A Multimodal Approach to Detect the Tuberculosis Using Deep Learning Technique

Saranya K. G.[1], Akalya A.[2]

[1]Associate Professor, [2]Student, Department of Computer Science, PSG College of Technology, Coimbatore, Tamil Nadu, India

E-mail: [1]24mz31@psgtech.ac.in

Abstract

The biggest challenges faced byTB (tuberculosis) patients are difficulty in managing TB particularly for those living in low socioeconomic areas and especially in rural communities. Chest X-rays are valuable in diagnosing TB because they are inexpensive and non-invasive for obtaining the data needed. There are many limitations to existing applications using deep learning algorithms to detect TB, including a limited number of images available for use (i.e., an X-ray or radiograph image of the chest), and these applications may require significant computational power. Additionally, many applications do not have a variety of characteristics to support the for diagnosis of TB. Existing deep learning TB detection methods do not provide interpretable output because they produce high rates of false negatives, especially when the disease is clinically unclear. This paper proposes a hybrid multimodal ensemble approach that uses both X-ray images and a systematic set of clinical symptoms to improve the accuracy of clinical-quality TB detection. By combining the advantages of the DenseNet and MobileNet architectures, this model is able to create additional radiographic features. It also uses graph neural networks to learn additional features related to the clinical symptoms, allowing the model to learn the contextual and clinical connections within the input data. The multimodal representations will be used to create overall diagnostic predictions using an attention-based ensemble to combine these representations. The overall prediction accuracy for the multimodal approach is 96% with precision/recall statistics nearly the same resulting in more robust and reduced false positive/false negative predictions compared to predictions from unimodal, image-based approaches. Examples of XAI techniques may be implemented to support

DOI: https://doi.org/10.36548/jscp.2025.4.004
Received: 04.12.2025, received in revised form: 03.01.2026, accepted: 20.01.2026, published: 31.01.2026

transparency and build trust in clinical decisions using visualizations (e.g. Grad-CAM) to illustrate areas of the lung that contribute to the diagnosis of the disease. The Unified Multimodal TB Screening method will provide useful, interpretable and computationally efficient solutions to help automate TB diagnoses with the technologies needed to implement screening in real world, resource-constrained settings.

Keywords: Tuberculosis Detection, Deep Learning, Hybrid Ensemble Model, DenseNet, MobileNet, Graph Neural Network (GNN), Explainable AI (XAI), Grad-CAM, Chest X-ray, Medical Image Analysis.

1. Introduction

Mycobacterium tuberculosis is the common cause of tuberculosis (TB), a contagious bacterial infection that primarily affects the lungs but can also spread to the kidneys, spine and brain. Despite being preventable and treatable, tuberculosis (TB) is still a major global public health concern, especially in lower and middle-income nations where delayed diagnosis increases TB transmission and raises death rates. Millions of people are affected by new cases of tuberculosis every year, according to the global health report, demonstrating the need for a precise and prompt diagnostic system.

The non-invasive and relatively low cost of a chest X-ray is significant. When evaluating TB disease processes, clinicians depend on the radiographic appearance of lung transparency, valve damage, and consolidation (areas of dense tissue) on chest X-rays. Even though skilled radiologists have time to manually check the chest X-rays (CXRs), various factors such as the CXR's quality and the disease's stage can complicate the process. In rural areas without a qualified radiologist, timely and accurate diagnosis is further constrained. Convolutional neural networks (CNNs) and deep learning among other types of artificial intelligence, have modified computer-aided diagnostic (CAD) systems to identify tuberculosis.

CNN-based methods better in extracting discriminative features from chest X-ray images, increasing TB diagnosis accuracy while minimizing human labor. However, the majority of existing deep-learning models operate in a single-mode manner, utilizing only imaging data, without considering important clinical indicators that doctors commonly use to diagnose the disease. These indicators include fever, night sweats, weight loss, chronic cough, and chest pain. In real-world clinical settings, radiologic results are rarely used in isolation to diagnose tuberculosis. To create a reliable diagnosis, clinicians depend on the integration of radiographic findings with patient symptoms and demographics.

When radiographic results are unclear and such clinical contextual factors are absent, the low reliability of all-image-based models will inhibit their clinical acceptance, despite reports of high classification rates. One of the two main problems with implementing deep learning techniques in the medical setting is the models' interpretability. The majority of effective convolutional neural networks (CNNs) operate as "black boxes" used to make decisions. The adoption of these models will be delayed by the lack of trust caused by inadequate privacy. Because they enable feature-based and visual explanations of model predictions, XAI techniques are important.

The purpose of the present study is to propose an alternative multimodal deep learning model that utilizes data from structured clinical symptoms and the visual image of the X-ray to accurately identify the diagnosis of tuberculosis. Dense Network and Mobile Network architectures, which are accurate and highlight significant features in the radiograph, were utilized in the visual section of the model. In addition, the features of Graph Neural Networks and the attention-based process are employed to effectively represent the characteristics of the patients, as the data comprises the symptoms and demographic attributes of the patients. Furthermore, the proposed model utilizes the attention-based approach to effectively combine all the multimodal features, as it is imperative to make a diagnosis of the patient. This addresses the gap between image analysis and the reasoning process, as seen in the Grad-CAM visualization process.

2. Literature Survey

Initial studies on automated detection of tuberculosis (TB) utilized conventional machine learning methods on chest X-ray (CXR) images. Such methods usually included hand-designed techniques for extracting features, such as texture descriptors, shape-based features, edge-based features, and gradient-based representations, followed by classical classifiers such as Support Vector Machines (SVM), Random Forests (RF), and k-Nearest Neighbors (k-NN). Naing and Htike [19] presented an overall classification of the first computer-aided diagnosis systems and focused on the significance of integrating texture and structural features in the identification of TB. Although these methods showed feasible performance on controlled datasets, they heavily relied on manually constructed features, greatly neglecting their applicability to the diverse current imaging conditions and groups. Another stochastic learning-based artificial neural network model to detect TB, suggested by Urooj et al. [8], aimed to ensure lower computational complexity and training time compared with deep CNNs.

The model is faster and uses fewer resources but exhibited poor sensitivity to the delicate pathology of lung opacities and focal lesions, despite their subtlety. In a similar vein, Mehrotra et al. [9] examined the hybridization of traditional machine learning classifiers with CNN-based features and enhanced classification accuracy at the cost of increased system complexity and inference time. In general, conventional machine learning methods were reliable in clinical settings; however, their limited ability to represent features restricted their diagnostic utility in practice.

Convolutional Neural Networks (CNNs) have become one of the most important methods in deep learning, resulting in the extensive use of artificial intelligence (AI) in detecting TB using chest X-ray images. CNNs can automatically learn hierarchical feature representations, making them highly appropriate for detecting complex radiographic patterns related to pulmonary diseases. Huy and Lin [5] designed an improved structure of DenseNet with a Convolutional Block Attention Module (CBAM), which enhanced the model's ability to focus on clinically important areas of the lungs. The proposed CBAM-DenseNet model demonstrated a high level of classification accuracy and better localization of lesions; however, it has high input image requirements for high-resolution images and demands powerful computers, making it unsuitable for low-resource settings. Sharma et al. [7] presented CNN-based TB detection models that included a visualization method with activation maps to highlight regions of infectivity and enhance readability.

Although the mentioned visualization techniques were found to increase clinical knowledge, the models were heavily dependent on large annotated data sets and extensive preprocessing. Munadi et al. [10] used image enhancement methods like contrast stretching and edge sharpening before the training of the CNN, which enhanced the visibility of features but also raised the chances of overfitting and computational cost. Xu and Yuan [15] proposed four coordinate-attention CNNs to learn spatial correlations across the CXR images, which provided better localization accuracy but at the expense of a more complex model. Even though the CNN-based TB detection models demonstrate a high level of performance, they are mostly unimodal, as they use imaging data only. Such a limitation compromises diagnostic strength, especially in cases where the radiographic appearance is obscure or atypical, and leads to an increased risk of false negatives when there is a lack of focus on clinical context.

To address the drawbacks of single-model architecture, multiple works have examined ensemble and hybrid deep learning approaches to TB detection. Ensemble techniques are based on predictions that integrate several models or represent variations of features to enhance generalization and decrease variance. Malik et al. [14] suggested a fusion model that combines handcrafted features and deep CNN embeddings, showing enhanced resilience to differences in image acquisition conditions. In their paper, Meraj et al. [16] presented a modality-specific ensemble that utilized various model representations, such as full CXR images, segmented lung regions, and edge-enhanced images, which demonstrated higher cross-dataset generalization. There have also been studies applying Bayesian deep learning in the quantification of predictive uncertainty. In Bairagi et al. [17], Bayesian CNNs using Monte Carlo dropout were employed to predict uncertainty in TB, offering diagnostic outputs with confidence. Although uncertainty modeling enhanced reliability, it increased inference time and computational cost.

Lung segmentation was integrated into classification by Rahman et al. [11] to concentrate learning on lung parts, thus enhancing detection accuracy. Yadav and Jadhav [12] proved that segmentation-based preprocessing generates significant improvements in the classification of pulmonary diseases by alleviating background noise. Despite the general improvement in the direction of ensemble and hybrid models over single CNN architecture, they have higher computational costs and the disadvantage of being difficult to explain, thus making their applicability to real-world healthcare contexts, especially under resource-constrained conditions, more challenging.

Recent studies have focused on designing lightweight deep learning architectures that can be deployed on embedded and Internet of Things (IoT) devices. An optimized CNN that can recognize respiratory diseases, such as TB, on Raspberry Pi was suggested by Bosale and Patnaik [13] as an IoT-deployable device. These models can be used to support low-cost and remote healthcare environments that utilize AI-assisted diagnosis, as demonstrated by these lightweight models. Nonetheless, these systems also have limitations, including limited memory, lower model capacity, and difficulties in retraining and maintenance, which can adversely impact the long-term reliability of diagnostics.

2.1 Research Gaps and Objectives

Even though previous researchers report high accuracy in TB classification with deep learning, the majority of existing systems are dedicated to optimizing image-level results, and the issues of clinical reliability and interpretability are not addressed. Clinically, the diagnosis of TB is achieved through the process of incorporating TB clinical manifestations with regard to radiographic characteristics, patient symptoms, and demographics. Single-mode systems used in automated frameworks have widened the gap between the AI approach to making predictions and real-world diagnostic reasoning.

Moreover, existing ensemble methods enhance accuracy by sacrificing computational complexity; thus, they cannot be used in environments with limited resources. Additionally, limited focus has been placed on minimizing the risk of false negatives, which is an important aspect of TB screening, as the outcomes of undetected diagnoses are grave. The main goals of the research are:

- To achieve a multimodal deep learning system that combines chest X-ray images with organized clinical symptom records to detect TB.
- To utilize DenseNet and MobileNet architectures to simultaneously exploit each other to extract radiographic features and benefit from their respective levels of efficiency.
- To simulate inter-symptom relationships and individual patient clinical patterns based on Graph Neural Networks (GNNs) and attention-based mechanisms.
- To develop an attention-based fusion and ensemble strategy that enhances robustness and manages computational costs.
- To minimize the probability of false negatives and improve diagnostic accuracy in relation to unimodal image-based systems.

3. Existing System

The existing tuberculosis detection mechanism using the proposed study framework introduces the term CBAMWDNet, which stands for Convolutional Block Attention Module combined with Wide DenseNet. This deep learning model aims to detect tuberculosis with the help of automated techniques using chest X-ray images. By combining the Wide DenseNet model and the Convolutional Block Attention Model, the system aims to provide a more specific feature representation using the attention mechanism on the lung regions.

The channel attention module focuses on the informative features of the channel by using a combination of global average pooling, maximum pooling, and their shared multilayer perceptron. The spatial attention module mainly focuses on the spatial location of informative lung features using a spatial attention map through a series of convolution operations. The sequential attention mechanism employed by this model enables it to ignore irrelevant background information while maximizing relevant disease data about the lung. The Wide DenseNet component of this model can be viewed as an extension of the conventional DenseNet model in that it introduces width into convolution operations.

All the layers are connected by feature maps from the prior layers, making it efficient in terms of feature map usage. This high level of connectivity allows the model to detect low-level textures and high-level structural abnormalities of tuberculosis, including cavities and consolidations. CBAMWDNet has been applied using publicly available databases based on the Montgomery County, Shenzhen Hospital, and Qatar University Chest X-ray databases, which comprised 5,000 images, of which 3,906 were normal and 1,094 were TB-positive. The model achieved an accuracy of 98.8% in the classification stage of the results, indicating that it performed well when applied at the image level. Although the model is quite accurate, it comes with some limitations and disadvantages regarding its practical application in real-life scenarios.

Firstly, the data used is based solely on images, while clinical symptoms and demographics, which are essential components of TB testing procedures, are not considered. Secondly, models based on images have a higher rate of false negatives, especially when unusual patterns are involved. Thirdly, Wide DenseNet, which is based on attention, is computationally expensive, which might hinder its practical application in resource-intensive healthcare.

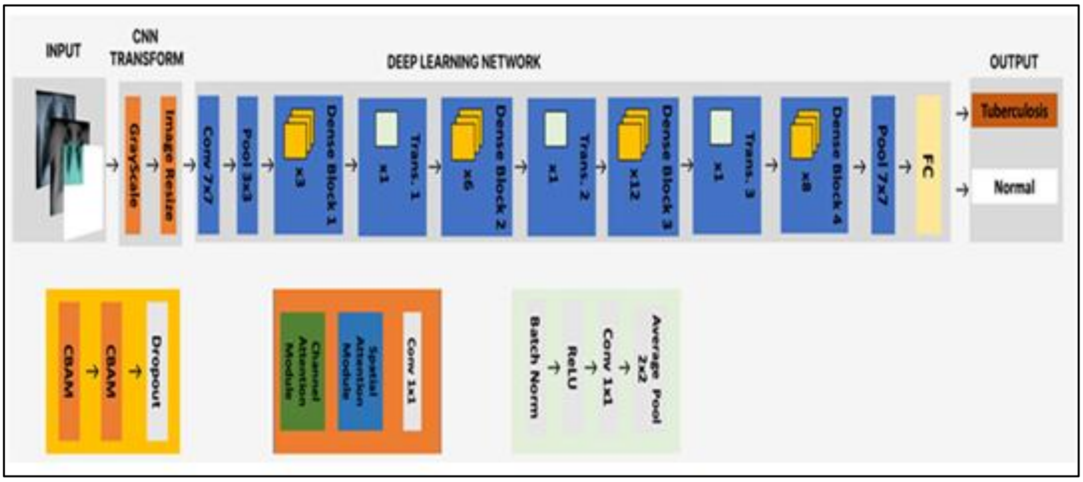

Figure 1. Architecture of the CBAMWDNet [5]

The CBAMWDNet model architecture is shown in Figure 1, which illustrates how the convolutional layers have been combined with the CBAM attention mechanism as a means of refining feature extraction and therefore classification performance.

A. Convolutional Block Attention Module

The CBAM method aims to improve convolutional neural networks by highlighting the most important features of an image. It consists of two submodules used in parallel; Spatial and Channel. Channel Attention evaluates the importance of feature maps using global average and max pooling to create a common network that generates the attention value. Spatial Attention uses channel pooling and convolution to determine the position of relevant image features resulting in a spatial attention map. This method transmits input to the Channel Attention module to increase the relevant channels. The processed output is analyzed by the Spatial Attention module to identify regions relevant to an image. Integration helps the model focus on lungs in chest X-rays for diagnosis, minimizing background noise and improving detection accuracy and interpretability.

B. Wide DenseNet

Wide DenseNet, which varies from the standard DenseNet network includes both the depth and breadth of the network layers. Each level in this structure allows input from all previous layers, increasing the reuse of features or data flow throughout the network. A convolutional filter containing several convolutional filters and pooling procedures is applied to the input chest X-ray image followed by four Dense Blocks. The model's high level of connectivity allows access to both low and high-level lung patterns, and it is therefore widely employed in the detection of hidden tuberculosis infections and structural variations.

C. Dense Blocks and Feature Propagation

Dense blocks are required to provide effective transmission in the network. The output and input are transmitted to each level of the same block in the layer. This architecture enables effective reuse of acquired features, increases gradient flow and handles the issue of vanishing gradients. In this situation, the details collected in previous layers will be combined with the creation of new features for better image representation. This continuous propagation helps the model to capture higher features and complex lung textures to recognize early TB symptoms in chest X-ray images.

4. Proposed System

The proposed design incorporates chest X-ray images and clinical data to detect tuberculosis (TB) accurately. Preprocessing of images and patient data is performed through augmentation and encoding. DenseNet and Transformer methods are applied to extract features, which are processed using a CNN model. An attention-based fusion integrates visual and clinical attributes to increase the accuracy of predictions. Accuracy and sensitivity measures are used to assess the performance of the model, providing the reliability in disease predictions and enhancing diagnostic assistance in the detection of TB.

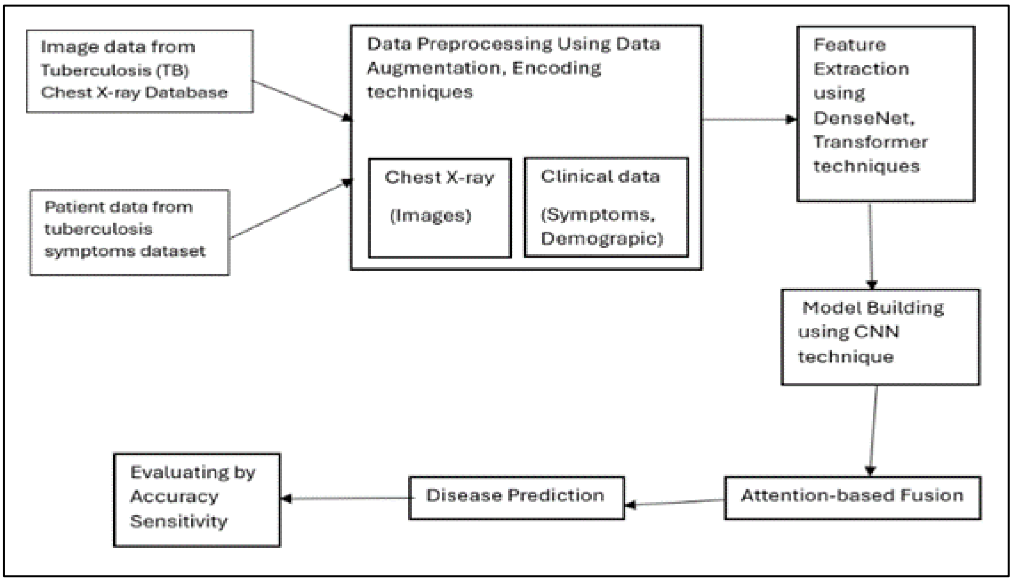
Image data from Tuberculosis (TB) Chest X-ray Database
Patient data from tuberculosis symptoms dataset
Data Preprocessing Using Data Augmentation, Encoding techniques
Chest X-ray (Images)
Clinical data (Symptoms, Demograpic)
Feature Extraction using DenseNet, Transformer techniques
Model Building using CNN technique
Attention-based Fusion
Disease Prediction
Evaluating by Accuracy Sensitivity

Figure 2. Proposed System of Architecture

Figure 2 illustrates the architecture of the proposed system and demonstrates how it is based on the successive stages of data preprocessing and feature extraction, model training and eventual classification to accurately identify the disease.

A. DenseNet

DenseNet is a recent deep learning network applied in image classification. Unlike classical convolutional neural networks, DenseNet connects all the layers with all the prior layers and unlike other networks, the learnings made by the early layers can be reused by the subsequent layers. It has a high degree of connection which enhances gradient flow, reduces the loss of gradient and enables more efficient propagation of features.

DenseNet has recorded exceptional performance in medical image analysis particularly in the analysis of chest X-rays for disease detection. DenseNet was employed in this project to extract complex information from the TB chest X-ray images which include lung abnormalities and lesion patterns that are significant for the appropriate diagnosis of tuberculosis. It has a high learning ability for hierarchical characteristics, making it very specific and sensitive to detect TB.

B. MobileNet

MobileNet is a small-scale convolutional neural network that focuses on computational efficiency and speed, making it suitable for use in mobile or low-resource settings. MobileNet uses depthwise separable convolutions to significantly reduce the number of parameters and computation compared to standard CNNs while maintaining high accuracy. In this project, MobileNet was used to perform real-time TB detection enabling it to make inferences quickly, regardless of the powerful GPUs. This renders it specifically applicable for implementation in remote or resource-limited clinical environments, in where urgent TB screening is needed.

C. Graph Neural Networks (GNNs)

Graph Neural Networks (GNNs) are a type of specialized neural network that takes graph-like data as its input, thus allowing the modeling of the relationships between the objects represented as vertices in the graph. In a medical setting, the data on a patient (e.g. age, gender, medical history, symptoms, etc.) can be described in a graph where the nodes represent the information on the patient and the edges represent the correlations or interactions.

GNNs are trained not only on the properties of individual nodes and their topology but also on the relationships between nodes, making them efficient in predicting the risk and progression of disease. In the current paper, GNNs were observed to process clinical data of TB patients, a part of which involved the interrelation between symptoms and demographic variables, allowing the model to be predictive.

D. Transformers

Transformers are deep learning models that were created to process natural language texts; however, the attention mechanism employed by such models can also solve structured and tabular data. The self-attention mechanism enables the model to prioritize the relevance of every feature, which, in turn, allows it to concentrate on the most useful clinical indicators.

Transformers were used in this project to identify clinical attributes of TB related critical features like persistent cough, long lasting fever, weight loss and night sweats. By focusing on these major traits, the model will be able to make better predictions and enhance the overall accuracy of the diagnosis.

E. Data Preprocessing

The standardization of all X-ray images of the chest under the scan to a range of [0,1] before the training process is essential to achieve the desired uniformity in the intensity scale and faster convergence. Symptom data that are categorical, such as fever, cough and fatigue are converted into binary or numerical formats to fit into the model. To solve the problem of class imbalance between TB-positive and normal samples, data augmentation and oversampling are used to boost generalization and eliminate the model's bias towards majority classes.

Figure 3 presents a bar graph indicating the distribution of the samples across the various classes selected in data balancing with the objective of representing the classes equally and preventing biases in the model during training.

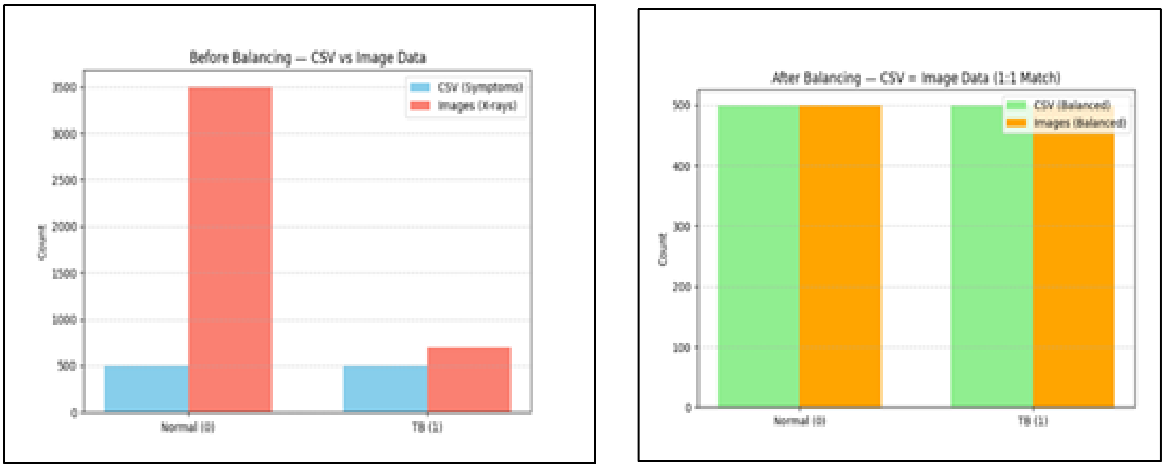

Figure 3. Bar Graph of Data Balancing

F. Importing Necessary Libraries

This section shows how vital Python libraries required to perform the multimodal tuberculosis detection model were imported. The design, training, and testing of deep learning models are performed using TensorFlow and Keras, while the numerical operations and data structures are handled using NumPy and Pandas. Scikit-learn assists in preprocessing data and feature scaling, and Keras can be trained more effectively as with callbacks, which can be utilized to implement early stopping and checkpoint models to prevent overfitting.

- **Pandas**

A library for data analysis that is supported by helpful data structures such as DataFrames which facilitate effective processing and manipulation of structured data. Most of its applications include reading, cleaning, and sorting clinical and image metadata.

- **numPy**

A fundamental scientific computing library of numerical Python. It facilitates mathematical operations, matrices, and effective management of multi-dimensional arrays that are involved in deep learning applications.

- **TensorFlow**

An effective open-source deep learning system created by Google. It offers systems to create, train and deploy machine learning and deep learning models effectively, such as CNNs and Transformers.

- **tensorflow.keras.layers & tensorflow.keras.models**

Layers: In Dense, Conv2D and Dropout, one could specify the structural specifics of experiments using layer models. This is applied to create, compile, and manage the framework of the neural network.

- **tensorflow.keras.preprocessing.image**

Image loading, augmentation and preprocessing tools are offered. It assists in preprocessing image datasets to be used in the training of models through resizing, normalization, and batch creation.

- **sklearn.preprocessing.StandardScaler**

This is a part of the Scikit-learn library that is used to standardize numerical data by scaling features to zero mean and unit variance. This guarantees stable and faster convergence of the model.

- **tensorflow.keras.utils.to_categorical**

It converts the class labels (integers) into one-hot coded vectors which is essential in multi-class classification tasks when using deep learning.

G. Building the Multimodal

The multimodal tuberculosis detection system provided comprises a variety of deep learning models (DenseNet, MobileNet, Transformer, and Graph Neural Network (GNN)) to handle patient clinical information as well as the images of the chest X-ray. DenseNet is a densely connected convolutional network that enhances the transportation of features by linking all layers to the following layer. This enables the model to reuse the learned features.

This has an advantage in retrieving both low-level data such as edges and textures and high-level data such as the boundaries of the lungs, nodules, and lesions in the chest X-rays. MobileNet is an optimized convolutional architecture with depthwise separable convolutions and is effectively applied to the analysis of chest X-rays in real time, with high precision and low computational expenses, making it suitable for cases with limited hardware. At the same time, the Transformer model operates using structured clinical data, e.g., age, gender, symptoms, and medical history of a patient.

The mechanistic features of the self-attention mechanism highlight its most important clinically significant properties, such as persistent cough, prolonged fever, and weight loss, as well as the intricate interactions between a large number of attributes. Additionally, the GNN considers patient attributes as connected nodes in a graph, and the attributes and characteristics of a node and its relationships are learned. This enables the system to understand non-linear relationships, such as combinations of symptoms that are connected with the diagnosis of TB.

The branches of DenseNet and MobileNet (image-based and clinical data, respectively) are then combined with the help of an attention-based fusion layer to balance the importance of visual and clinical data. In an ensemble approach, optimized weighting is again employed in conjunction with organized noise control to further consolidate the predictions of all the models, achieving higher stability, reduced overfitting, and improved diagnostic tolerance. The interplay of these factors forms an effective multimodal scheme, which can accurately and reliably detect TB and be applied to actual clinical problems.

Algorithm

```
Input: Chest X-ray image dataset and clinical data (symptoms, age, gender, etc.) from
Open-i
Output: Predicted tuberculosis class (TB / Non-TB)

Step 1: Import required libraries (TensorFlow, Keras, NumPy, Pandas, Scikit-learn,
etc.).
Step 2: Register custom Keras layers to fix Transformer serialization (GetItem()
function).
Step 3: Define file paths for models and test datasets.
Step 4: Load the test dataset and preprocess:
    • Resize chest X-ray images to 224×224.
    • Normalize pixel values between 0–1.
    • Standardize clinical attributes using StandardScaler().
Step 5: Load pre-trained models DenseNet, MobileNet, Transformer and GNN from
storage paths.
Step 6: For each model:
    • Perform prediction on the test set.
    • Store predicted probability outputs.
Step 7: Initialize ensemble parameters:
    • Weights for models (DenseNet=0.3, MobileNet=0.3, Transformer=0.2, GNN=0.2)
    • Noise factor = 0.15, Temperature = 1.8, Dropout = 0.25, Label flip rate = 0.03
Step 8: For each test sample:
    a. Randomly drop one model prediction (dropout).
```

```
        b. Apply temperature scaling to soften probabilities.
        c. Add structured noise for uncertainty.
        d. Compute weighted sum of model outputs.
    Step 9: Apply label smoothing and occasional label flipping to simulate uncertainty.
    Step 10: Compute final ensemble prediction $y_{pred} = arg\ max$ (avg_pred_prob).

      Step 11: The model's effectiveness is measured using Accuracy, Precision, Recall, F1-score, along with the Confusion Matrix.
Step 12: Plot accuracy comparison graph between individual models and the multimodal ensemble
```

H. Fusion and Ensemble Integration

Outputs of both methods, image-based (DenseNet and MobileNet) and clinical-based (GNN and Transformer), are combined with an attention-based fusion layer. This fusion system is dynamic because it balances the input of the visual and clinical characteristics, independent of the final decision. In order to ensure strength and generalization further, a weighted ensemble approach is employed. All four models make predictions that are consolidated with the help of optimized weights in accordance with model reliability.

During ensemble optimization, controlled regularization methods such as uncertainty management and temperature scaling are used to stabilize predictions and improve overfitting. These mechanisms are not applied in the process of final clinical inference but in model training and ensemble calibration. The multimodal ensemble structure allows the system to provide consistent and trusted predictions with computational efficiency and explainability.

Algorithm

```
Input: Predicted probabilities from DenseNet, MobileNet, Transformer and GNN
Output: Final optimized ensemble prediction

  Step 1: Assign model weights based on performance reliability.
  Step 2: Initialize ensemble prediction vector is 0.
  Step 3: For each test sample:
      a. For each model prediction $p_i$:
          • Apply dropout probability $p_{drop}$to randomly skip one model.
      • Perform temperature scaling: $p_i = p_i^{(1/T)} / \sum p_i^{(1/T)}$.
          • Add structured uniform noise $n \in [-\eta, \eta]$.
```

```
            • Clip and normalize probability distribution.
            • Aggregate weighted probabilities:
                $avg\_pred\_prob += w_i \times p_i$.
        b. Normalize final probability sum.
Step 4: Apply label smoothing:
        $p' = (1 - \epsilon)p + \frac{\epsilon}{2}$.
Step 5: Randomly flip a small fraction of labels (rate = 0.03) to reduce overfitting.
Step 6: Select class with maximum probability as final prediction.
Step 7: Evaluate ensemble using accuracy and confusion matrix
```

5. Experimental Results

The multimodal TB detection model, which includes DenseNet, MobileNet, GNNs, and Transformers, has been shown to be accurate on data from chest X-rays and clinical data. DenseNet was used to produce significant image features, and MobileNet was used to generate images with limited storage. In Transformers, the connection between patient details was investigated using GNNs, which highlighted clinical factors like age, symptoms, and gender. The integrated technique has been found to be more sensitive, accurate, and has a higher F1-score than individual models. Early loss of collaboration and learning schedule decreased overfitting while increasing durability. Overall, it is possible to determine that the presented model is reliable, efficient, and adaptable for use in real life for both TB screening and decision support in situations where healthcare resources are limited.

A. Dataset Description

The proposed multimodal TB detection system is evaluated using a combination of publicly accessible chest X-ray datasets for training and validation, along with a separate real-world clinical dataset for testing. Chest X-ray scans have been collected from three different datasets: Montgomery County, Shenzhen Hospital, and Qatar University. These datasets contain 5,000 chest X-ray images of tuberculosis-positive and normal individuals, with duplicate samples processed and removed during the preparation step. These images are used to train image-based deep learning models such as DenseNet and MobileNet. The training data has been separated into 80% patient-wise groups to avoid data loss.

An automated testing dataset was created from the [20] NIH Chest X-ray dataset to evaluate the proposed system's generalization in real-world scenarios. The NIH dataset is used for testing. This dataset was selected because it includes additional clinical and demographic data that are essential for multimodal analysis but are not available in other public TB datasets. The clinical characteristics collected from the NIH dataset include patient age, gender, and clinically relevant results, which are used for creating the structured clinical feature representations in the proposed system. This dataset design enables effective feature learning from a range of publicly available datasets while also offering external validation of previously available real-world clinical data. The proposed technique improves reliability and usability for real-world TB detection by separating training and testing resources, using accurate clinical data during evaluation.

B. Ensemble Confusion Matrix

The Multimodal Ensemble Confusion Matrix shows the model's accuracy in identifying between non-TB and TB cases. Among the overall samples, 85 cases were non- tuberculosis accurately recognized but four cases were incorrectly diagnosed as TB. Similarly, 85 samples of tuberculosis were correctly recognized, with only three incorrectly labeled as non-TB. The model's overall accuracy is 96% showing that it is an appropriate method for classifying both groups. The rate of true positive and true negative values indicates the ensemble model has a high degree of dependability, reducing the number of false positives and false negatives showing that the model is effective in medical image classification tasks.

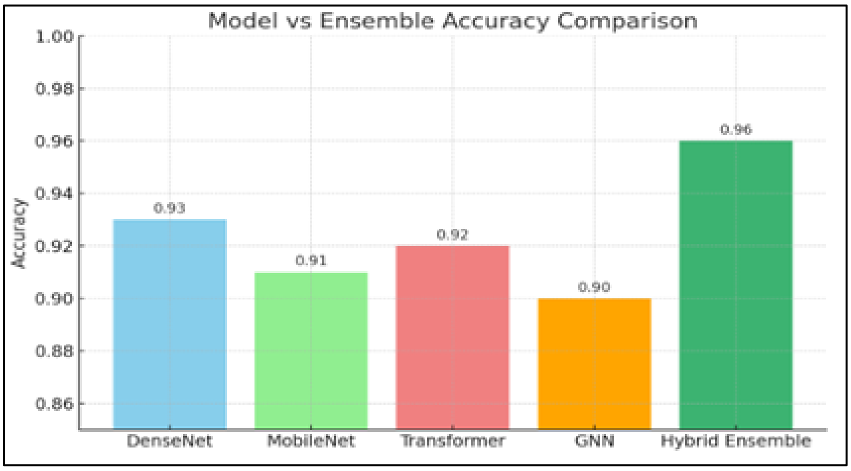

Figure 4. Comparison of Graph Model

The comparison chart of different models is represented in Figure 4 where the performance indicators such as accuracy, precision and recall is indicated to determine the effectiveness of the proposed system.

- **Accuracy**

Accuracy shows general accuracy of the model at both TB-positive and TB-negative. It is divided by the number of samples is classified from total number of samples. The value of accuracy is high, which means the model is suitable to reveal the difference between infected and healthy patients.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where:

- TP (True Positive): Predicted right TB cases.
- TN (True Negative): Predicted correctly non-TB cases.
- FP (False Positive): Cases of Non-TB that get wrongly expressed as TB.
- FN (False Negative): Cases of TB that are mistakenly identified to be non-TB

- **Precision**

Precision determines the number of samples that are predicted to be TB and those that are TB-positive. It is an important measure when there is a high cost of false positives and there is a need to work on the reliability of the positive prediction.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

High accuracy will ensure that the model does not place normal chest X-rays in the classification of TB which will matter in deciding clinically and limit unnecessary medical procedures.

- **Recall (Sensitivity)**

Recall, or sensitivity, or true positive rate, is used to assess the capacity of the model to correctly detect all the real instances of TB. It quantifies how well the model recognizes positive samples examples of the data set.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

The small recall value indicates that the system has the capability of identifying TB cases with low level of false negative and necessary to control the disease at its inception and control the population health.

- **F1-Score**

It is a mean of recall and precision, and characterizes a trade-off between recall and precision. It is mostly applicable in cases of class imbalance since it takes into account both false positives and false negatives during the methodology.

$$\text{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (4)$$

High F1-score proves that the model finds the excellent equilibrium of sensitivity and accuracy, which provides stable and reliable diagnostic results.

- **Confusion Matrix**

The confusion matrix provides a detailed visual representation of the classification results comparing the predicted labels to the actual datasets. It displays the distribution of true positives, true negatives, false positives and false negatives utilized to identify error patterns in the model. The confusion matrix for this project shows the proposed ensemble model accurately recognized the majority of TB and non-TB samples demonstrating that it is adaptable and dependable.

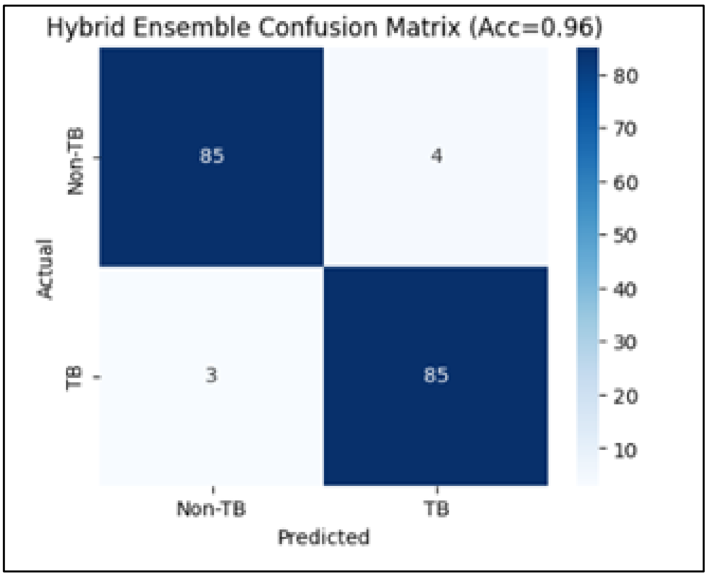

Figure 5. Confusion Matrix

Figure 5 represents the confusion matrix and it shows the success that the model has achieved in classifying the data by comparing the labels achieved by prediction with the actual ones to both measure the accuracy and the distribution of error.

Table 1. Accuracy of Multimodal

	Precision	Recall	F1-Score	Support
Non-TB	0.97	0.96	0.96	89
TB	0.96	0.97	0.96	88
Accuracy			0.96	177
Macro avg	0.96	0.96	0.96	177
Weighted avg	0.96	0.96	0.96	177

Table 1 illustrates the accuracy of multimodal. The Multimodal TB detection achieved a high precision of 0.96%-0.97%, an F1-score of 0.96%, an accuracy of 96%. This model used the additional data combining chest X-ray images and clinical data from DenseNet, MobileNet, Transformer and GNN. Attention-based fusion methods and the use of ensembles enabled feature learning, reduced overfitting and improved generalization demonstrating the availability of a flexible, accurate and clinically accessible system may be used to real-world TB diagnosis. Table 2 represents the ablation study results.

Table 2. Ablation Study Results

Model Configuration	Accuracy (%)	Precision	Recall	F1-score
DenseNet (Image only)	90.8	0.91	0.90	0.90
DenseNet + MobileNet	93.1	0.93	0.93	0.93
Imagemodels+Transformer	94.5	0.95	0.95	0.95
Proposed Multimodal Ensemble Model	96.0	0.96	0.97	0.96

The ablation results including clinical data into image-based data in deep learning algorithms will improve TB detection ability. A combination of transformer-based learning with GNN-based feature connection improves robustness and generalizability. The ensemble fusion method improves performance showing effective use of the proposed multimodal architecture.

6. Future Enhancements

Explainable Artificial Intelligence (XAI)

- Grad-CAM++, SHAP and LIME will be used as three different quantification and visualization methodologies to predict model results.
- The most important lung regions will be highlighted and the role of clinical features in TB detection will be demonstrated.
- The model will improve the transparency and clinicians' dependence on the system will be improved, leading the way for the system's reliability and implementation in clinical practice.

Capability Detection for Multiple Diseases

- The system will be improved further by the use of a multilabel classification technique and the utilization of huge datasets resulting in the detection of diseases such as pneumonia, COVID-19 and lung fibrosis.
- Task classification: The existing TB model may be converted into a lung disease detection system creating a multifunctional diagnostic tool.
- The patient will recognize the clinical value and efficiency of the healthcare system due to various lung problems may be recognized immediately through the combination of symptoms and Chest X-ray inputs.

7. Conclusion

This work describes an efficient multimodal deep learning architecture for detecting TB that addresses basic challenges with existing diagnostic algorithms based on images. The proposed system combines radiographic results with real-world clinical and demographic data to process actual clinical diagnostic procedures more accurately compared to earlier techniques. The combination of DenseNet and MobileNet enables the extraction of valuable data with high accuracy, allowing Graph Neural Networks and Transformer attention mechanisms to model symptom connections effectively and with patient-specific medical significance. The most significant feature of the proposed system is its complex evaluation method, which depends on a separate NIH Chest X-ray dataset with actual clinical data for external verification. This concept has been shown to be more adaptable than existing designs that have been evaluated on selected datasets. Empirical studies have shown that the proposed system can achieve acceptable accuracy levels, significantly improve recall, and achieve low false-negative rates. This is a critical aspect in the field of tuberculosis evaluation, as false negatives have the potential to cause serious public health problems.

Furthermore, the implementation of explainable AI algorithms increases data transparency by focusing on clinically significant parts of the lungs and emphasizing important symptoms in the decision-making process. Overall, the proposed multimodal ensemble system improves current TB detection models based on diagnostic reliability, interpretability, and clinical significance, demonstrating its applicability and high quality for real-world testing.

References

[1] Khaing, Aung Phyo. "A Study on Economic Burden of Tuberculosis Patients (Case Study: South Dagon Township, Yangon Region)(Aung Phyo Khaing, 2024)." PhD diss., MERAL Portal, 2024.

[2] Galbusera, Fabio, and Andrea Cina. "Image annotation and curation in radiology: an overview for machine learning practitioners." European Radiology Experimental 8, no. 1 (2024): 11.

[3] Hansun, Seng, Ahmadreza Argha, Siaw-Teng Liaw, Branko G. Celler, and Guy B. Marks. "Machine and deep learning for tuberculosis detection on chest X-rays: Systematic literature review." Journal of medical Internet research 25 (2023): e43154.

[4] Heiliger, Lars, Anjany Sekuboyina, Bjoern Menze, Jan Egger, and Jens Kleesiek. "Beyond medical imaging-a review of multimodal deep learning in radiology." TechRxiv 19103432 (2022).

[5] Huy, Vo Trong Quang, and Chih-Min Lin. "An improved densenet deep neural network model for tuberculosis detection using chest x-ray images." IEEE Access 11 (2023): 42839-42849.

[6] Kotei, Evans, and Ramkumar Thirunavukarasu. "A comprehensive review on advancement in deep learning techniques for automatic detection of tuberculosis from chest X-ray images." Archives of computational methods in engineering 31, no. 1 (2024): 455-474.

[7] Sharma, Vinayak, Sachin Kumar Gupta, and Kaushal Kumar Shukla. "Deep learning models for tuberculosis detection and infected region visualization in chest X-ray images." Intelligent Medicine 4, no. 2 (2024): 104-113.

[8] Urooj, Shabana, S. Suchitra, Lalitha Krishnasamy, Neelam Sharma, and Nitish Pathak. "Stochastic learning-based artificial neural network model for an automatic tuberculosis detection system using chest X-ray images." IEEE Access 10 (2022): 103632-103643.

[9] Mehrrotraa, Rajat, M. A. Ansari, Rajeev Agrawal, Pragati Tripathi, Md Belal Bin Heyat, Mohammed Al-Sarem, Abdullah Yahya Mohammed Muaad, Wamda Abdelrahman Elhag Nagmeldin, Abdelzahir Abdelmaboud, and Faisal Saeed. "Ensembling of efficient deep convolutional networks and machine learning algorithms for resource effective detection of tuberculosis using thoracic (chest) radiography." IEEE Access 10 (2022): 85442-85458.

[10] Munadi, Khairul, Kahlil Muchtar, Novi Maulina, and Biswajeet Pradhan. "Image enhancement for tuberculosis detection using deep learning." IEEE Access 8 (2020): 217897-217907.

[11] Alsdurf, H., B. Empringham, C. Miller, and A. Zwerling. "Tuberculosis screening costs and cost-effectiveness in high-risk groups: a systematic review." BMC infectious diseases 21, no. 1 (2021): 935.

[12] Sharma, Anubhav, Karamjeet Singh, and Deepika Koundal. "A novel fusion based convolutional neural network approach for classification of COVID-19 from chest X-ray images." Biomedical Signal Processing and Control 77 (2022): 103778.

[13] Bhosale, Yogesh H., and K. Sridhar Patnaik. "IoT deployable lightweight deep learning application for COVID-19 detection with lung diseases using RaspberryPi." In 2022 International conference on IoT and blockchain technology (ICIBT), pp. 1-6. IEEE, 2022.

[14] Malik, Hassaan, Tayyaba Anees, Muhammad Umar Chaudhry, Radomir Gono, Michał Jasiński, Zbigniew Leonowicz, and Petr Bernat. "A novel fusion model of hand-crafted features with deep convolutional neural networks for classification of several chest diseases using X-ray images." IEEE Access 11 (2023): 39243-39268.

[15] Xu, Tianhao, and Zhenming Yuan. "Convolution neural network with coordinate attention for the automatic detection of pulmonary tuberculosis images on chest x-rays." IEEE Access 10 (2022): 86710-86717.

[16] Uslu, Fatmatülzehra, and Anil A. Bharath. "TMS-Net: A segmentation network coupled with a run-time quality control method for robust cardiac image segmentation." Computers in Biology and Medicine 152 (2023): 106422.

[17] Abideen, Zain Ul, Mubeen Ghafoor, Kamran Munir, Madeeha Saqib, Ata Ullah, Tehseen Zia, Syed Ali Tariq, Ghufran Ahmed, and Asma Zahra. "Uncertainty assisted robust tuberculosis identification with bayesian convolutional neural networks." Ieee Access 8 (2020): 22812-22825.

[18] Rahman, Tawsifur, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F. Islam, Rashid Mazhar, Tahir Hamid et al. "Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization." Ieee Access 8 (2020): 191586-191601.

[19] Hooda, Rahul, Ajay Mittal, and Sanjeev Sofat. "Tuberculosis detection from chest radiographs: a comprehensive survey on computer-aided diagnosis techniques." Current Medical Imaging Reviews 14, no. 4 (2018): 506-520.

[20] National Library of Medicine. (2024). Open-i: Open Access Biomedical Image Search Engine (Tuberculosis Chest X-ray Dataset). U.S. National Library of Medicine, National Institutes of Health. Available at: https://openi.nlm.nih.gov/