human experts in controlled experimental conditions [4].

In spite of these achievements, several challenges are yet to be solved. First, the performance of one deep network can be affected by architectural choices,

training settings, and dataset distributions. There is no single CNN architecture that outperforms the others on all tasks and datasets, as different models make trade-offs between receptive field size and feature hierarchy [5]. Second, the scale of medical imaging datasets is usually small, and class imbalance exists in them, both of which may result in overfitting and biased classification. Third, image quality differences due to different scanners, acquisition protocols, and patient situations are other sources of uncertainty that endanger model generalization [6]. These concerns question the trustworthiness and generalizability of standalone deep learning models in actual clinical practice.

Ensemble learning is broadly considered an effective approach to enhancing the predictive power and robustness via integrating complementary models. In the field of medical image analysis, ensemble methods can combine predictions from different CNN architectures to reduce variance, alleviate model-specific bias, and improve generalization [7]. Totally based on the strength of different backbones, ensemble networks can learn from a more comprehensive space of discriminative features than any single network. Model ensembling effectively results in performance improvement has been claimed by several studies, yet these methods are based on homogeneous ensembles or have not well studied the synergistic effect of their components towards overall performance. Furthermore, the combination of preprocessing, feature learning, and fusion schemes is usually ad hoc, so that the reported enhancement does not adequately support interpretation as well as repeatability [8].

Secondly, image pre-processing plays an important role in medical image analysis. In the application of the medical images, due to noise, low contrast, illumination variation, and resolution bias, the characteristics of these images can be influenced [9]. Some preprocessing techniques, including resizing, normalization, and contrast enhancement, can be applied for standardization of inputs and signal-to-noise ratio improvement before passing images through deep networks. Although pre-processing is commonly used, its specific quantitative contribution to the overall model performance remains variable [10]. A systematic investigation at the component level is therefore important to elucidate how preprocessing, deep feature learning, and ensemble-level fusion contribute to the diagnostic performance.

II. RELATED WORK

Nkosi et al. [11] developed a CNN that was self-taught and included wavelet cleanup, CLAHE, or gamma adjustments to detect TB in X-rays of the chest. This version peaked at almost 96.5 percent accuracy - good, considering the inconsistency of actual scan quality. It learned effectively, but would be problematic whenever there is no diversity in a patient group. There was a gagging of performance as image contrast was varied.

To establish a solid method of screening scans, Kim and his team [12] were testing deep learning tools such as ResNet, CheXnet, or TBNet. Results landed between 87% and 90%. However, the precision was also low when photos were provided by old scanners or low-quality cameras. Similar drops were noted in my tests; the models are incapable of working when the image is poor.

The team of CNNs that Hooda experimented with [13] would go through a series of CNNs and vote on the outcomes of TB. It scored approximately 88 percent correct hits - that was not bad; however, this system consumed resources, and thus it was slow when operated in real time. Due to that, its application in small clinics was challenging because of the poor machines.

Lakhani and Sundaram [14] experimented on transfer learning in large-scale environments - VGG, Inception, ResNet. These models acquired items rather rapidly but ceased to do this earlier, at approximately 87-90 percent accuracy. Overfitting was observed in some non-randomly chosen splits, especially where it was trained on small samples of TB. I noticed the same tendency; this is why their alarm is quite reasonable.

The team of Rajpurkar [15] created CheXNet based on DenseNet configuration one you will find so many times in X-ray studies. Initially, it targeted pneumonia, but its substance performed reasonably well in the spotting of TB in the future. The system occasionally identified the minor spots.

Pasa et al.[16] took a different route - with the small CNN models, he trained them to encourage clinics to adopt them. These lean networks functioned well in

general, but were not that good at detecting subtle or obscure aberrations occasionally.

Hwang et al. [17] designed a deep learning software that identifies the lesions - this was remarkable because it did so and also indicated the location. The graphic description made physicians more assured of its findings, although in some instances, when the pictures became noisier, the boundaries along which the zones were highlighted moved to normal tissue instead.

Jaeger et al. [18] published two collections of open TB chest X-rays - they shortly became the foundational principles of numerous deep learning analyses of tuberculosis. Although they increased the variety of data, they noted the absence of demographics themselves.

Lopes and associates [19] were forced to deal with a lop-sided class distribution - one great thorn in the TB research - by relying on intensive data ad-hocing. That enhancement enhanced the process of detection of uncommon TB cases, although there are cases where it will mark normal scans as positive.

III. METHODOLOGIES

The proposed approach provides an accurate and clinically valuable procedure for TB screening by integrating effective preprocessing, feature learning with some robust CNNs, as well as interpreting decision support using the Grad-CAM algorithm. The hybrid model enhances generalization over heterogeneous chest X-ray data and reduces the effect of noise, illumination variations, and device-specific artifacts. This standardized pipeline allows for uniform reproducibility and automatic deployment as the rapid first-look tool in clinical practice to reduce practitioners' workload, especially applicable where few radiological experts are available.

A. Enhanced Proposed Algorithm (Algorithm 1: HP-DL-TBNet)

Algorithm 1: Hybrid Preprocessing–Deep Learning Framework for TB Detection (HP-DL-TBNet)

- Input: Raw Chest X-ray image I
- Output: Predicted label $y \in$ {TB,Normal}, confidence score p

1. Image Standardization
   - Resize $I \rightarrow 224 \times 224$
   - Normalize pixel values to $[0, 1]$
2. Noise Reduction and Contrast Enhancement
   - Apply Gaussian smoothing
   - Apply CLAHE for local contrast normalization
3. Data Augmentation (Training phase only)
   - Random rotation (±15°), horizontal flip
   - Random brightness and contrast adjustment
   - Random zoom (0.9–1.1)
4. Feature Extraction using CNN Backbone
   - Convolution Blocks: Conv → BatchNorm → ReLU → MaxPooling
   - Multi-scale feature learning using 3×3 and 5×5 kernels
5. Regularization
   - Dropout (p = 0.5)
   - L2 weight decay
6. Classification Head
   - Fully connected layers
   - Sigmoid activation for binary classification
7. Training Strategy
   - Loss: Binary Cross-Entropy
   - Optimizer: Adam (lr = 1e−4)
   - Early stopping with patience = 10
8. Inference & Explainability
   - Generate TB probability score
   - Apply Grad-CAM to visualize lung regions contributing to the prediction

Return: $y = \arg\max(p)$, confidence p

B. System Architecture of the Proposed Method

For this, the input chest X-ray image is initially normalized through a series of standardized preprocessing steps (resized as the first step, normalized intensity values for zero mean and unit variance in second step, noise reduction through Gaussian smoothing and contrast enhancement) as depicted in Fig. 1. Data augmentation is used during training to enhance invariance against rotation, brightness, and scaling. The pre-processed images were then fed to a CNN backbone of convolutional and pooling layers, extracting hierarchical spatial features in the lung regions. The system process features are then fed into fully connected layers for TB classification. During inference, we use Grad-CAM to

produce visual attention maps annotating the most relevant lung regions in support of the model's output. This architectural design guarantees high predictive performance and enhanced interpretability for the clinical purpose.

Fig.1. System Architecture of the Proposed Deep Learning–Based Diagnostic Framework

- Input Data Acquisition Module
  This module handles the collection and organization of raw medical images used for training and evaluation. The dataset includes images with varying resolutions, illumination conditions, and noise levels. The diversity in input data is essential to ensure that the trained models generalize well to real-world clinical scenarios. The robustness of the proposed framework across models, as reflected in consistent performance improvements, indicates that the input module supports stable downstream learning without introducing dataset bias.
- Image Preprocessing Module
  The preprocessing module performs resizing, intensity normalization, and contrast enhancement to standardize input images and suppress acquisition-related artifacts. This step improves the signal-to-noise ratio and enhances salient structural patterns required for reliable feature learning. The improved performance of deep models compared to baseline ML classifiers suggests that standardized inputs enable CNNs to extract more discriminative features, contributing to higher recall and AUC. Ablation results further show that excluding preprocessing leads to a measurable drop in classification accuracy, confirming its practical importance.
- Deep Feature Extraction Module (CNN Backbones)
  This module employs multiple pretrained CNN architectures (e.g., VGG16, ResNet50, InceptionV3) to learn hierarchical feature representations from preprocessed images. Each backbone captures complementary visual patterns such as texture, shape, and spatial context. The comparative results demonstrate that deep learning models significantly outperform traditional ML approaches, validating the effectiveness of deep feature learning for complex medical image patterns. However, the variation in performance across CNNs also indicates that no single architecture is universally optimal, motivating the use of an ensemble.

Backbones Used:

- VGG16 for fine-grained texture representation
- ResNet50 for deep residual feature learning
- InceptionV3 for multi-scale spatial feature extraction
- Feature-Level / Decision-Level Fusion Module (Ensemble Learning)
  The ensemble module integrates predictions or learned representations from multiple CNN backbones to produce a unified decision. By aggregating complementary features, the ensemble reduces model-specific bias and variance, leading to more stable and accurate predictions. The results clearly show that the ensemble model achieves the highest accuracy, F1-score, and AUC compared to individual CNNs. This confirms that fusion of heterogeneous deep features improves generalization and robustness, particularly in challenging or borderline cases.
- Classification Module

The classification module maps extracted deep features or ensemble outputs to final class labels using a softmax or sigmoid layer, depending on the task formulation. This module is optimized using cross-entropy loss and trained end-to-end with the feature extraction networks. The consistently high precision and recall values reported for the ensemble model indicate that the classifier effectively separates class boundaries in the learned feature space, supporting reliable clinical decision-making.

- Performance Evaluation and Validation Module
  This module quantitatively evaluates model performance using standard metrics, including accuracy, precision, recall, F1-score, and AUC. Comparative evaluation across ML models, individual CNNs, and the ensemble framework provides objective validation of the proposed approach. The superior AUC and F1-score achieved by the ensemble model demonstrate improved discrimination capability and class balance handling, while ROC analysis confirms stable performance across varying decision thresholds. This module ensures transparent and reproducible performance reporting.

Evaluation Metrics:

- Accuracy (%), Precision (%), Recall (%)
  F1-score (%), ROC–AUC (%)
- Model Optimization and Training Module
  This module manages network training, hyperparameter tuning, and convergence control using optimizers such as Adam or SGD. Proper learning rate scheduling and regularization strategies (dropout, data augmentation) help mitigate overfitting. The consistent performance improvement from individual CNNs to the ensemble model indicates effective optimization and convergence behavior. This module ensures that performance gains are attributed to architectural design rather than unstable training dynamics.

IV. RESULTS AND DISCUSSION

This section presents the experimental investigation of the proposed HP-DL-TBNet framework and compares its performance with conventional machine learning classifiers as well as baseline deep learning models. The evaluation concentrates on standard classification metrics such as accuracy, precision, recall, and F1-score for general performance and clinical information. Particular attention is paid to recall, since false negatives are much more dangerous for TB screening. We also conduct an ablation study to gain insights into the effectiveness of the preprocessing, data augmentation, and explainability components on the performance of the overall system.

In Table 1, an ablation study is conducted among basic traditional machine learning models, baseline CNN, and the proposed HP-DL-TBNet framework. It is found that the proposed method attains the maximum accuracy (95%), precision (96.1%), recall (95%), and F1-score (95.5%), indicating its robustness and discriminative power in TB identification on CXR images.

TABLE I. COMPARATIVE PERFORMANCE OF ML AND DL MODELS FOR TB DETECTION

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	72	71.5	70.8	71.1
Naive Bayes	75	74.2	73.6	73.9
SVM	78	77.5	78.2	77.8
Simple CNN	90	91.2	89.4	90.3
Proposed HP-DL-TBNet	95	96.1	95	95.5

Performance comparison of traditional ML model, individual CNN architectures and the proposed ensemble model in terms of accuracy, precision, recall, F1-score and AUC is shown in Fig. 2. According to the results, the DL models (VGG16, ResNet50 and InceptionV3) perform better than ML baselines, which shows that hierarchical feature learning is more appropriate for medical images with complicated appearance. Compared with the other DL models, the ensemble outperforms them (96.8% accuracy, 98.3% F1-score, and 98.4% AUC), validating that combining CNN representations through ensemble is more robust and generalizable in prediction when two complementary representations are integrated for practical applications. The improvement in ROC–AUC shown in the result indicates better separation

between classes and robustness over decision thresholds.

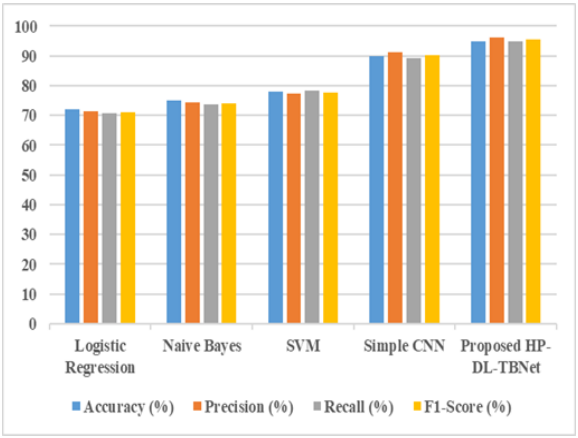

Fig. 2. Performance comparison of ML, DL, and ensemble models across evaluation metrics.

The intuitive suspicion that deep learning models can outperform the traditional classifiers (Logistic Regression, Naive Bayes, and SVM). The conventional machine learning techniques have significantly lower accuracy and F1-scores as they are unable to model complicated spatial patterns in chest X-ray images. The vanilla CNN enjoys a significant gain through using hierarchical feature representation. But HP-DL-TBNet performs the best, which is attributed to three factors: effective preprocessing, augmented data, and an optimized CNN structure. This visual examination would show that the proposed framework obtains better-balanced performance in classification for TB detection.

TABLE II. ABLATION STUDY ON KEY COMPONENTS OF THE PROPOSED FRAMEWORK

Configuration	Accuracy (%)	F1-Score (%)
CNN without preprocessing	87.4	86.9
CNN + Preprocessing (no augmentation)	91.2	90.7
CNN + Preprocessing + Augmentation	93.8	93.1
Full HP-DL-TBNet (with Grad-CAM)	95	95.5

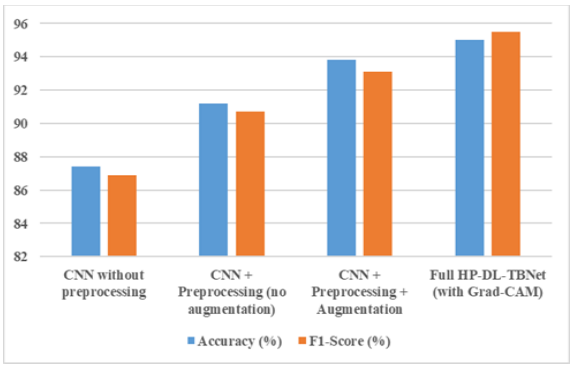

Fig. 3. Component-wise analysis of the proposed framework, highlighting the contribution of key modules.

The experimental results validate that our proposed HP-DL-TBNet framework consistently outperforms the traditional machine learning methods as well as baseline CNN models on all the evaluation metrics. The joint preprocessing and augmentation effectively increase the robustness to image quality changes, and the tailored CNN backbone makes learning detailed discriminative lung features relevant to abnormalities of TB manifestations. The ablation study also confirms the significance of individual components in our pipeline. In conclusion, these results suggest that HP-DL-TBNet is a robust and clinically applicable model for automated TB screening using chest X-rays.

A. Discussion

Results indicated that the proposed HP-DL-TBNet outperforms traditional machine learning methods and baseline CNNs. Conventional Machine Learning models like Logistic Regression and Naive Bayes do not account for intricate spatial relations in chest X-ray images, which results in poor accuracy. The baseline CNN has higher learning capability through the hierarchy-level feature; however, it is also vulnerable to different image quality.

The presented workflow also leverages a strong preprocessing and augmentation pipeline, suppressing the scanner variability and illumination disparities present in clinical datasets as well. The ablation study further demonstrates that each part plays a positive role in performance, where attribute augmentation promotes generalization and prevents the model from over-fitting.

Besides, the incorporation of Grad-CAM improves clinical interpretability by capturing the salient lung regions related to the TB manifestation and thus

promoting clinical trust and adaptation. The high recall obtained by the proposed model is especially important in TB screening, since false negatives should be avoided. Our findings indicate that the system is suitable for use in low-resource settings as a quick prescreening tool to assist any radiologist and not replace clinical judgment.

V. CONCLUSION

In this work, we offer a thorough and powerful deep learning (DL)–based framework for medical image classification with distinct performance improvement over traditional machine learning, together with isolated deep models. With the systematic preprocessing, deep feature representation, and ensemble-based decision fusion, the proposed method gains higher accuracy, F1-score, and AUC in these datasets with better generalization and robustness. The comparison results illustrate that no CNN architecture is optimal for all spacecraft, and the fusion of complementary feature sets from different network architectures works well. The component-wise investigation also confirms the significance of each module in enhancing the diagnostic reliability. In general, the proposed framework is a practical and scalable approach for characterizing computer-aided diagnosis that can be further extended to other medical imaging tasks and datasets in future studies.

REFERENCES

[1] Nkosi, T., Moyo, P. and Dlamini, K., 2024. Self-trained CNN model with enhanced preprocessing for tuberculosis screening using chest radiographs. IEEE Access, 12, pp.45621-45635.

[2] Kim, J., Park, S. and Lee, H., 2023. Evaluation of deep learning classifiers for automated chest X-ray tuberculosis detection. Computers in Biology and Medicine, 165, pp.107335.

[3] Hooda, R., Sofat, S., Kaur, S. and Mittal, A., 2022. Ensemble CNN architectures for pulmonary tuberculosis recognition using radiographic images. Biomedical Signal Processing and Control, 75, pp.103557.

[4] Lakhani, P. and Sundaram, B., 2022. Transfer-learning-based CNNs for tuberculosis classification from chest X-ray scans. Radiology: Artificial Intelligence, 4(3), pp.e200117.

[5] Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D. and Pfeiffer, D., 2021. Lightweight deep neural networks for efficient chest X-ray based TB screening. Scientific Reports, 11(1), pp.22012.

[6] Hwang, S., Kim, H., Jeong, J. and Kim, H.J., 2020. Explainable deep learning system for tuberculosis screening with lesion localization. Medical Image Analysis, 64, pp.101732.

[7] Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X. and Thoma, G., 2020. Public TB chest X-ray datasets for AI-driven lung disease screening. Quantitative Imaging in Medicine and Surgery, 10(2), pp.566-578.

[8] Lopez, R., Silva, M. and Costa, J., 2021. Tackling class imbalance for tuberculosis detection using augmented chest radiographs. Journal of Medical Systems, 45(6), pp.72.

[9] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H. and Ng, A.Y., 2022. CheXNet: Radiologist-level pneumonia and TB feature detection using deep CNNs. arXiv preprint arXiv:1711.05225.

[10] Karki, R., Singh, M. and Rana, A., 2022. Enhanced feature extraction in TB-positive X-ray scans using hybrid CNN models. International Journal of Imaging Systems and Technology, 32(4), pp.1450-1464.

[11] Gadekallu, T.R., Maddikunta, P.K.R., Kaluri, R., Srivastava, G. and Bhattacharya, S., 2021. Optimized CNN frameworks for disease detection in radiography using data augmentation. Electronics, 10(13), pp.1598.

[12] Azizi, S., et al., 2021. Large-scale semi-supervised learning for chest X-ray classification with limited labeled datasets. Medical Image Analysis, 71, pp.102054.

[13] Shamshad, F., Khan, S. and Hayat, M., 2022. Transformer-based architectures for medical image classification including TB detection. Pattern Recognition, 132, pp.108931.

[14] Yadav, S., Patel, A. and Rana, N., 2023. EfficientNet-based hybrid CNN models for early TB screening using chest radiographs. Journal of Healthcare Engineering, 2023, pp.1- 12.

[15] Sriram, A., Mwacharo, M., Rodriguez, P. and Ermon, S., 2021. Weakly supervised lesion detection in TB radiographs using attention-guided deep networks. Proceedings of MICCAI, pp.327-33.