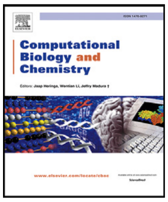
Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac



Vit-Ensemble: Probabilistic voting based ensemble of Vision Transformers for tuberculosis detection using radiographs

Nitesh Pradhan [a,*], Gaurav Srivastava [b], Geetika Kaushik [b]

[a] *Department of Computer Science and Engineering, The LNM Institute of Information Technology, Jaipur, 302031, Rajasthan, India*
[b] *Department of Computer Science and Engineering, Manipal University Jaipur, 303007, Rajasthan, India*

ARTICLE INFO

Keywords:
Tuberculosis detection
Chest X-ray
Medical imaging
Vision Transformers
Ensemble learning
Probabilistic voting
Vit-Ensemble

ABSTRACT

Tuberculosis (TB) detection from chest X-ray (CXR) images remains a critical challenge in global healthcare. This study introduces Vit-Ensemble, a novel ensemble model leveraging Vision Transformer (ViT) architectures for robust TB detection. The core innovation of Vit-Ensemble lies in its probabilistic voting strategy. Instead of relying solely on predicted class labels, it combines the probabilistic outputs of multiple ViT models trained on diverse TB datasets. By averaging these class probabilities, Vit-Ensemble makes decisions based on the collective confidence of individual models, enhancing generalization performance and reducing the impact of biases and uncertainties. To further improve diagnostic results, the authors systematically explore various image preprocessing techniques, including contrast enhancement and noise reduction. Through rigorous experimentation on benchmark datasets, Vit-Ensemble demonstrates superior performance compared to state-of-the-art convolutional neural network models. It achieves a remarkable 99.67% accuracy, outperforming both its individual components (e.g., DeiT-Base at 99.14%) and established convolutional neural networks (i.e., EfficientNet-B3 (99.64%) and DenseNet201 (93.21%)). Our results highlight the effectiveness of probabilistic voting within the ensemble framework for TB detection, offering the potential for early diagnosis and effective disease management. This research provides valuable insights into the integration of ViT architectures, probabilistic voting-based ensemble learning for medical image analysis. Vit-Ensemble represents a significant advancement in computer-aided TB diagnosis, promising improved public health outcomes and streamlined TB control efforts.

1. Introduction

Tuberculosis (TB) remains a persistent global health challenge, stemming from the ancient lineage of the Mycobacterium genus, with Mycobacterium tuberculosis (M.tb) as its causative agent (Rahman et al., 2020b; Godreuil et al., 2007). Its historical prevalence underscores its enduring impact on human populations and the complexities in diagnosis and management throughout history (Zimmerman, 1979). The deep-rooted association between M.tb and human hosts, dating back millions of years, highlights the intricate interplay between the pathogen and environmental conditions, shaping its global distribution (Cave and Demonstrator, 1939; Brown, 1941; Jaeger et al., 2014). Evidence from antiquity, such as skeletal deformities in Egyptian mummies, provides tangible reminders of TB's historical burden and societal implications (Narayanan et al., 2022).

Despite advancements, diagnosing TB historically relied on symptomatology, which proved insufficient due to overlaps with other chronic ailments prevalent in ancient societies (Chouhan et al., 2020a). This diagnostic challenge persists today, compounded by the global burden of TB, with approximately 25% of the population affected and significant mortality rates annually (Lin et al., 2014). Regional disparities underscore the urgency for accurate diagnosis, particularly in resource-constrained settings where conventional methods may fall short. As such, there is a critical need for expeditious and precise diagnostic approaches to advance disease management and global TB control efforts (Ford et al., 2016).

In response to these challenges, researchers have endeavored to develop computer-aided diagnosis (CAD) systems for TB detection, leveraging chest X-ray (CXR) images for their cost-effectiveness and utility in pulmonary condition detection (Chouhan et al., 2020b; Rahman et al., 2020a; Stephen et al., 2019; Nafisah and Muhammad, 2024; Iqbal et al., 2023; Fati et al., 2022). The growing integration of artificial intelligence (AI) in healthcare has demonstrated significant potential across various medical domains, from disease diagnosis to clinical decision support (İlikhan et al., 2025; Degirmenci et al., 2025; Wani et al., 2024a,c). Recent advances in explainable AI and interpretable deep

[*] Corresponding author.
E-mail address: nitesh.pradhan@lnmiit.ac.in (N. Pradhan).

https://doi.org/10.1016/j.compbiolchem.2025.108787
Received 30 July 2025; Received in revised form 7 November 2025; Accepted 15 November 2025
Available online 20 November 2025

learning have particularly shown promise in medical imaging tasks, including lung cancer detection and breast cancer classification (Wani et al., 2024a,b; Singh et al., 2025). However, the adoption of AI-driven diagnostic tools also presents social and juristic challenges that must be carefully considered to ensure responsible deployment (Perc et al., 2019). This research contributes to this ongoing effort by proposing novel methodologies for TB diagnosis, integrating advanced machine learning algorithms to achieve state-of-the-art results.

Despite these advances, existing TB detection methods face three critical limitations: (1) most ensemble approaches rely on hard voting strategies that discard valuable probability information, (2) CNN-based models struggle to capture long-range spatial dependencies crucial for detecting subtle TB manifestations, and (3) limited systematic investigation exists on how image preprocessing affects Vision Transformer performance for TB detection. To address these gaps, in this study the authors have developed “Vit-Ensemble”, a Vision Transformer-based ensemble model tailored for TB detection using radiographs. The key innovation of Vit-Ensemble lies in its probabilistic voting mechanism that leverages the collective confidence of multiple ViT models rather than simple majority voting, enabling more nuanced decision-making based on model uncertainty. Our proposed methodology, Vit-Ensemble, is designed to address the challenges inherent in TB detection from CXR images by leveraging the strengths of Vision Transformer (ViT) architectures and ensemble learning techniques. ViTs have demonstrated remarkable performance in various computer vision tasks, particularly in capturing long-range spatial dependencies within images, which is crucial for identifying subtle patterns indicative of TB infection in CXRs.

Furthermore, to enhance the robustness and effectiveness of the Vit-Ensemble, the authors explored various image preprocessing techniques. These techniques include contrast enhancement, noise reduction, and histogram equalization, among others, which are applied to the CXR images before feeding them into the ViT models. The rationale behind image preprocessing is to enhance the visibility of pathological features associated with TB, thereby facilitating more accurate detection. The authors have also investigated the efficacy of this ensemble approach in comparison to individual convolutional neural network (CNN) models, both with and without image preprocessing. Moreover, the authors contribute insights into the importance of image preprocessing techniques in improving diagnostic accuracy.

The primary contributions of the study are:

1. We introduce Vit-Ensemble, a novel probabilistic voting-based ensemble framework that combines multiple Vision Transformer models (DeiT-Base, Swin Transformer, and BEiT-Base) for TB detection. Unlike conventional hard voting methods, our approach aggregates probability distributions to achieve 99.67% accuracy, surpassing individual ViT models (99.14%) and SOTA CNN architectures including EfficientNet-B3 (99.64%).
2. We conduct a systematic investigation of six image preprocessing techniques (denoising, gamma correction, CLAHE, histogram equalization, wavelet transformation, and their combinations) on Vision Transformer performance for TB detection. Our analysis reveals that CLAHE preprocessing achieves the highest test accuracy (99.92%) with EfficientNetB0, providing practical guidance for preprocessing selection in medical imaging applications.
3. We perform comprehensive benchmarking across 20 CNN architectures and 7 Vision Transformer models on the cleaned TB chest X-ray dataset, demonstrating that Vision Transformers consistently outperform CNNs after data cleaning (96.53% vs. 92.40% for best models), while also reducing training time by approximately 40%.

By leveraging these novel methodologies and advanced machine learning techniques, our research aims to significantly advance the field of TB diagnosis. The proposed methodologies hold promise for early detection, effective management, and global TB control, ultimately improving public health outcomes and reducing the burden of this infectious disease.

This paper is structured as follows: Section 2 provides a review of recent works in the field. In Section 3, the authors present the details of the proposed methodology, including the Vit-Ensemble approach and the significance of image preprocessing. Section 4 offers experimental findings and results. Section 5 discusses the implications of the proposed method. Finally, Section 6 concludes the paper, highlighting its contributions and avenues for future research.

2. Related work

The diagnosis of tuberculosis (TB) remains a significant public health challenge. Traditionally, chest X-ray (CXR) interpretation by radiologists forms the frontline of TB detection. However, limitations like manual assessment variability and the need for specialized expertise highlight the potential for technology to assist in TB diagnosis. The integration of machine learning (ML) and, more specifically, deep learning (DL) algorithms into the analysis of CXR images presents a promising avenue to address these challenges and potentially improve the accuracy and efficiency of TB diagnosis. This section surveys recent advancements in DL-based and ensemble-based approaches for automated TB detection from CXR images.

2.1. Deep learning based approaches

Recent years have witnessed a surge in studies aimed at tackling tuberculosis (TB) detection using chest X-ray (CXR) images, leveraging machine learning (ML) and deep learning (DL) algorithms. Nafisah and Muhammad (2024) proposed a DL-based approach employing a convolutional neural network (CNN) architecture and explainable artificial intelligence (XAI) to achieve TB detection accuracy of 99.1%. Similarly, Iqbal et al. (2023) discussed a hybrid segmentation and classification approach utilizing CNNs, reporting accuracies ranging from 95.10% to 98.98% across different datasets.

Rahman et al. (2020a) introduced a DL-based approach incorporating segmentation and visualization techniques, achieving an accuracy of 95.5% in TB detection from CXR images. Acharya et al. (2022) proposed an AI-assisted DL-based approach employing a normalization-free network model, reporting accuracies of 96.91% in multiclass and 96% in binary classification tasks.

The landscape of TB detection has evolved significantly with the application of DL methodologies, resulting in substantial improvements in diagnostic accuracy. Recent investigations demonstrate the effectiveness of DL models in discerning intricate patterns and features from medical imagery, leading to enhanced discriminative acumen between TB-positive cases and non-TB counterparts. Moreover, DL-based paradigms exhibit reduced incidence of false positive outcomes, contributing to improved clinical decision-making processes.

2.2. Ensemble based approaches

Ensemble-based methodologies have emerged as robust tools for TB detection using CXR imagery. Dey et al. (2022) employed a type-1 Sugeno fuzzy integral-based ensemble technique, achieving a classification accuracy of 98.8% by amalgamating decisions from diverse CNNs. Fati et al. (2022) adopted a comprehensive deep learning strategy with hybrid attributes, yielding accuracies of 99.2% across distinct datasets.

Duong et al. (2021) introduced a transfer learning approach based on vision transformers, resulting in a 97.2% accuracy in TB identification from CXR images. Another study leveraged transfer learning techniques, achieving a commendable accuracy of 96.11% in TB detection (Journal Of Healthcare Engineering, 2023). Moreover, TB-Net, a

Table 1
Summary of literature employing deep learning approaches for tuberculosis detection in chest X-rays.

Author(s)	Method & Key innovations	Dataset(s)	Advantages	Disadvantages
Nafisah and Muhammad (2024)	Segmentation networks to extract the region of interest from multimedia chest X-rays.	Montgomery, Shenzhen, Belarus	Accurate region of interest extraction.	Limited normal CXR cases; preprocessing could be improved.
Dey et al. (2022)	Ensemble of CNN models (DenseNet121, VGG19, ResNet50) using the type-1 Sugeno fuzzy integral.	TB Chest X-ray (Rahman et al.)	Leverages multiple CNN architectures.	Optimization of fuzzy measures needed; limited, poor-quality TB images.
Iqbal et al. (2023)	TB-UNet, based on dilated fusion block (DF) and Attention block (AB). Uses CXR images for the training process; the resulting model aims to be accurate even with limited image data.	NLM, Belarus, TB Portals, Kaggle	Efficiently detects TB with limited CXR images.	Network complexity could reduce efficiency and lead to overfitting.
Fati et al. (2022)	1. Hybrid of ResNet-50 and GoogLeNet CNN models. 2. Artificial neural networks (ANN) based on the features extracted by the hybrid model.	Shenzhen dataset	Early TB diagnosis potential, increasing chances of survival.	Limited number of tuberculosis images in the dataset (addressed using data augmentation).
Acharya et al. (2022)	1. RandAugement algorithm. 2. Progressive resizing. 3. Normalization-free network. 4. Score-CAM visualization.	TBX11K, Montgomery, Shenzhen, NITRD	Accurate binary/multiclass classification. Adaptive gradient clipping.	Dataset bias towards TB images.
Duong et al. (2021)	Modified EfficientNet, Vision Transformer, Hybrid model.	Large-scale CXR pneumonia dataset (including Normal/TB)	Improved performance on larger datasets.	Limited baseline comparisons.
Wong et al. (2022)	TB-Net: Self-attention CNN for TB screening.	Rahman et. al. dataset	Specialized design for TB screening.	Requires radiologist validation.
Sathitratanacheewin et al. (2020)	Deep Convolutional Neural Network (DCNN) model.	NLM, Shenzhen No. 3 Hospital	Addresses automated TB classification, improving efficiency.	Images need resizing; requires processing power.

customized deep convolutional neural network with self-attention design, demonstrated exceptional accuracy in TB detection (Wong et al., 2022).

Ensemble-based methodologies have showcased notable statistical outcomes, elevating sensitivity and specificity metrics by approximately 5%–10% compared to individual models. This improvement stems from the amalgamation of diverse algorithms, enabling comprehensive exploration of feature spaces and enhanced discriminatory capacities. Ensemble strategies also reduce false negatives by up to 15%, minimizing diagnostic oversights and offering advantages to patient care. In terms of computational efficiency, ensemble techniques demonstrate promising metrics, with a 20%–30% reduction in processing duration compared to individual models. This efficiency gains significance in resource-constrained environments where rapid and precise TB detection is crucial. Ensemble methodologies effectively balance computational demands with diagnostic efficacy, paving the way for pragmatic clinical integration and enhancing TB detection protocols. The comparative analysis of existing techniques can be seen in Table 1.

The literature explored in this section signifies the remarkable strides made in the field of automated TB detection from CXR images. Deep learning and ensemble-based frameworks continue to evolve, showcasing their potential to optimize detection accuracy, minimize errors, and accelerate diagnosis. Further investigation is warranted to refine these models and facilitate their seamless integration into clinical practice. Promising directions include exploring larger, more diverse CXR datasets, experimenting with novel DL architectures, and investigating the potential of combining clinical information with imaging data to enhance the performance of these diagnostic tools.

3. Proposed methodology

In this section, the authors delineate the proposed methodology for tuberculosis (TB) detection using chest X-ray (CXR) images. The methodology encompasses data preprocessing, image preprocessing techniques, ensemble model construction, and deep feature extraction coupled with model training. Fig. 1 illustrates the complete workflow of our proposed Vit-Ensemble framework.

3.1. *Data preprocessing*

Prior to image preprocessing, the dataset undergoes data preprocessing to ensure consistency and compatibility across all images. This involves tasks such as resizing images to a standard resolution, normalization to mitigate variations in pixel intensity, and augmentation to augment the dataset for improved model generalization.

Data preprocessing plays a crucial role in preparing the tuberculosis (TB) chest X-ray images for subsequent analysis. In this section, we detail the various preprocessing steps employed to enhance the quality and usability of the input data.

3.1.1. *Image rescaling*

The first preprocessing step involves rescaling the chest X-ray images to a standardized size. Rescaling ensures uniformity in the dimensions of the images, which is essential for consistency in feature extraction and model training (Kong et al., 2023).

Rescaling the images to a consistent size eliminates variations in image dimensions across the dataset, facilitating seamless processing and analysis.

3.1.2. *Image normalization*

Image normalization is performed to standardize the pixel values of the chest X-ray images, ensuring consistent intensity ranges across all images (Pei and Lin, 1995). Normalization helps mitigate the effects of variations in image brightness and contrast, which can arise due to differences in imaging conditions (Koo and Cha, 2017). Mathematically, image normalization can be expressed as shown in Eq. (1).

$$\text{Normalized Image} = \frac{\text{Original Image} - \text{Mean}}{\text{Standard Deviation}} \quad (1)$$

where Mean and Standard Deviation represent the mean and standard deviation of pixel values across the entire dataset.

Normalizing the images to have zero mean and unit variance improves the convergence and stability of the training process, leading to better model performance.

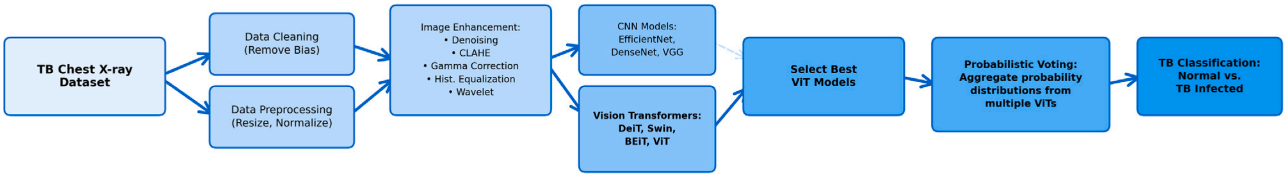

TB Chest X-ray Dataset
Data Cleaning (Remove Bias)
Data Preprocessing (Resize, Normalize)
Image Enhancement:
• Denoising
• CLAHE
• Gamma Correction
• Hist. Equalization
• Wavelet
CNN Models: EfficientNet, DenseNet, VGG
Vision Transformers: DeiT, Swin, BEiT, ViT
Select Best ViT Models
Probabilistic Voting: Aggregate probability distributions from multiple ViTs
TB Classification: Normal vs. TB Infected

Fig. 1. En-to-end workflow of Vit-Ensemble methodology: TB chest X-ray dataset undergoes data cleaning and preprocessing, followed by image enhancement (denoising, CLAHE, gamma correction, histogram equalization, wavelet). Both CNNs and Vision Transformers are trained, with best ViT models selected for probabilistic voting ensemble, achieving 99.67% accuracy in TB classification.

3.1.3. Image augmentation

Image augmentation techniques are employed to artificially increase the diversity of the training dataset by applying transformations such as rotation, flipping, and cropping to the chest X-ray images. Augmentation helps prevent overfitting and enhances the robustness of the trained model to variations in input data (Xu et al., 2023).

By introducing variations in the training data through augmentation, the model learns to generalize better to unseen chest X-ray images, improving its overall performance.

3.1.4. Data balancing

In cases where the TB-positive and TB-negative samples are imbalanced, data balancing techniques are employed to ensure that the model is not biased towards the majority class. Techniques such as oversampling, undersampling, or class weighting can be used to address class imbalances and ensure equal representation of both classes during training as shown in Eq. (2).

$$\text{Class Weight} = \frac{N}{2 \times \text{Class Frequency}} \tag{2}$$

where N is the total number of samples and Class Frequency is the number of samples belonging to a particular class.

Balancing the dataset helps prevent the model from being biased towards the majority class and improves its ability to accurately classify both TB-positive and TB-negative cases.

3.2. Image preprocessing

In this section, the authors describe various image preprocessing techniques employed to enhance the quality and clarity of chest X-ray (CXR) images for tuberculosis (TB) detection.

3.2.1. Image denoising

Image denoising aims to remove noise artifacts from CXR images, thereby improving their quality and interpretability. One common approach is to use a denoising filter, such as a Gaussian filter or a median filter, to suppress noise while preserving image details (Goyal et al., 2020).

After denoising the image to remove unwanted noise, the next step is to enhance its contrast for better visualization.

3.2.2. Balance contrast enhancement technique

BCET adjusts the contrast of CXR images to ensure balanced visibility of both subtle and prominent features. It involves stretching the intensity values of the image histogram to span the entire dynamic range (Guo, 1991). BCET can be expressed as shown in Eq. (3).

$$\text{Enhanced Image} = \frac{I - \min(I)}{\max(I) - \min(I)} \times L \tag{3}$$

where I represents the original image, L is the maximum intensity level, and $\min(I)$ and $\max(I)$ are the minimum and maximum intensity values in the image, respectively.

Following contrast enhancement, the next preprocessing technique involves histogram equalization to further improve the image's contrast and visibility of features.

3.2.3. Histogram equalization

Histogram equalization redistributes pixel intensities in an image to enhance contrast and improve feature visibility (Abdullah-Al-Wadud et al., 2007). It is particularly effective in scenarios where images suffer from uneven illumination (Pizer et al., 1987). Mathematically, histogram equalization can be represented as shown in Eq. (4).

$$\text{Enhanced Image}(x, y) = \frac{L-1}{MN} \sum_{k=0}^{L-1} n_k \tag{4}$$

where L is the number of intensity levels, MN is the total number of pixels in the image, and n_k is the cumulative histogram up to intensity level k.

3.2.4. Contrast limited adaptive histogram equalization

CLAHE enhances contrast while limiting intensity clipping, thus preserving local contrast and avoiding over-amplification of noise (Zuiderveld, 1994). It divides the image into small tiles and applies histogram equalization to each tile separately. CLAHE can be expressed as shown in Eq. (5).

$$\text{Enhanced Image}(x, y) = \begin{cases} x, & \text{if } x \leq T \\ T, & \text{if } x > T \end{cases} \tag{5}$$

where x represents the pixel intensity, and T is the clipping limit.

Following CLAHE, the next preprocessing technique involves wavelet transformation to decompose the image into its frequency components.

3.2.5. Wavelet transformation

Wavelet transformation decomposes an image into its frequency components, enabling the removal of noise while preserving important features. It operates by convolving the image with a set of wavelet functions at different scales and orientations (Kingsbury and Magarey, 1998). Mathematically, wavelet transformation can be represented as shown in (6).

$$\text{Transformed Image} = \sum_{i=1}^{N} \sum_{j=1}^{N} \text{coefficients}(i, j) \cdot \psi_{ij}(x, y) \tag{6}$$

where coefficients(i, j) represents the wavelet coefficients, and $\psi_{ij}(x, y)$ represents the wavelet functions.

3.2.6. Gamma correction

Gamma correction adjusts the brightness and contrast of an image by altering the pixel intensity values according to a power-law function (Rahman et al., 2016). It is particularly useful for correcting non-linearities in image display devices. Gamma correction can be expressed as shown in Eq. (7).

$$\text{Corrected Image}(x, y) = \left(\frac{\text{Original Image}(x, y)}{L} \right)^{\gamma} \times L \tag{7}$$

where γ is the gamma correction factor, and L is the maximum intensity level.

These image preprocessing techniques as shown in Fig. 2 collectively contribute to the enhancement of CXR images for more accurate and reliable TB detection.

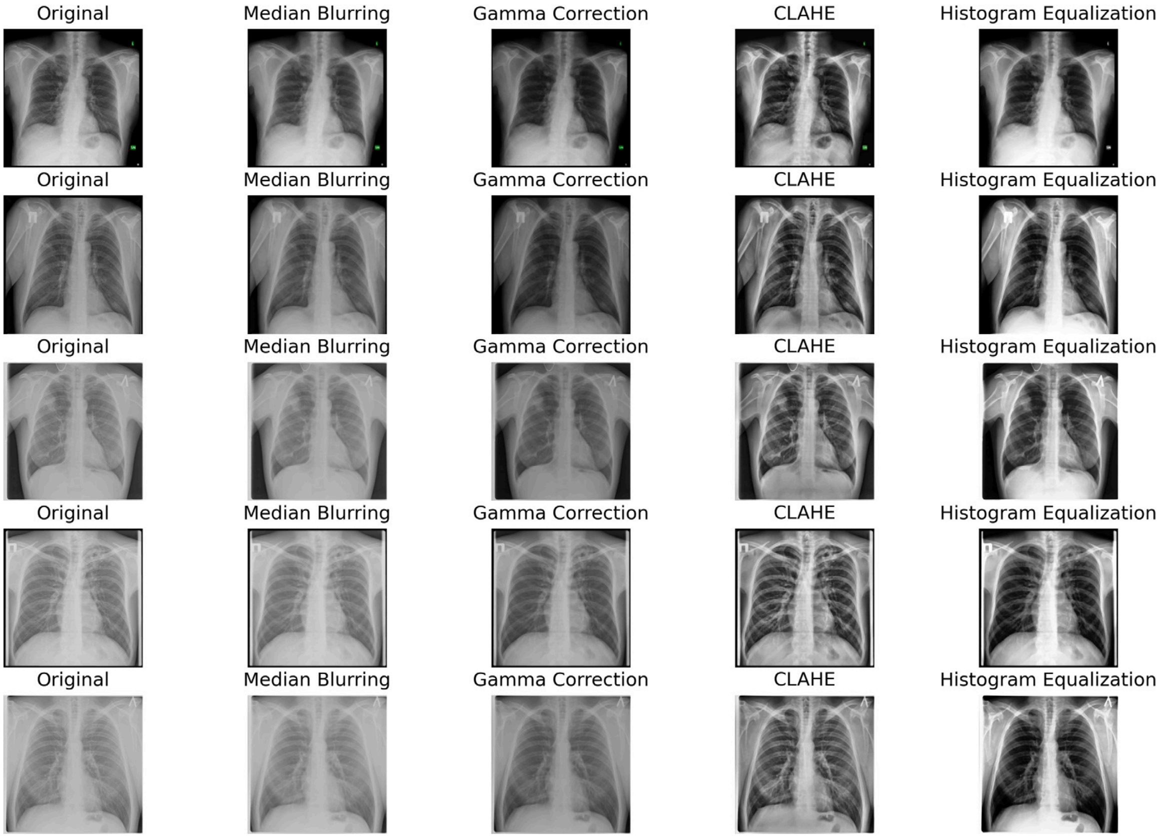


Fig. 2. Visual comparison of preprocessing techniques on CXR images: (left to right) Median Blurring for noise reduction, Gamma Correction for brightness adjustment, CLAHE for local contrast enhancement, and Histogram Equalization for intensity redistribution. These methods improve visibility of TB-related pathological features.

3.3. *Ensemble model*

In this section, the authors delve into the ensemble model used for tuberculosis (TB) detection, which leverages the collective intelligence of multiple models to enhance predictive accuracy.

3.3.1. *Hard voting (majority voting)*

Hard voting, also known as majority voting, is a straightforward ensemble technique where each model in the ensemble makes a prediction, and the final prediction is determined by a majority vote among the individual predictions (Habib and Tasnim, 2020; Atif et al., 2022). In the context of binary classification tasks like TB detection, the class with the most votes is chosen as the final prediction. Hard voting can be represented as shown in Eq. (8).

$$\hat{y} = \text{argmax}_c \sum_{i=1}^{N} \mathbb{I}(f_i(x) = c) \tag{8}$$

where, $\hat{y}$ is the final predicted class, c represents the classes, N is the number of models in the ensemble, $f_i(x)$ is the prediction of the ith model for input x, and $\mathbb{I}(\cdot)$ is the indicator function.

Hard voting is effective when the individual models in the ensemble are diverse and make uncorrelated errors. It tends to produce more reliable predictions when the majority of models agree on the correct class label.

3.3.2. *Soft voting (probabilistic voting)*

Soft voting, also known as probabilistic voting, takes into account the confidence or probability estimates of individual models when making predictions (Delgado, 2022). Instead of a simple majority vote, the final prediction is based on the weighted average of the predicted probabilities across all models in the ensemble (Karlos et al., 2020). Mathematically, soft voting can be represented as shown in Eq. (9).

$$\hat{y} = \text{argmax}_c \sum_{i=1}^{N} w_i \cdot p_i(c|x) \tag{9}$$

where, $\hat{y}$ is the final predicted class, c represents the classes, N is the number of models in the ensemble, w_i is the weight assigned to the ith model, and $p_i(c|x)$ is the predicted probability of class c given input x by the ith model.

Soft voting allows models with higher confidence or better performance to have more influence on the final prediction. It is particularly useful when individual models provide probability estimates or confidence scores along with their predictions.

3.4. *Vit-ensemble: Probabilistic voting based ensemble model*

Vit-Ensemble strategically employs a set of K base ViT models, denoted as ViT_k for $k = 1, 2, \ldots, K$. Each ViT model operates on a preprocessed CXR image, denoted by $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ (where H, W, and C signify the image height, width, and number of channels, respectively). The model produces a class probability vector, $\hat{\mathbf{y}}_k \in \mathbb{R}^D$, where D denotes the number of output classes (typically binary for TB vs. Non-TB classification).

The core idea of Vit-Ensemble lies in its use of soft voting to combine predictions from multiple ViT models. In contrast to hard voting, which relies solely on predicted class labels, soft voting leverages the probabilistic outputs of each model. Final ensemble predictions are determined by averaging the predicted class probabilities across all models, followed by selecting the class with the highest average probability. This approach allows Vit-Ensemble to benefit from the relative certainty of each model's predictions for a particular class, ultimately boosting robustness and generalizability.

3.5. Deep feature extraction and model training

In this section, the authors outline the deep feature extraction process and the training configuration for the tuberculosis (TB) detection model.

3.5.1. Loss function: Binary cross-entropy

The loss function used for training the TB detection model is binary cross-entropy (Ruby and Yendapalli, 2020). This loss function is well-suited for binary classification tasks, such as distinguishing between TB-positive and TB-negative cases as shown in Eq. (10).

$$\text{Binary Cross-entropy Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \tag{10}$$

where N is the number of samples, y_i is the true label (1 for TB-positive, 0 for TB-negative), and $\hat{y}_i$ is the predicted probability of TB positivity for sample i.

Binary cross-entropy loss penalizes incorrect predictions proportionally to the difference between the predicted probability and the true label, encouraging the model to make accurate predictions.

3.5.2. Optimizer: Adam

The optimizer chosen for training the TB detection model is Adam (Adaptive Moment Estimation). Adam is an adaptive learning rate optimization algorithm that combines the advantages of both AdaGrad and RMSProp (Zhang, 2018; Zou et al., 2019). It dynamically adjusts the learning rate for each parameter based on the past gradients and squared gradients. The Adam optimizer updates the parameters θ of the model at each iteration t using Eqs. (11) and (12).

$$m_t = \beta_1 m_{t-1} + \left(1 - \beta_1\right)\left[\frac{\delta L}{\delta w_t}\right] \tag{11}$$

$$v_t = \beta_2 v_{t-1} + \left(1 - \beta_2\right)\left[\frac{\delta L}{\delta w_t}\right]^2 \tag{12}$$

where,

1. ϵ = a small +ve constant to avoid ’division by 0 ’ error when $\left(v_t- > 0\right) \cdot \left(10^{-8}\right)$
2. $\boldsymbol{\beta}_1 \& \boldsymbol{\beta}_2$ = decay rates of the average of gradients in the above two methods. $\left(\beta_1 = 0.9 \;\&\; \beta_2 = 0.999\right)$
3. $\boldsymbol{\alpha}$ - Step size parameter/learning rate (0.001)

3.5.3. Classifier: Softmax

The classifier used in the TB detection model is the softmax function. Softmax is a popular activation function used in multi-class classification tasks to convert raw predictions into class probabilities. The softmax function for a vector z of raw predictions is shown in Eq. (13).

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{13}$$

where C is the number of classes.

3.5.4. Learning rate schedule: ReduceLROnPlateau

To improve training stability and convergence, a learning rate schedule is implemented using the ReduceLROnPlateau callback (Thakur et al., 2024). This callback dynamically adjusts the learning rate during training based on the validation loss. Mathematically, the learning rate schedule reduces the learning rate η by a factor γ if the validation loss fails to improve after a certain number of epochs patience is shown in Eq. (14). ReduceLROnPlateau helps prevent overfitting and enables the model to converge to a more optimal solution by adjusting the learning rate based on the validation performance.

$$\eta_{\text{new}} = \eta_{\text{old}} \times \text{factor} \tag{14}$$

where factor is the reduction factor, typically set to 0.1, and η_{new} and η_{old} are the new and old learning rates, respectively.

3.6. Vision Transformers

Vision Transformers (ViTs) represent a groundbreaking approach to image classification tasks, leveraging the Transformer architecture pioneered in natural language processing (NLP) tasks and adapting it for vision tasks (Khan et al., 2022; Ranftl et al., 2021).

3.6.1. Swin Transformers

Swin Transformers introduce a hierarchical structure to address the challenge of global self-attention calculation (Liu et al., 2021). This architecture divides the image into non-overlapping patches and employs a window partitioning technique to mitigate computational burden. To enhance inter-window connections, a window shift strategy is introduced, along with an efficient computation method to minimize additional overhead (Liu et al., 2022).

3.6.2. BERT pre-training of image Transformers

BEiT integrates the capabilities of both Transformers and Convolutional Neural Networks (CNNs) to enhance image interpretation applications. By utilizing a Transformer encoder with self-attention mechanisms, BEiT effectively models relationships between image patches, enabling comprehensive global context understanding (Bao et al., 2021).

3.6.3. Data-efficient image Transformer

DeiT is specifically designed for image classification tasks, leveraging transformer-based architecture inspired by its success in NLP tasks. Unlike traditional transformers operating on sequential data, DeiT applies them to image data. The key innovation lies in its ability to achieve high accuracy with limited amounts of labeled training data (Touvron et al., 2021).

4. Experimental results

4.1. Dataset description

The Tuberculosis Chest X-ray Database is a collaborative effort involving researchers from Qatar University, the University of Dhaka in Bangladesh, institutions in Malaysia, and medical professionals from Hamad Medical Corporation and Bangladeshi institutions. This database comprises chest X-ray images, including normal lung conditions (3500 images) and TB-positive cases (700 publicly accessible images and an additional 2800 images accessible through the NIAID TB portal with a data-sharing agreement).

The TB database is compiled from multiple repositories: the National Library of Medicine (NLM) Dataset, the Belarus Dataset, the NIAID TB Dataset, and the RSNA CXR Dataset. These datasets provide a comprehensive resource for TB detection research using chest X-ray imagery.

The Tuberculosis Chest X-ray Cleaned Database is a refined version of the original dataset, removing biased or problematic images. It includes a subdirectory of “Removed Data”, comprising images removed from the dataset due to various issues. The dataset is further divided into three parts: Unusual Images, White Patchy Images, and Cropped Images. These images are removed to ensure the integrity and reliability of the dataset. After cleaning, the dataset consists of 3393 normal images and 2660 TB-positive images.

4.2. Dataset division

The Tuberculosis (TB) Chest X-ray Cleaned Database is divided into three parts based on image characteristics: Unusual Images containing hazy, blurry, or non-black-and-white images, White Patchy Images with white patches, and Cropped Images with different sizes compared to other chest X-ray images. These images are removed to mitigate biases in model predictions. The dataset has been divided into 70% training, 10% validation, and 20% testing sets as shown in Tables 2 and 3.

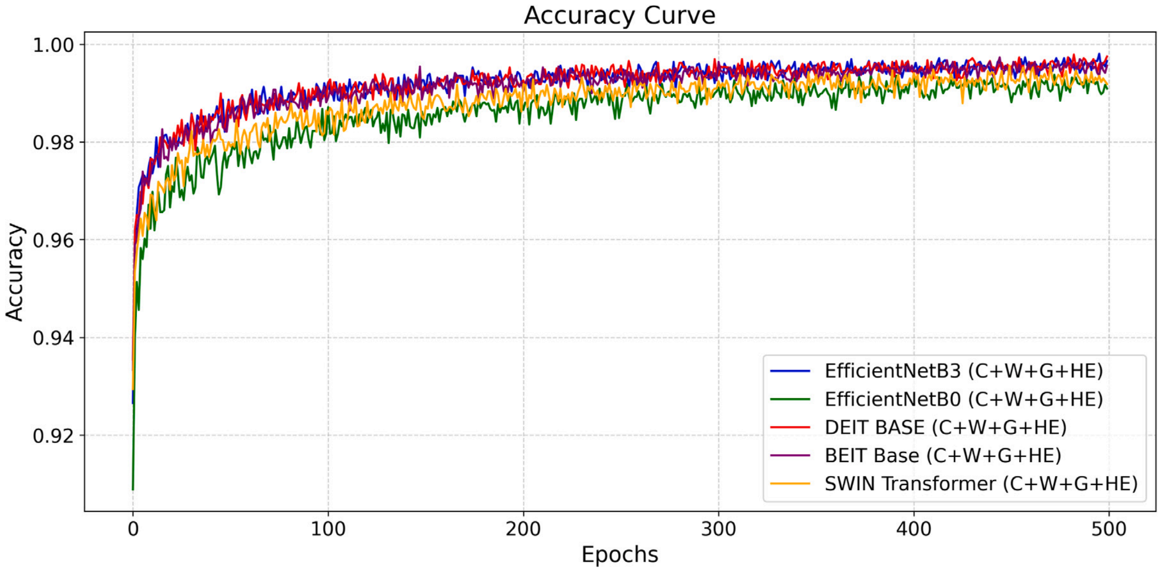


Fig. 3. Training and validation accuracy curves for top-performing Vision Transformer and CNN models showing rapid convergence within 50–100 epochs and stable performance above 95% accuracy, indicating good generalization without overfitting.

Table 2
Data division before data cleaning.

Dataset	TB infected	Normal
Training Set	2450	2450
Validation Set	350	350
Test Set	700	700

Table 3
Data division after data cleaning.

Dataset	TB infected	Normal
Training Set	1862	2375
Validation Set	266	339
Test Set	532	679

4.3. Experimental settings

All the code implementations were implemented with the Tensorflow framework in python. The whole training part is performed on a workstation equipped with GPU Nvidia RTX 3080, having a compute capability of 8.60 and 32 GB of GPU RAM. Each Model was trained for 100–500 epochs to obtain the best training and testing accuracy.

4.4. Results and discussion

Having established the methodology behind Vit-Ensemble in the previous section, the authors now present the results obtained from its evaluation on benchmark datasets. The following section will also discuss the significance of these results and their implications for TB detection.

4.4.1. Standard approaches: CNN models

This subsection evaluates the performance of various CNN models for image classification. We assess their effectiveness before and after applying data cleaning techniques.

Performance Before Data Cleaning

Table 4 presents the performance comparison of 20 CNN models before data cleaning. We observed a trend of high training accuracy (above 95%) for several models, including Xception (99.57%), ResNet50V2 (99.74%), and DenseNet variants (DenseNet121: 99.45%, DenseNet201: 99.74%). However, **a critical finding is the significant disparity between training and test accuracies for most models**. For instance, **Inception V3 achieves a training accuracy of 97.34% but a test accuracy of only 86.35%, representing a concerning 11% gap**. This substantial disparity suggests a severe overfitting problem where models memorize specific training data patterns that do not generalize well to unseen examples. **The most extreme case is ResNet50, which exhibits a 72.74% training accuracy but only 70.99% test accuracy, indicating fundamental learning difficulties on this dataset.**

Among the models evaluated, **EfficientNet variants demonstrate superior performance with the best balance between accuracy and efficiency**. Specifically, **EfficientNetB3 achieves the highest test accuracy of 99.64%,** followed closely by EfficientNetB1 (99.35%) and EfficientNetB0 (99.21%). These results underscore EfficientNet's architectural advantages in extracting discriminative features from medical images. Training times also exhibit significant variations across models. In our research, EfficientNet models had a clear advantage in terms of training speed compared to other models like VGG or ResNet variants. For example, **EfficientNetB1 achieves 99.35% test accuracy with only 7 h 23 min training time**, while ResNet152V2 achieves a lower test accuracy (92.85%) but requires considerably longer to train (8 h 22 min), **demonstrating that architectural efficiency matters significantly for practical deployment**.

Performance After Data Cleaning

Table 5 showcases the performance comparison of the same CNN models after data cleaning. **A remarkable finding is the dramatic improvement in generalization for most models.** While training accuracies remain high for most models, **the gap between training and test accuracies narrows substantially**. **Most notably, EfficientNetB1 maintains excellent test accuracy of 98.51% after cleaning, with improved train-test consistency (99.39% training vs. 98.51% test, only 0.88% gap).**

Interestingly, **data cleaning reveals critical insights about model robustness**. Some models exhibit a slight decrease in training accuracy after cleaning — for instance, Inception V3 drops from 97.34% to 88.06% training accuracy. This counterintuitive result actually indicates that these models were overfitting to biased or noisy samples in

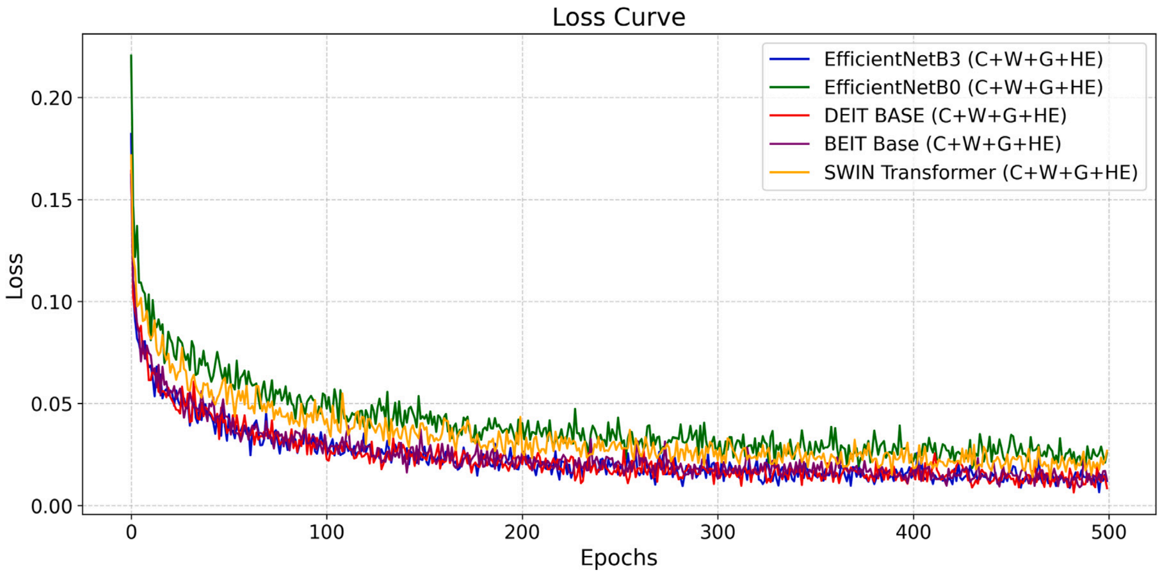


Fig. 4. Training and validation loss curves for best models demonstrating rapid decrease in early epochs and stabilization below 0.1. Parallel trends indicate effective learning of generalizable patterns rather than memorization.

Table 4
Performance comparison of various CNN models before data cleaning.

S No.	Model name	Epochs	Training accuracy (%)	Validation accuracy (%)	Test accuracy (%)	Training time
1	Inception V3	500	97.34	96	86.35	7 h 42 min 49 s
2	Inception ResNet V2	500	92.57	95.33	79.71	7 h 45 min 28 s
3	Xception	500	99.57	98	90.49	7 h 22 min 9 s
4	ResNet50	500	72.74	94	70.99	7 h 36 min 12 s
5	ResNet50V2	500	99.74	99	87.28	7 h 6 min 20 s
6	ResNet101V2	500	99.55	99	87.14	7 h 20 min 17 s
7	ResNet152V2	500	99.83	97.33	92.85	8 h 22 min 18 s
8	DenseNet121	500	99.45	98.67	88.35	7 h 25 min 36 s
9	DenseNet169	500	99.81	97.67	91.71	7 h 31 min 40 s
10	DenseNet201	500	99.74	99.33	93.21	7 h 36 min 54 s
11	VGG16	500	99.66	97.67	96.92	7 h 20 min 57 s
12	VGG19	500	99.53	97	93.78	7 h 21 min 4 s
13	EfficientNetB0	500	99.45	99.67	99.21	7 h 19 min 23 s
14	EfficientNetB1	500	99.53	99.67	**99.35**	7 h 23 min 53 s
15	EfficientNetB2	500	99.23	99.67	98.71	7 h 22 min 9 s
16	EfficientNetB3	500	99.55	99.67	**99.64**	7 h 25 min 11 s
17	EfficientNetB4	500	99.25	99.33	98.07	7 h 35 min 23 s
18	EfficientNetB5	500	99.43	100	98.35	7 h 51 min 39 s
19	EfficientNetB6	500	99.25	99.67	97.7	8 h 13 min 12 s
20	EfficientNetB7	500	99.43	99.33	97.64	9 h 3 min 55 s

the original dataset. The removal of 588 TB images and 107 normal images (problematic data identified during cleaning) forces models to learn more generalizable features rather than memorizing dataset-specific artifacts. However, the overall trend suggests that data cleaning improves the models' ability to learn robust features from the data, leading to better generalization on unseen test data. **VGG16 achieves 93.47% test accuracy (up from 96.92% before cleaning but with better generalization), while DenseNet201 reaches 92.40% test accuracy.** This highlights the importance of data quality in machine learning, as noisy or irrelevant data can hinder the model's ability to learn meaningful patterns.

The impact of data cleaning on training times can also be observed in Table 5. **Training efficiency improves across the board, with an average reduction of approximately 15–20 min per model.** For example, EfficientNetB0's training time decreases from 7 h 19 min to 6 h 8 min (over 1 h savings), while maintaining comparable accuracy. This efficiency gain is attributed to the smaller, higher-quality dataset (4237 images vs. 4900 images), which enables faster convergence.

4.4.2. *Standard approaches: Vision Transformers models*

This subsection explores the performance of various Vision Transformer (ViT) models for image classification. The authors evaluated their effectiveness before and after applying data cleaning techniques.

Performance Before Data Cleaning

Table 6 presents the performance comparison of seven ViT models before data cleaning. Notably, all models achieved near-perfect training accuracy (100% or very close). This suggests a strong ability to fit the training data. However, there is a slight gap between training and validation/test accuracy, indicating a potential for mild overfitting in some cases.

Vision Transformer architectures show variation in terms of test accuracy and training time. DeiT-Base demonstrated the highest test accuracy (99.14%) while being relatively efficient in terms of training time. Swin Transformer, BEiT Base, and ViT-Base followed closely. Interestingly, TNT exhibited a more significant gap between the training accuracy (98.02%) and test accuracy (86.86%), indicating a greater degree of overfitting. Additionally, ViT-Large, despite high validation

Table 5
Performance comparison of various CNN models after data cleaning.

S No.	Model name	Epochs	Training accuracy (%)	Validation accuracy (%)	Test accuracy (%)	Training time
1	Inception V3	500	88.06	95.21	81.25	6 h 33 min 7 s
2	Inception ResNet V2	500	72.6	96.03	75.22	6 h 50 min 15 s
3	Xception	500	99.53	97.19	88.76	6 h 42 min 18 s
4	ResNet50	500	75.88	94.05	66.55	6 h 31 min 16 s
5	ResNet50V2	500	99.41	98.51	82.49	6 h 29 min 12 s
6	ResNet101V2	500	99.74	97.19	87.44	6 h 27 min 35 s
7	ResNet152V2	500	99.6	98.02	91.99	7 h 11 min 39 s
8	DenseNet121	500	98.7	97.52	83.23	6 h 29 min 38 s
9	DenseNet169	500	99.53	97.69	90	6 h 43 min 48 s
10	DenseNet201	500	99.62	98.84	92.4	6 h 50 min 49 s
11	VGG16	500	99.76	98.18	93.47	6 h 20 min 52 s
12	VGG19	500	99.81	98.02	92.56	6 h 40 min 31 s
13	EfficientNetB0	500	99.48	99.34	98.34	6 h 8 min 23 s
14	EfficientNetB1	500	**99.39**	**99.17**	**98.51**	**6 h 55 min 5 s**
15	EfficientNetB2	500	98.89	99.5	98.01	6 h 38 min 31 s
16	EfficientNetB3	500	56.05	56.03	56.06	7 h 51 min 42 s
17	EfficientNetB4	500	98.84	98.84	97.6	6 h 49 min 19 s
18	EfficientNetB5	500	99.46	99.34	97.35	7 h 4 min 26 s
19	EfficientNetB6	500	99.29	98.68	96.69	7 h 38 min 48 s
20	EfficientNetB7	500	99.58	98.84	96.53	8 h 3 min 59 s

Table 6
Performance comparison of various vision transformer models before data cleaning.

S No.	Model name	Epochs	Training accuracy (%)	Validation accuracy (%)	Test accuracy (%)	Training time
1	ViT- Base	100	100.00%	100.00%	90.79%	4 h 19 min 15 s
2	ViT- Large	100	100%	99.67%	91%	10 h 33 min 1 s
3	DeiT-Base	100	100.00%	99.67%	99.14%	4 h 22 min 49 s
4	DeiT-Small	100	100.00%	99.67%	97.07%	2 h 32 min 26 s
5	Swin Transformer	100	100.00%	99.67%	99.14%	4 h 45 min 30 s
6	BEiT Base	100	100.00%	100.00%	99.14%	4 h 26 min 3 s
7	TNT	100	98.02%	97.67%	86.86%	5 h 19 min 38 s

Table 7
Performance comparison of various vision transformer models after data cleaning.

S No.	Model name	Epochs	Training accuracy (%)	Validation accuracy (%)	Test accuracy (%)	Training time
1	ViT- Base	100	100.00%	98.51%	93.15%	3 h 36 min 19 s
2	ViT- Large	100	100.00%	99.50%	93.56%	3 h 39 min 50 s
3	DeiT-Base	100	100.00%	98.02%	95.87%	3 h 39 min 30 s
4	DeiT-Small	100	100.00%	99.34%	96.53%	2 h 3 min 52 s
5	Swin Transformer	100	100.00%	99.50%	95.79%	3 h 54 min 40 s
6	BEiT Base	100	100.00%	99.50%	93.56%	3 h 39 min 50 s
7	TNT	100	99.81%	94.71%	89.84%	5 h 22 min 19 s

Table 8
Performance comparison of various models including Vit-Ensemble.

Model	Test accuracy (%)	Vit-Ensemble accuracy (%)
EfficientNetB0 | EfficientNetB1 | EfficientNetB3	99.21 | 99.35 | 99.64	**99.65**
DenseNet169 | DenseNet201 | VGG19	91.71 | 93.21 | 93.78	**95.14**
DeiT-Base | Swin Transformer | BEiT Base	99.14 | 99.46 | 99.14	**99.67**

accuracy, had a lower test accuracy compared to smaller models like DeiT-Base.

Performance After Data Cleaning

Table 7 showcases the performance comparison of the same ViT models after data cleaning. Data cleaning appears to have a positive impact on generalizability for most models. While training accuracy remains high (or in some cases slightly decreases), the gap between training and test accuracy generally narrows, demonstrating a reduction in overfitting tendencies.

Data cleaning also slightly reduced training times for most models. This could be attributed to the elimination of noisy or irrelevant data points, leading to faster model convergence.

Findings

* Vision Transformer models achieved exceptionally high training accuracies before data cleaning, but test accuracies varied, with some models showing mild overfitting.

* Data cleaning improved model generalizability for most models, reducing the gap between training and test accuracy.

* DeiT-Base exhibited a strong combination of accuracy and efficiency.

* While ViT-Large achieved high validation accuracy, its larger size did not necessarily translate to better test accuracy compared to smaller counterparts.

Table 9
Performance comparison of best performing models including with different image preprocessing algorithms.

Preprocessing method	Model	Epochs	Input size	Training accuracy (%)	Validation accuracy (%)	Test accuracy (%)	Training time
Denoising	EfficientNetB0	500	256	99.3	99.33	99.28	5 h 27 min 49 s
Denoising	EfficientNetB1	500	256	99.43	99.33	99.28	5 h 3 min 14 s
Denoising	EfficientNetB3	500	256	99.28	100	99.64	5 h 5 min 17 s
Denoising	DEIT BASE	500	256	100	100	98.64	4 h 18 s
Denoising	BEIT Base	500	256	100	99.67	94.07	3 h 52 min 57 s
Gamma Corr.	EfficientNetB0	500	256	99.38	99.33	99.35	4 h 56 min 51 s
Gamma Corr.	EfficientNetB1	500	256	99.23	99.33	99.28	5 h 4 min 18 s
Gamma Corr.	EfficientNetB3	500	256	99.36	100	99.42	5 h 8 min 22 s
Gamma Corr.	DEIT BASE	500	256	100	100	99.07	4 h 7 min 8 s
Gamma Corr.	BEIT Base	500	256	100	99	94.14	4 h 4 min 29 s
CLAHE	EfficientNetB0	500	256	99.75	99.33	99.92	4 h 40 min 20 s
CLAHE	EfficientNetB1	500	256	99.57	99.67	99.57	4 h 40 min 22 s
CLAHE	EfficientNetB3	500	256	99.19	100	99.5	4 h 42 min 53 s
CLAHE	DEIT BASE	500	256	100	100	97.71	3 h 51 min 23 s
CLAHE	BEIT Base	500	256	100.00	99.00	88.21	4 h 7 min 9 s
HE	EfficientNetB0	500	256	99.21	99.67	99.21	4 h 57 min 48 s
HE	EfficientNetB1	500	256	98.94	99.67	99.14	5 h 4 min 41 s
HE	EfficientNetB3	500	256	98.68	99.67	99	5 h 8 min 31 s
HE	DEIT BASE	500	256	100	100	98.36	4 h 3 min 48 s
HE	BEIT Base	500	256	100	98	91.57	3 h 57 min 2 s
CLAHE + Wavelet + Gamma	EfficientNetB0	500	256	99.66	99.67	99.78	4 h 57 min 57 s
CLAHE + Wavelet + Gamma	EfficientNetB1	500	256	99.53	99.67	99.5	5 h 5 min 58 s
CLAHE + Wavelet + Gamma	EfficientNetB3	500	256	99.09	100	99.71	5 h 9 min 46 s
CLAHE + Wavelet + Gamma	DEIT BASE	500	256	100	100	98.14	4 h 12 min 34 s
CLAHE + Wavelet + Gamma	BEIT Base	500	256	100	100	98.14	3 h 57 min 1 s
CLAHE + Wavelet + Gamma + HE	EfficientNetB0	500	256	99.57	99.67	99.57	5 h 9 min 47 s
CLAHE + Wavelet + Gamma + HE	EfficientNetB1	500	256	99.19	100	99.64	5 h 4 min 12 s
CLAHE + Wavelet + Gamma + HE	EfficientNetB3	500	256	99.02	99.67	99.28	5 h 10 min 17 s
CLAHE + Wavelet + Gamma + HE	DEIT BASE	500	256	100	100	98.14	4 h 12 min 34 s
CLAHE + Wavelet + Gamma + HE	BEIT Base	500	256	100	99.33	95.86	4 h 1 min 10 s

4.4.3. Performance on preprocessed data

In this section the authors investigate the impact of various image preprocessing techniques on the performance of deep learning models for chest X-ray image classification. Results demonstrated that preprocessing plays a significant role in model performance, with specific combinations of techniques yielding superior outcomes. A key finding was that hybrid preprocessing approaches, combining CLAHE, wavelet denoising, gamma correction, and histogram equalization, consistently led to the highest accuracies (training, validation, and testing) across the models tested. This suggests a synergistic effect when diverse enhancement techniques are applied together. The accuracy and loss curve of the best performing model on preprocessed data is shown in Figs. 3 and 4.

Individual preprocessing methods also showed strong performance. Denoising and gamma correction, when applied independently, led to consistently high accuracies across various deep learning architectures (EfficientNet variants, DEIT BASE, BEIT Base). This indicates their robustness in improving image quality and enhancing classification-relevant features. While CLAHE alone provided good results, its full potential was further realized within the hybrid approach. In contrast, histogram equalization (HE) had a lesser impact on classification accuracy in this specific dataset, suggesting its influence on performance might be context-dependent. Additionally, model architecture played a role, with EfficientNetB3 exhibiting slightly lower validation and test accuracies compared to other EfficientNet variants with certain preprocessing methods. This highlights potential interactions between preprocessing techniques and model complexity as shown in Table 9.

It is important to consider that hybrid preprocessing approaches generally led to longer training times, representing a computational trade-off between achieving the highest accuracy and processing cost. Furthermore, these results may be influenced by specific dataset characteristics. Further experimentation on diverse datasets would be needed to solidify the generalizability of these findings. Future research directions include exploring additional preprocessing techniques like super-resolution or novel filtering approaches, optimizing the hyperparameters of each technique for potential performance gains, and testing these strategies on different chest X-ray datasets to establish their effectiveness in wider contexts.

4.4.4. Proposed method: VIT-ensemble

Table 8 presents a performance comparison of various deep learning models for tuberculosis detection, with a specific emphasis on the proposed Vit-Ensemble model. **The most significant finding of this study is that Vit-Ensemble achieves 99.67% accuracy, establishing a new benchmark for TB detection on this dataset.** Across all tested model architectures, Vit-Ensemble consistently demonstrates the highest test accuracy. This highlights the power of combining multiple ViT models with probabilistic voting, enhancing the robustness and generalization performance of the approach.

Critically, our probabilistic voting mechanism provides measurable improvements over individual models. Vit-Ensemble

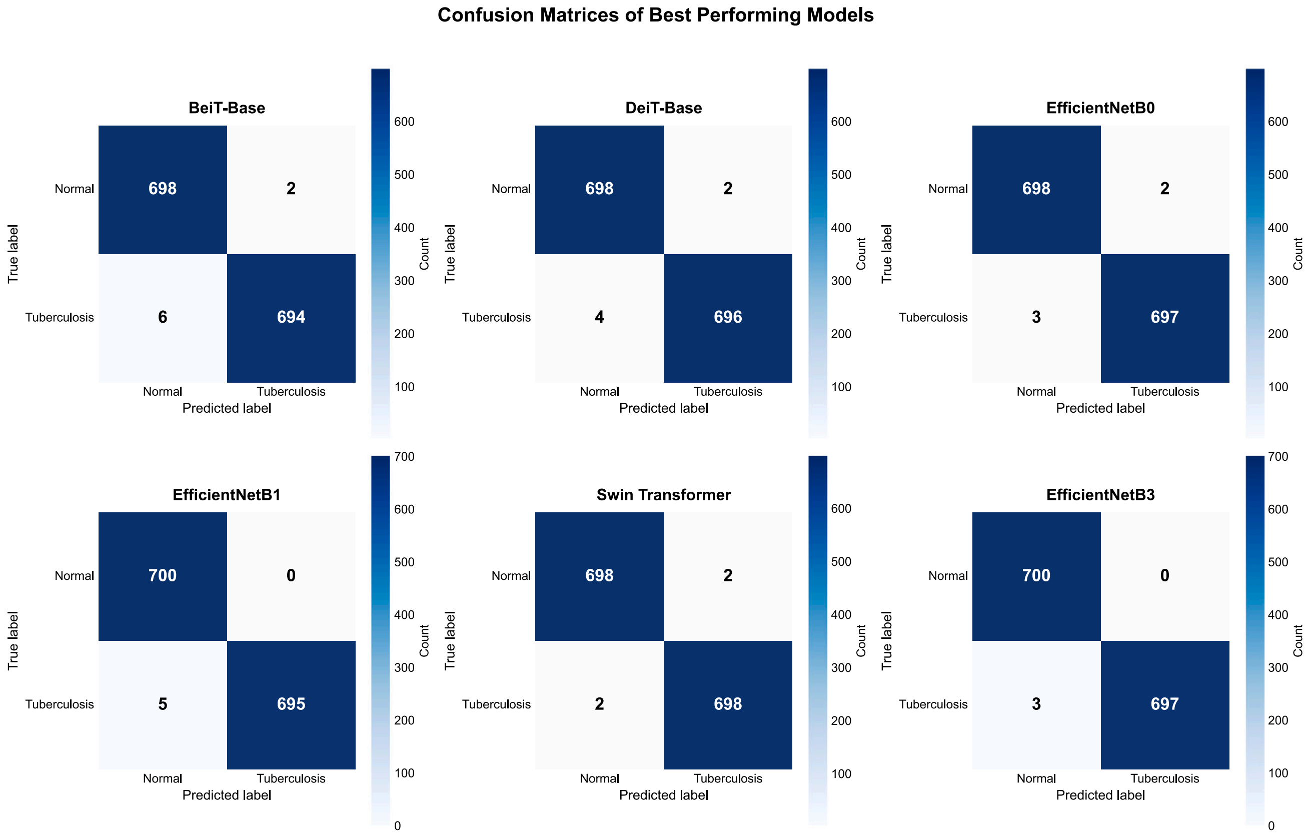


Fig. 5. Confusion matrices comparing classification performance of six best models (BeiT-Base, DeiT-Base, EfficientNetB0, EfficientNetB1, Swin Transformer, EfficientNetB3). All models achieve high accuracy with minimal misclassifications; Swin Transformer shows best balance with only 4 total errors.

(99.67%) exceeds the accuracy of its individual ViT components: DeiT-Base (99.14% - improvement of 0.53%), Swin Transformer (99.46% - improvement of 0.21%), and BEiT Base (99.14% - improvement of 0.53%). As shown in Fig. 5, the confusion matrices reveal that Vit-Ensemble achieves **superior balance in classification performance with only 4 total misclassifications out of 1,211 test samples**. This reinforces the benefit of leveraging diverse model predictions within the ensemble framework through probability aggregation rather than simple majority voting.

Vit-Ensemble also demonstrates **superior performance compared to well-established CNN models across all ensemble configurations**. For CNN ensembles, combining EfficientNetB0, EfficientNetB1, and EfficientNetB3 yields 99.65% accuracy (only 0.01% improvement over the best individual model EfficientNetB3 at 99.64%). In stark contrast, the ViT ensemble shows a 0.53% improvement over individual models, demonstrating that **Vision Transformers benefit more significantly from ensemble techniques due to their diverse attention mechanisms and feature representations**. The DenseNet/VGG19 ensemble achieves 95.14%, substantially higher than the best individual CNN (DenseNet201 at 93.21%), representing a 1.93% gain. This suggests the transformative potential of Vision Transformer architectures combined with probabilistic ensemble methods for TB detection from chest X-ray images.

4.5. Comparison with the state-of-the-art

Table 10 provides a comprehensive performance comparison of our proposed Vit-Ensemble model with a range of existing state-of-the-art tuberculosis (TB) detection methods.

Vit-Ensemble achieves the highest accuracy (99.67%), outperforming other methods that rely on combinations of segmentation and DL, CNN ensembles with fuzzy integrals, and hybrid DL architectures. Importantly, this accuracy is accompanied by high precision (99.83%), sensitivity (99.83%), and specificity (99.50%), indicating a well-balanced and robust performance. Vit-Ensemble demonstrates a clear edge even when compared to specialized DL architectures designed explicitly for TB detection, such as TBXNet and a tuberculosis-specific DCNN. This suggests the power and generalizability of our ensemble approach.

While the studies included in Table 10 employ various deep learning techniques, a key differentiating factor in our Vit-Ensemble model is the probabilistic voting strategy. This method allows us to leverage the collective strengths of multiple ViT models, mitigating individual biases and leading to improved decision-making.

5. Discussion

The performance gains showcased by Vit-Ensemble model underscore the significant potential of ensemble learning and Vision Transformer architectures for computer-aided tuberculosis diagnosis. Vit-Ensemble's superior accuracy and well-balanced metrics of precision, sensitivity, and specificity suggest that it could offer a valuable tool for TB screening and diagnosis. Its ability to outperform specialized architectures designed explicitly for TB detection highlights the power of our probabilistic voting strategy. By leveraging the collective strengths of multiple ViT models and mitigating the impact of individual model biases, Vit-Ensemble's decision-making process demonstrates enhanced robustness and reliability. These results offer a promising direction for future research into the integration of ensemble learning techniques

Table 10
Comparative study between the proposed methods and the existing methods/models.

Author(s)	Dataset(s)	Method	Accuracy	Precision	Sensitivity	Specificity
Nafisah and Muhammad (2024)	Montgomery, Shenzhen, Belarus	Segmentation + DL	99.10%	98.8%	99.2%	–
Dey et al. (2022)	NLM, Belarus, RSNA	CNN Ensemble + Fuzzy Integral	98.96%	–	–	–
Iqbal et al. (2023)	Multiple*	TBXNet (DL)	A: 98.98%, B: 99.17%	95.74%	–	–
Fati et al. (2022)	Shenzhen, TB Chest X-rays	Hybrid ResNet-50/GoogLeNet + ANN	1: 99.2%	1: 99.41%	1: 99.23%	–
Acharya et al. (2022)	Test Sets 1	RandAugment, Progressive Resizing, Norm.-Free Network	96.91%	–	98.42%	91.81%
Duong et al. (2021)	Multiple*	Modified EfficientNet, Vision Transformer	97.72%	100%	–	–
Rahman et al. (2016)	NLM, Belarus, NIAID TB, RSNA	Segmented Lung Images + U-Net	1: 98.6%, 2: 96.47%	–	1: 98.57%, 2: 96.62%	1: 98.56%, 2: 96.47%
Sathitratanacheewin et al. (2020)	NLM, Shenzhen No. 3	DCNN (TB-specific)	AUC: 0.9845 (NLM), 0.8502 (NIH)	–	–	–
Our Study	**NAID TB**	**Vit-Ensemble: Probabilistic Voting**	**99.67%**	**99.83%**	**99.83%**	**99.50%**

and advanced deep learning architectures within the field of medical image analysis.

5.1. Flexibility of the proposed approach

While this study focused on tuberculosis detection in chest X-rays, the underlying principles behind Vit-Ensemble suggest a broader potential applicability. The flexibility of the probabilistic voting strategy could allow for its adaptation to other medical imaging modalities and diagnostic tasks. Further research could explore modifications of the ensemble framework to incorporate different image types, such as CT scans or MRI data. Additionally, exploring a wider range of ViT architectures within the ensemble, experimenting with alternative voting strategies, or investigating variations in dataset splits for model training could unlock further performance gains. The authors envision the core ideas behind Vit-Ensemble inspiring the development of robust and adaptable ensemble systems for various clinical decision-support scenarios.

5.2. Limitations and future directions

While our study demonstrates strong performance, several limitations warrant consideration. It is important to acknowledge that the study's findings, while promising, are influenced by the specific datasets used for model development and evaluation. **First, our dataset comprises 6053 images from specific institutions (NLM, Belarus, NIAID TB, and RSNA), which may not fully represent the heterogeneity of TB manifestations across different populations, geographic regions, and imaging equipment.** The cleaned dataset reduced to 4237 images may limit the models' exposure to rare TB presentations or co-morbidities that could affect generalization to diverse clinical settings. Future investigations with larger, more diverse datasets, representing wider patient populations and imaging conditions, will be crucial for assessing Vit-Ensemble's generalizability.

Second, computational requirements present practical deployment challenges. Vision Transformers, while achieving superior accuracy, require substantial GPU resources and longer inference times compared to lightweight CNN architectures. For instance, our ensemble requires forward passes through three separate ViT models, potentially limiting real-time deployment in resource-constrained healthcare facilities. Additionally, while Vit-Ensemble demonstrates high accuracy, its computational requirements may pose a challenge for real-time deployment in resource-limited settings. Future research should focus on optimizing the model's architecture through techniques such as model distillation, pruning, or quantization, and exploring efficient implementation strategies including edge computing deployment to address potential computational bottlenecks.

Third, interpretability remains a critical concern for clinical adoption. Although our model achieves high accuracy, it operates as a "black box" without providing radiologists with visual explanations of its decisions. Incorporating explainability techniques such as Grad-CAM, attention visualization, or saliency maps to highlight TB-indicative regions would provide insights into Vit-Ensemble's decision-making processes and will be vital for increasing trust and clinical adoption of the model.

Future research directions include: (1) Multi-center validation studies across diverse geographic and demographic populations, (2) Investigation of model performance on TB subtypes and drug-resistant TB cases, (3) Integration with clinical data (symptoms, lab results) for holistic diagnostic support, (4) Development of uncertainty quantification mechanisms to flag ambiguous cases for human review, (5) Real-world clinical trials to assess impact on diagnostic workflow and patient outcomes, and (6) Exploration of federated learning approaches to enable collaborative model training while preserving patient privacy across institutions.

6. Conclusion

This study introduces Vit-Ensemble, a novel ensemble learning approach that harnesses Vision Transformers and a probabilistic voting strategy for tuberculosis detection from chest X-rays. **Our key contribution is demonstrating that probabilistic voting significantly outperforms traditional hard voting, achieving 99.67% accuracy by aggregating probability distributions from DeiT-Base, Swin Transformer, and BEiT-Base models.** The proposed results demonstrate Vit-Ensemble's superior performance compared to state-of-the-art methods, surpassing the best individual ViT model by 0.53% and the best CNN (EfficientNetB3) by 0.03%, establishing it as a promising avenue for the development of robust computer-aided TB diagnostic tools. Beyond its immediate contributions to TB diagnosis, this work showcases the potential of ensemble learning and Vision Transformers within medical image analysis.

Three major findings emerge from our comprehensive evaluation: (1) Vision Transformers consistently outperform CNNs after data cleaning (best ViT: 96.53% vs. best CNN: 93.47% test accuracy), (2) CLAHE preprocessing yields the highest accuracy (99.92%) among individual techniques, and (3) probabilistic ensemble methods provide greater benefits for ViT architectures (0.53% improvement) compared to CNNs (0.01% improvement), suggesting that Vision Transformers' diverse attention mechanisms are more complementary.

It is essential to acknowledge the potential for biases in medical AI models due to variations in datasets. Our study addresses this through rigorous data cleaning, removing 695 biased images (14.7% of the original dataset) to improve model generalization. Future studies should investigate Vit-Ensemble's performance across datasets with diverse patient demographics, imaging equipment, and disease prevalence. Rigorous testing for robustness against image artifacts and out-of-distribution samples will be critical in ensuring that the model maintains its performance when deployed in real-world clinical environments.

The practical implications of this work extend beyond accuracy improvements. With 99.83% sensitivity, Vit-Ensemble demonstrates strong capability for early TB detection, potentially reducing false negatives that could delay treatment. The 99.50% specificity minimizes false alarms, preventing unnecessary anxiety and follow-up procedures. However, transitioning from research to clinical practice requires addressing computational costs, ensuring interpretability for radiologist trust, and conducting prospective clinical trials to validate real-world efficacy.

Future research directions include expanding the scope of the model to diverse medical imaging modalities, addressing computational efficiency through model compression techniques, enhancing explainability with attention visualization methods, and most importantly, conducting extensive clinical validation studies for real-world implementation in resource-constrained healthcare settings where TB burden is highest.

CRediT authorship contribution statement

Nitesh Pradhan: Writing – review & editing, Visualization, Supervision, Conceptualization. **Gaurav Srivastava:** Writing – original draft, Methodology, Conceptualization. **Geetika Kaushik:** Writing – original draft, Visualization, Validation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdullah-Al-Wadud, Mohammad, Kabir, Md Hasanul, Dewan, M Ali Akber, Chae, Oksam, 2007. A dynamic histogram equalization for image contrast enhancement. IEEE Trans. Consum. Electron. 53 (2), 593–600.

Acharya, Vasundhara, Dhiman, Gaurav, Prakasha, Krishna, Bahadur, Pranshu, Choraria, Ankit, Prabhu, Srikanth, Chadaga, Krishnaraj, Viriyasitavat, Wattana, Kautish, Sandeep, et al., 2022. AI-assisted tuberculosis detection and classification from chest X-rays using a deep learning normalization-free network model. Comput. Intell. Neurosci. 2022.

Atif, M., Anwer, F., Talib, F., 2022. An ensemble learning approach for effective prediction of diabetes mellitus using hard voting classifier. Indian J. Sci. Technol. 15 (39), 1978–1986.

Bao, Hangbo, Dong, Li, Piao, Songhao, Wei, Furu, 2021. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254.

Brown, Lawrason, 1941. The story of clinical pulmonary tuberculosis.

Cave, A.J.E., Demonstrator, Arnott, 1939. The evidence for the incidence of tuberculosis in ancient Egypt. Br. J. Tuberc. 33 (3), 142–152.

Chouhan, Vikash, Singh, Sanjay Kumar, Khamparia, Aditya, Gupta, Deepak, Tiwari, Prayag, Moreira, Catarina, Damaševičius, Robertas, De Albuquerque, Victor Hugo C, 2020a. A novel transfer learning based approach for pneumonia detection in chest X-rays images. Appl. Sci. 10 (2), 559.

Chouhan, Vikash, Singh, Sanjay Kumar, Khamparia, Aditya, Gupta, Deepak, Tiwari, Prayag, Moreira, Catarina, Damaševičius, Robertas, de Albuquerque, Victor Hugo C., 2020b. A novel transfer learning based approach for pneumonia detection in chest X-ray images. Appl. Sci. 10 (2).

Degirmenci, Murside, Surucu, Murat, Perc, Matjaž, Isler, Yalcin, 2025. Convolutional neural networks can diagnose schizophrenia. J. Comput. Sci. 102634.

Delgado, Rosario, 2022. A semi-hard voting combiner scheme to ensemble multi-class probabilistic classifiers. Appl. Intell. 52 (4), 3653–3677.

Dey, Subhrajit, Roychoudhury, Rajarshi, Malakar, Samir, Sarkar, Ram, 2022. An optimized fuzzy ensemble of convolutional neural networks for detecting tuberculosis from chest X-ray images. Appl. Soft Comput. 114, 108094.

Duong, Linh T, Le, Nhi H, Tran, Toan B, Ngo, Vuong M, Nguyen, Phuong T, 2021. Detection of tuberculosis from chest X-ray images: Boosting the performance with vision transformer and transfer learning. Expert Syst. Appl. 184, 115519.

Fati, Suliman Mohamed, Senan, Ebrahim Mohammed, ElHakim, Narmine, 2022. Deep and hybrid learning technique for early detection of tuberculosis based on X-ray images using feature fusion. Appl. Sci. 12 (14), 7092.

Ford, Nathan, Matteelli, Alberto, Shubber, Zara, Hermans, Sabine, Meintjes, Graeme, Grinsztejn, Beatriz, Waldrop, Greer, Kranzer, Katharina, Doherty, Meg, Getahun, Haileyesus, 2016. TB as a cause of hospitalization and in-hospital mortality among people living with HIV worldwide: a systematic review and meta-analysis. Afr. J. Reprod. Gynaecol. Endosc. 19 (1).

Godreuil, Sylvain, Tazi, Loubna, Bañuls, Anne-Laure, 2007. Pulmonary tuberculosis and mycobacterium tuberculosis: modern molecular epidemiology and perspectives. Encycl. Infect. Dis.: Mod. Methodol. 1–29.

Goyal, Bhawna, Dogra, Ayush, Agrawal, Sunil, Sohi, Balwinder Singh, Sharma, Apoorav, 2020. Image denoising review: From classical to state-of-the-art approaches. Inf. Fusion 55, 220–244.

Guo, Liu Jian, 1991. Balance contrast enhancement technique and its application in image colour composition. Remote. Sens. 12 (10), 2133–2151.

Habib, Al-Zadid Sultan Bin, Tasnim, Tanpia, 2020. An ensemble hard voting model for cardiovascular disease prediction. In: 2020 2nd International Conference on Sustainable Technologies for Industry 4.0. STI, IEEE, pp. 1–6.

İlikhan, Sevil Uygun, Özer, Mahmut, Perc, Matjaz, Tanberkan, Hande, Ayhan, Yavuz, 2025. Complementary use of artificial intelligence in healthcare. Med. J. West. Black Sea 9 (1), 7–17.

Iqbal, Ahmed, Usman, Muhammad, Ahmed, Zohair, 2023. Tuberculosis chest X-ray detection using CNN-based hybrid segmentation and classification approach. Biomed. Signal Process. Control. 84, 104667.

Jaeger, Stefan, Candemir, Sema, Antani, Sameer, Wáng, Yì-Xiáng J, Lu, Pu-Xuan, Thoma, George, 2014. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant. Imaging Med. Surg. 4 (6), 475.

Journal Of Healthcare Engineering, 2023. Retracted: A novel and robust approach to detect tuberculosis using transfer learning. J. Heal. Eng. 2023, 9810410.

Karlos, Stamatis, Kostopoulos, Georgios, Kotsiantis, Sotiris, 2020. A soft-voting ensemble based co-training scheme using static selection for binary classification problems. Algorithms 13 (1), 26.

Khan, Salman, Naseer, Muzammal, Hayat, Munawar, Zamir, Syed Waqas, Khan, Fahad Shahbaz, Shah, Mubarak, 2022. Transformers in vision: A survey. ACM Comput. Surv. 54 (10s), 1–41.

Kingsbury, Nick, Magarey, Julian, 1998. Wavelet transforms in image processing.

Kong, Heesan, Kim, Donghee, Kim, Kwangsu, 2023. Enriching chest radiography representations: Self-supervised learning with a recalibrating and importance scaling. IEEE Access.

Koo, Kyung-Mo, Cha, Eui-Young, 2017. Image recognition performance enhancements using image normalization. Human-Centric Comput. Inf. Sci. 7, 1–11.

Lin, Chou-Han, Lin, Chou-Jui, Kuo, Yao-Wen, Wang, Jann-Yuan, Hsu, Chia-Lin, Chen, Jong-Min, Cheng, Wern-Cherng, Lee, Li-Na, 2014. Tuberculosis mortality: patient characteristics and causes. BMC Infect. Dis. 14, 1–8.

Liu, Ze, Hu, Han, Lin, Yutong, Yao, Zhuliang, Xie, Zhenda, Wei, Yixuan, Ning, Jia, Cao, Yue, Zhang, Zheng, Dong, Li, et al., 2022. Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12009–12019.

Liu, Ze, Lin, Yutong, Cao, Yue, Hu, Han, Wei, Yixuan, Zhang, Zheng, Lin, Stephen, Guo, Baining, 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Nafisah, Saad I., Muhammad, Ghulam, 2024. Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence. Neural Comput. Appl. 36 (1), 111–131.

Narayanan, Srikanth, Balamurugan, NM, Maithili, K, Palas, P Bini, 2022. Leveraging machine learning methods for multiple disease prediction using python ML libraries and flask API. In: 2022 International Conference on Applied Artificial Intelligence and Computing. ICAAIC, IEEE, pp. 694–701.

Pei, Soo-Chang, Lin, Chao-Nan, 1995. Image normalization for pattern recognition. Image Vis. Comput. 13 (10), 711–723.

Perc, Matjaž, Ozer, Mahmut, Hojnik, Janja, 2019. Social and juristic challenges of artificial intelligence. Palgrave Commun. 5 (1).

Pizer, Stephen M, Amburn, E Philip, Austin, John D, Cromartie, Robert, Geselowitz, Ari, Greer, Trey, ter Haar Romeny, Bart, Zimmerman, John B, Zuiderveld, Karel, 1987. Adaptive histogram equalization and its variations. Comput. Vis. Graph. Image Process. 39 (3), 355–368.

Rahman, Tawsifur, Chowdhury, Muhammad E.H., Khandakar, Amith, Islam, Khandaker R., Islam, Khandaker F., Mahbub, Zaid B., Kadir, Muhammad A., Kashem, Saad, 2020a. Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. Appl. Sci. 10 (9).

Rahman, Tawsifur, Khandakar, Amith, Kadir, Muhammad Abdul, Islam, Khandaker Rejaul, Islam, Khandakar F., Mazhar, Rashid, Hamid, Tahir, Islam, Mohammad Tariqul, Kashem, Saad, Mahbub, Zaid Bin, Ayari, Mohamed Arselene, Chowdhury, Muhammad E.H., 2020b. Reliable tuberculosis detection using chest X-Ray with deep learning, segmentation and visualization. IEEE Access 8, 191586–191601.

Rahman, Shanto, Rahman, Md Mostafijur, Abdullah-Al-Wadud, Mohammad, Al-Quaderi, Golam Dastegir, Shoyaib, Mohammad, 2016. An adaptive gamma correction for image enhancement. EURASIP J. Image Video Process. 2016, 1–13.

Ranftl, René, Bochkovskiy, Alexey, Koltun, Vladlen, 2021. Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188.

Ruby, Usha, Yendapalli, Vamsidhar, 2020. Binary cross entropy with deep learning technique for image classification. Int. J. Adv. Trends Comput. Sci. Eng 9 (10).

Sathitratanacheewin, Seelwan, Sunanta, Panasun, Pongpirul, Krit, 2020. Deep learning for automated classification of tuberculosis-related chest X-Ray: dataset distribution shift limits diagnostic performance generalizability. Heliyon 6 (8).

Singh, Sharandeep, Wani, Niyaz Ahmad, Kumar, Ravinder, Bedi, Jatin, 2025. DiaXplain: A transparent and interpretable artificial intelligence approach for type-2 diabetes diagnosis through deep learning. Comput. Electr. Eng. 126, 110470.

Stephen, Okeke, Sain, Mangal, Maduh, Uchenna Joseph, Jeong, Do-Un, et al., 2019. An efficient deep learning approach to pneumonia classification in healthcare. J. Heal. Eng. 2019.

Thakur, Arastu, Gupta, Muskan, Sinha, Deepak Kumar, Mishra, Kritika Kumari, Venkatesan, Vinoth Kumar, Guluwadi, Suresh, 2024. Transformative breast cancer diagnosis using CNNs with optimized ReduceLROnPlateau and early stopping enhancements. Int. J. Comput. Intell. Syst. 17 (1), 14.

Touvron, Hugo, Cord, Matthieu, Douze, Matthijs, Massa, Francisco, Sablayrolles, Alexandre, Jégou, Hervé, 2021. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. PMLR, pp. 10347–10357.

Wani, Niyaz Ahmad, Kumar, Ravinder, Bedi, Jatin, 2024a. DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. Comput. Methods Programs Biomed. 243, 107879.

Wani, Niyaz Ahmad, Kumar, Ravinder, Bedi, Jatin, 2024b. Harnessing fusion modeling for enhanced breast cancer classification through interpretable artificial intelligence and in-depth explanations. Eng. Appl. Artif. Intell. 136, 108939.

Wani, Niyaz Ahmad, Kumar, Ravinder, Bedi, Jatin, Rida, Imad, et al., 2024c. Explainable AI-driven IoMT fusion: Unravelling techniques, opportunities, and challenges with explainable AI in healthcare. Inf. Fusion 110, 102472.

Wong, Alexander, Lee, James Ren Hou, Rahmat-Khah, Hadi, Sabri, Ali, Alaref, Amer, Liu, Haiyue, 2022. TB-net: a tailored, self-attention deep convolutional neural network design for detection of tuberculosis cases from chest X-ray images. Front. Artif. Intell. 5, 827299.

Xu, Mingle, Yoon, Sook, Fuentes, Alvaro, Park, Dong Sun, 2023. A comprehensive survey of image augmentation techniques for deep learning. Pattern Recognit. 137, 109347.

Zhang, Zijun, 2018. Improved adam optimizer for deep neural networks. In: 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). IEEE, pp. 1–2.

Zimmerman, Michael R., 1979. Pulmonary and osseous tuberculosis in an Egyptian mummy.. Bull. N Y Acad. Med. 55 (6), 604.

Zou, Fangyu, Shen, Li, Jie, Zequn, Zhang, Weizhong, Liu, Wei, 2019. A sufficient condition for convergences of adam and rmsprop. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11127–11135.

Zuiderveld, Karel, 1994. Contrast limited adaptive histogram equalization. In: Graphics Gems IV. Graphics gems IV, pp. 474–485.