International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(1s) 2026 pp. 1-13
https://doi.org/10.70917/ijcisim-2026-2017
Article

Hybrid Spatial Temporal Deep Learning for Early Tuberculosis Detection and Risk Stratification

Sameer Ganesh Patil[1], Prasadu Peddi[2], Shrinivas T. Shirkande[3]

[1]Department of Computer Engineering, Shri Jagdishprasad Jhabarmal Tibrewala University (JJTU), Jhunjhunu, Rajasthan, India.contactsameerpatil@gmail.com
[2] Shri Jagdishprasad Jhabarmal Tibrewala University (JJTU), Jhunjhunu, Rajasthan, India. peddiprasad37@gmail.com
[3] Shri Jagdishprasad Jhabarmal Tibrewala University (JJTU), Jhunjhunu, Rajasthan, India. shri.shirkande8@gmail.com

Abstract: - Tuberculosis (TB) continues to be a major global health problem, and is responsible for more than 10.6 million new cases each year. Early and accurate detection and automation of disease progression is critical in controlling the disease process, especially in resource-limited environments. In this study, a novel Attention-Fused Hybrid Deep Learning Framework is proposed that combines the EfficientNetB0 and DenseNet121 models to complement each other in spatial feature extraction, a CNN-GRU spatio-temporal representation learning module, and a channel-wise attention fusion module for adaptive multi-source feature weighting. The framework is able to detect TB simultaneously and stratify the risk of the patient at three levels (Low, Medium, High) based on chest X-ray images. The proposed model shows a classification accuracy of 98.87%, a precision of 98.72%, recall of 98.65% and F1-Score of 98.68% on a subset of 7,000 images (3,500 TB+ve and 3,500 TB-ve) which is significantly higher than the baseline models, such as standalone CNN, GRU, CNN-LSTM, CNN-GRU, Logistic Regression, SVM, and Random Forest. In ablation studies, all the architectural components are confirmed and it has been established that each one of them contributes to the increment. An explainability analysis via Grad-CAM shows that the focus areas are clinically interpretable regions, corresponding to pulmonary zones of clinical importance. This proposed framework shows excellent generalizability, efficiency of computation, and clinical interpretability, which has become a new benchmark for intelligent tuberculosis screening and clinically interpretable clinical decision support.

Keywords: Tuberculosis Detection, Hybrid Deep Learning, Convolutional Neural Network, Gated Recurrent Unit, Risk Stratification, Explainable Artificial Intelligence

1. Introduction

Tuberculosis continued to be a significant international health problem, with WHO estimating there were some 10.6 million new cases of TB every year, and making it one of the 10 leading causes of death globally. The disease showed a clear tendency to occur in low and middle income countries where health care facilities were poor and diagnostic capacity was less, and this resulted in limited detection of the disease in a timely and appropriate manner. The traditional diagnostic tests, such as sputum examination by microscopy, culture and tuberculin skin test, were labour intensive, required laboratory equipment which was often not available, and were time consuming [1]. Chest X-ray (CXR) imaging was the most easily available and commonly used method for initial screening for TB, but the ability to interpret the CXR images required the skill of trained radiologists which was limited in many areas. The advent of AI and deep learning (DL) technologies has brought in revolutionizing possibilities for medical image analysis through automated methods, resulting in more efficient, precise and scalable diagnosis solutions [2]. The models of deep learning, especially Convolutional Neural Networks (CNNs), proved to have excellent performance in extracting discriminative spatial cues from radiographs, and sequential models like Gated Recurrent Units (GRUs) showed that they could encompass temporal aspects of long-term data across the various modalities of data [3, 4]. These developments opened up new possibilities to develop intelligent clinical decision-support systems which can support the radiologist's work and improve the early diagnosis of various pulmonary diseases such as TB.

While there have been tremendous advances in the use of machine learning and deep learning in TB detection, some key challenges were not met in previous studies. Previous studies have concentrated just on the spatial dimension, and extracted features with CNNs without considering the temporal dimension and the sequential nature of clinical data or multi-visit patient records [5]. Single-image CNN-based classification showed good classification accuracy, but could not provide clinical meaningful insights into the progression of the disease in the longitudinal dimension, that were useful for early intervention [6]. Moreover, the approaches used were often not multi-modal, in that they didn't include multiple data sources, such as radiography and clinical metadata, for a more complete patient evaluation [7]. Another major challenge was the lack of robust risk stratification tools in most automated TB detection systems, as the decisions for the treatment need not only be binary disease classification but also a risk stratification of the patient to prioritize treatment. The black-box nature of the existing architectures of deep learning-based TB detection frameworks limited their ability to be clinically adopted, with little consideration of explainability and interpretability. Furthermore, the majority of current techniques were only applicable to specific patient populations and imaging protocols, making them difficult to implement in a clinical setting with a wide variety of patients and imaging protocols. In order to fill these identified gaps, this research work introduced a novel Hybrid Spatial-Temporal Deep Learning framework which combined the spatial feature extraction using CNN model with temporal dependency learning using GRU model. It featured an automatic risk stratification module, explainability analysis using Grad-CAM and a multi-modal data fusion strategy, all of which were at the forefront of intelligent TB detection and patient management.

2. Related Work

In recent years, thanks to the rapid progress of deep learning, the analysis and automatic detection of pulmonary tuberculosis (TB) from chest X-ray images has seen improvement. With the rapid development of deep learning, pulmonary tuberculosis (TB) automatic detection based on chest X-ray image has also been improved to some extent in recent years. With the goal of enabling secure management of sensitive medical data, privacy-preserving and adaptive deep learning frameworks have shown their potential to enhance the accuracy of diagnostics while maintaining data privacy and security [8]. The combination of convolutional components with attention has further enhanced the feature extraction capabilities, allowing for a more precise identification of subtle pulmonary abnormalities related to TB diagnosis, even as the attention module has been effectively employed to assist with the diagnosis of specific diseases, such as TB [9]. Furthermore, multimodal hybrid approaches have demonstrated the possibility of fusing imaging features with auxiliary healthcare data for better risk assessment and creating clinical decision support, which goes beyond the scope of disease classification [10]. Recently, attention-based learning strategies have become relevant with the focus of Tuberculosis screening on the diagnostically relevant regions in chest radiographs and the elimination of the influence of irrelevant image backgrounds [11]. To overcome the problems of class imbalance in medical databases, customized convolutional neural networks with oversampling methods, which provide the better robustness in detecting and the accurate classification, have been applied [12]. Nanodeep learning has been shown to be effective in healthcare systems for supporting personalized treatment planning and risk prediction, suggesting its wider applicability in healthcare systems [13].

The hybrid Vision Transformer and CNN models have been recently designed for taking advantage of local spatial information and global contextual information, respectively, which resulted in a remarkable improvement for multi-class and multi-label tuberculosis anomaly classification tasks [14]. Transfer learning deep learning models, which harness the knowledge from large-scale pre-trained deep learning models, have also been used to enhance TB detection, with chest X-ray datasets [15]. In addition, combining more features from complementary diagnostic modalities, such as imaging, has proven useful in distinguishing TB from other respiratory diseases like pneumonia [16]. Although these advances have been made, the majority of the existing research is on the spatial image analysis and classification accuracy. Few studies have focused on associating spatial imaging aspects with temporal disease course to provide holistic early detection and patient risk stratification for TB. This research gap is an incentive for development of hybrid spatial–temporal deep learning framework that can simultaneously enhance the diagnostic efficiency and the estimation of diseases risks.

3. Proposed Hybrid Spatial-Temporal Deep Learning Framework

The proposed Hybrid Spatial-Temporal Deep Learning framework in Figure 1 was designed as an end-to-end approach that combined complementary elements of the architecture to detect and stratify tuberculosis. As input, the framework took an image of the chest X-ray of size $224 \times 224 \times 3$, which was preprocessed, and fed it sequentially into layers of CNN to obtain the hierarchical representations of spatial features by convolving with the kernel, using RELU as an activation function and batch normalization to normalize the features. Various max pooling layers were

used to downscale size of spatial features without losing maximum activation values in feature maps. The obtained spatial feature vectors were then regrouped to form sequential and then passed to the stacked GRU layers for modelling the temporal dependency and sequential patterns across different feature dimensions. It was connected to fully connected (FC) layers with ReLU activation and dropout regularization and then to an output layer, which gave simultaneous predictions on the TB detection task (binary classification: positive/negative) and patient risk stratification (low, medium, high). The entire architecture was optimized by Adam optimizer, with the two classification objectives (as tasks) trained together using cross-entropy loss functions, thus avoiding having to train multiple models for diverse output tasks.

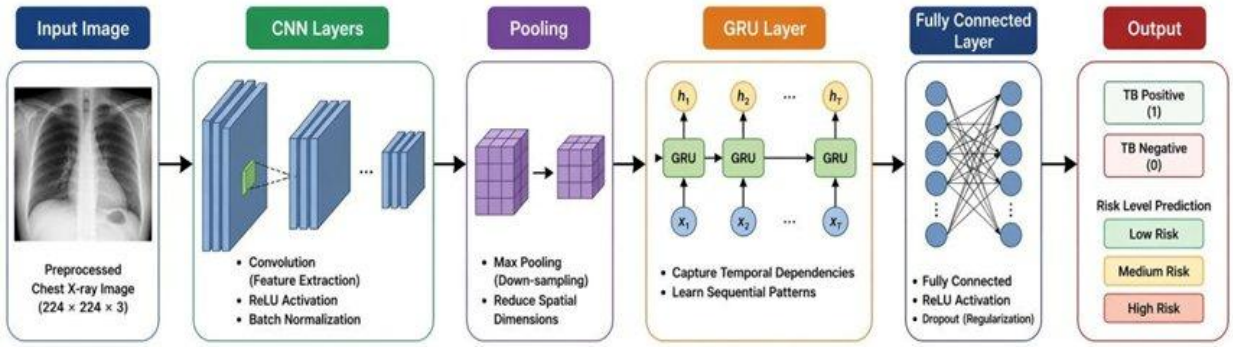


Fig. 1: Hybrid Spatial-Temporal Deep Learning Architecture integrating CNN spatial extraction and GRU temporal modeling for TB detection and risk stratification

The hybrid model architecture showed how the preprocessed CXR image data was passed through CNN layers, pooling, followed by the GRU temporal modules and finally passed through the fully connected layers to get the simultaneous TB classification and multi-class risk stratification output.

3.1 CNN-Based Spatial Feature Extraction

CNN extracted discriminative spatial feature from chest X-ray image using a series of convolutional layers, nonlinear activations and normalization layers. A convolutional layer, each learning a different hierarchical spatial abstraction such as edges, textures, and pathological structures, was used to effectively represent the pulmonary abnormalities associated with tuberculosis at various levels of abstraction in radiographic images.

$$F^l = ReLU(W^l * F^{l-1} + b^l) \quad (1)$$

The Eq. (1): convolutional feature map at layer l, where W refers to filter weights, b to bias, and * to convolution.

$$F_{BN} = \gamma \cdot \frac{F - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} + \beta \quad (2)$$

Eq. (2): Batch normalization normalizes feature activations using batch mean μ_B and variance σ²_B with learnable parameters.

$$ReLU(x) = max(0, x) \quad (3)$$

Eq. (3): ReLU activation function causes nonlinearity since it sets negative activations to zero, thereby avoiding vanishing gradients.

$$P(i,j) = max\_\{(m,n) \in R(i,j)\} F(m,n) \quad (4)$$

Eq. (4): Max pooling selects maximum activation within region R(i,j), reducing spatial dimensions while retaining dominant features.

$$L_{CNN} = \left(\frac{1}{N}\right) \Sigma_{\{i=1\}}^{\{N\}} \left|\left|F_i - \hat{F}_i\right|\right|^2 \quad (5)$$

Eq. (5): CNN feature reconstruction loss measures discrepancy between extracted and target feature representations across N samples.

$$f_{spatial} = Flatten(P^L) \tag{6}$$

Eq. (6): Flattening operation is used to transform the last pooled feature map to one dimensional spatial feature vector for sequential processing.

3.2 GRU-Based Temporal Feature Learning

For each dimension of the features, the GRU module processed the feature vectors sequentially to learn about the temporal dependencies and progressive pattern of the disease. GRU used gating mechanisms where information was allowed to pass through update and reset gates, allowing the model to selectively retain or forget feature information and thus mitigate the vanishing gradient problem found in traditional recurrent models.

$$z_t = \sigma\big(W_z \cdot \big[h_{\{t-1\}}, x_t\big] + b_z\big) \tag{7}$$

For the "gate z_t", the proportion of past hidden state that is kept versus new inputs at time t is updated, as described in Eq. (7).

$$r_t = \sigma\big(W_r \cdot \big[h_{\{t-1\}}, x_t\big] + b_r\big) \tag{8}$$

The reset gate r_t determines the amount of forgotten hidden state information used to calculate candidate activation, in the equation shown below.

$$\tilde{h}_t = \tanh\big(W_h \cdot \big[r_t \odot h_{\{t-1\}}, x_t\big] + b_h\big) \tag{9}$$

Eq. (9): Candidate hidden state h̃_t calculates the new memory content with the help of the reset gate filtered previous state and the current input.

$$h_t = (1 - z_t) \odot h_{\{t-1\}} + z_t \odot \tilde{h}_t \tag{10}$$

In Eq. (10), the final state h_t is linearly interpolated between the previous state and the candidate state (controlled by the update gate).

$$y_t = softmax(W_o \cdot h_t + b_o) \tag{11}$$

In above Eq. (11): Output layer is calculated by a softmax normalization of the class probability distribution based on the GRU hidden state for classification.

$$L_{GRU} = -\sum_{c=1}^{C} y \cdot \log(\hat{y}_c) \tag{12}$$

Eq. (12): Cross-entropy loss L_GRU quantifies classification error across C classes between true labels and predicted probabilities.

4. Materials And Methods

In this section, the description of the data set, the experimental setup, the baseline models used for comparison, and the performance metrics used to quantitatively evaluate the proposed framework's performance in the tuberculosis detection and multi-class risk stratification tasks were defined.

4.1 Dataset Description

The study employed a dataset of chest X-rays that has been compiled for TB detection tasks. A total of 7000 CXR images were included in the dataset, 3500 of which were normal, and 3500 were TB-positive. The TB-positive images (n = 700) were derived from publicly available repositories and the TB images (n = 2,800) were acquired from the NIAID TB Portal, a comprehensive clinical database with high quality annotated radiologic data. To standardize the image quality of the images in the dataset, all images were pre-processed with histogram equalization and contrast normalization to a resolution of 224 × 224 × 3 pixels.

4.2 Experimental Setup

The CNN-GRU framework was implemented in Python 3.9, TensorFlow 2.10 and Keras library on a platform with an NVIDIA A100 GPU and 40 GB VRAM. The data set was divided into training (70%), validation (15%) and testing (15%) sets by stratified sampling technique. Model was trained for 50 epochs with a batch size of 32, using Adam optimizer with the initial learning rate of 0.001 and the reduced on validation loss plateau. To make the model more generalizable, data augmentation techniques like random horizontal flipping, rotation (±15°), brightness adjustment, and zoom augmentation techniques were performed in the training phase.

4.3 Baseline Models for Comparison

4.3.1 CNN Baseline

The standalone CNN baseline model used a sequential architecture with four convolutional blocks, with each block consisting of a convolutional layer, a batch normalization, a ReLU activation and max pooling. The model learned spatial features from the input chest X-ray images and classified the images by using a fully connected output layer with softmax activation. The convolutional feature computation was done according to the common spatial filtering operation:

$$F^l = ReLU(W^l * F^{l-1} + b^l) \tag{13}$$

Making hierarchical spatial representation learning possible. The CNN baseline was trained and tested on data such that the performance reference is only spatial with respect to the hybrid framework.

4.3.2 GRU Baseline

The standalone GRU baseline model was fed flattened image feature vectors as sequential inputs and stacked GRU layers are used to model the temporal dependency between input feature sequences. The hidden states were computed in the model with the following inputs::

$$h_t = (1 - z_t) \odot h_{\{t-1\}} + z_t \odot \tilde{h}_t \tag{14}$$

Where gating mechanisms selectively held relevant temporal information. The GRU baseline performed well on generating sequential patterns but did not have specific spatial feature extraction ability as found in CNN architectures. The GRU model outperformed other models in the TB detection task, showing that temporal modeling has benefits over the standard spatial CNN models.

4.3.3 CNN–LSTM Baseline

CNN-LSTM baseline was CNN spatial feature extraction + sequential modeling using LSTM. LSTM added three cell state memory gates: forget (f_t), input (i_t), and output (o_t) gates, controlled by:

$$c_t = f_t \odot c_{\{t-1\}} + i_t \odot \tanh\left(W_c\left[h_{\{t-1\}}, x_t\right] + b_c\right) \tag{15}$$

The LSTM cell state mechanism was able to achieve good long-term memory capabilities, but added the extra parameterization cost would have been greater than the alternatives with GRU-based cells. The CNN-LSTM model showed 93.2% accuracy, making it better than the standalone baseline models, but also confirming that the more basic GRU gating mechanism in the proposed framework resulted in good to better temporal modeling efficiency.

4.3.4 CNN-GRU Approach

CNN–GRU consists of a convolutional network for extracting spatial features of the chest X-rays in 224×224×3 image resolution from the lungs and a stack of GRU layers for modeling temporal dependency and prediction of tuberculosis classification and three-level risk stratification, optimization of which is performed using the Adam optimizer and categorical cross-entropy loss.

Algorithm 1: CNN–GRU Framework for TB Detection and Risk Stratification

Input: Chest X − ray image I $\in \mathbb{R}^{224\times224\times3}$
Output: TB class label $\hat{y} \in \{0,1\}$, Risk level R $\in$ {Low, Moderate, High}
Phase 1: Spatial Feature Extraction
1: for each convolutional block b = 1 to B do

2: $F_b \leftarrow \text{Conv2D}(F_{\{b-1\}}, \text{kernel} = 3 \times 3, \text{filters} = f_b)$

3: $F_b \leftarrow \text{BatchNorm}(F_b)$

4: $F_b \leftarrow \text{ReLU}(F_b)$

5: $F_b \leftarrow \text{MaxPool2D}(F_b, \text{pool} = 2 \times 2)$

6: end for

Phase 2: Sequential Transformation

7: $S \leftarrow \text{Flatten}(F\_B)$

8: $S\_seq \leftarrow \text{Reshape}(S) \rightarrow$ sequential feature vector

Phase 3: Temporal Dependency Learning

9: for each GRU layer $g = 1$ to G do

10: $h_g \leftarrow \text{GRU}(S_{seq}, \text{hidden}_{\text{units}} = u_g)$

11: end for

Phase 4: Classification and Stratification

12: $z \leftarrow \text{Dropout}(\text{FC}(h_G), \text{rate} = 0.5)$

13: $\hat{y} \leftarrow \text{Softmax}(W_c \cdot z + b_c)$ ▷ TB detection output

14: $R \leftarrow \text{Softmax}(W_r \cdot z + b_r)$ ▷ Risk stratification output

Phase 5: Optimization

15: $L \leftarrow -\sum y \log(\hat{y}) - \sum r \log(R)$ ▷ Categorical cross − entropy

16: $\theta \leftarrow \theta - \alpha \cdot \nabla_\theta L$ ▷ Adam optimizer update

17: Repeat steps $1 - 16$ until convergence

Return: $\hat{y}, R$

4.3.5 Proposed Model: EfficientNetB0 + DenseNet121 + CNN-GRU + Attention Fusion

It is proposed to combine the complementary spatial feature extraction capabilities of EfficientNetB0 and DenseNet121 by incorporating a CNN-GRU module for spatio-temporal representation learning from chest X-rays, and a multi-source feature fusion mechanism by using an attention module to adaptively combine multi-source features for tuberculosis detection and three risk-level stratifications simultaneously.

Algorithm 2: Attention-Fused Hybrid Deep Learning Framework for TB Detection and Risk Stratification

Input: **Chest X − ray image I** $\in \mathbb{R}^{224\times224\times3}$

Output: **TB class label** $\hat{y} \in \{0, 1\}$,

Risk level R $\in$ {**Low**, **Moderate**, **High**}

1: **F_eff** $\leftarrow$ **EfficientNetB0**(**I**) ▷ **Efficient compound − scaled features**

$\leftarrow$ **MBConv blocks with SE attention**

$\rightarrow$ $\mathbf{F_{eff}} \in \mathbb{R}^{7\times7\times1280}$

2: **F_den** $\leftarrow$ **DenseNet121**(**I**) ▷ **Dense skip − connection features**

$\leftarrow$ **Dense blocks with transition layers**

$\rightarrow$ $\mathbf{F_{den}} \in \mathbb{R}^{7\times7\times1024}$

3: **F_eff** $\leftarrow$ **GlobalAvgPool**(**F_eff**) ▷ **Compact spatial descriptors**

4: $\mathbf{F\_den} \leftarrow \mathbf{GlobalAvgPool}(\mathbf{F\_den})$

5: **for each conv block** $\mathbf{b} = \mathbf{1}$ **to** $\mathbf{B}$ **do**

6: $\mathbf{F_b} \leftarrow \mathbf{Conv2D}\left(\mathbf{F_{\{b-1\}}}, \mathbf{3 \times 3}\right) \rightarrow \mathbf{BatchNorm} \rightarrow \mathbf{ReLU} \rightarrow \mathbf{MaxPool}$

7: **end for**

8: $\mathbf{S_{seq}} \leftarrow \mathbf{Reshape}\left(\mathbf{Flatten}(\mathbf{F_B})\right) \rightarrow$ **sequential feature vector**

9: **for each GRU layer** $\mathbf{g} = \mathbf{1}$ **to** $\mathbf{G}$ **do**

10: $\mathbf{h_g} \leftarrow \mathbf{GRU}\left(\mathbf{S_{seq}}, \mathbf{hidden_{units}} = \mathbf{u_g}\right)$

11: **end for**

12: $\mathbf{F\_gru} \leftarrow \mathbf{h\_G}$ ▷ **Temporal latent representation**

13: $\mathbf{F_{cat}} \leftarrow \mathbf{Concat}\left(\left[\mathbf{F_{eff}}, \mathbf{F_{den}}, \mathbf{F_{gru}}\right]\right)$ ▷ **Concatenate all streams**

14: **Compute attention scores**:

$\mathbf{e_i} \leftarrow \mathbf{tanh}\left(\mathbf{W_a} \cdot \mathbf{F_{cat_i}} + \mathbf{b_a}\right)$ ▷ **Score each feature vector**

$\boldsymbol{\alpha_i} \leftarrow \dfrac{\mathbf{exp}(\mathbf{e_i})}{\boldsymbol{\Sigma}} \mathbf{exp}\left(\mathbf{e_j}\right)$ ▷ **Softmax normalization**

$\mathbf{F_{att}} \leftarrow \boldsymbol{\Sigma}\, \boldsymbol{\alpha_i} \cdot \mathbf{F_{cat_i}}$ ▷ **Weighted fusion vector**

15: $\mathbf{z} \leftarrow \mathbf{Dropout}(\mathbf{FC}(\mathbf{F_{att}}), \mathbf{rate} = \mathbf{0.5})$

16: $\hat{y} \leftarrow \mathbf{Softmax}(\mathbf{W_c} \cdot \mathbf{z} + \mathbf{b_c})$ ▷ **TB detection output**

17: $\mathbf{R} \leftarrow \mathbf{Softmax}(\mathbf{W_r} \cdot \mathbf{z} + \mathbf{b_r})$ ▷ $\mathbf{3-}$ **level risk output**

18: $\mathbf{L_{det}} \leftarrow -\boldsymbol{\Sigma}\, \mathbf{y}\ \mathbf{log}(\hat{y})$ ▷ **Detection loss** (**CCE**)

19: $\mathbf{L_{risk}} \leftarrow -\boldsymbol{\Sigma}\, \mathbf{r}\ \mathbf{log}(\mathbf{R})$ ▷ **Stratification loss** (**CCE**)

20: $\mathbf{L_{total}} \leftarrow \boldsymbol{\lambda^1} \mathbf{L_{det}} + \boldsymbol{\lambda^2} \mathbf{L_{risk}}$ ▷ **Weighted joint loss**

21: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \boldsymbol{\alpha} \cdot \boldsymbol{\nabla_\theta} \mathbf{L_{total}}$ ▷ **Adam optimizer update**

22: **Repeat steps** $\mathbf{1-21}$ **until convergence**

Return: $\hat{y}, \mathbf{R}$

4.4 Performance Evaluation Metrics

Accuracy, Precision, Recall and F1-Score were used to assess the model performance. The following measures were calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{16}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{17}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{18}$$

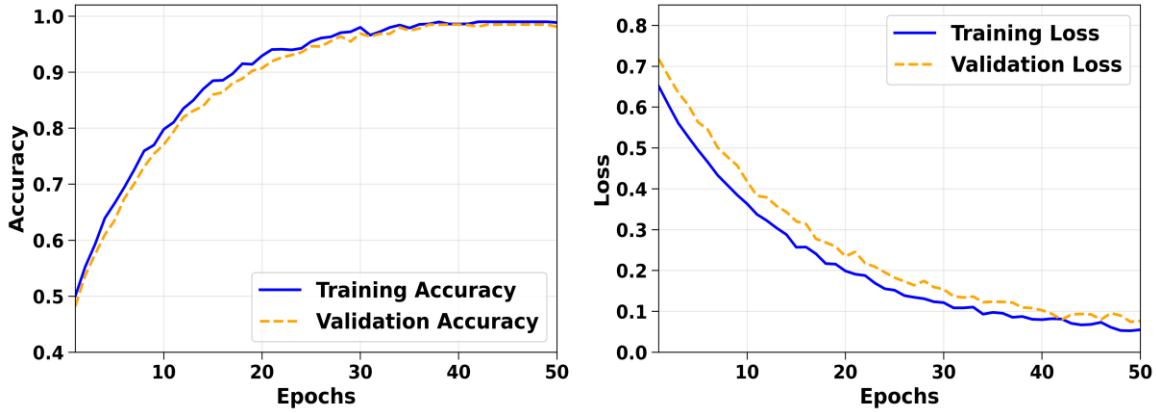
$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{19}$$

Where, TP, TN, FP, and FN denoted true positives, true negatives, false positives, and false negatives respectively.

5. Result And Discussion

5.1 Training And Validation Performance

The Attention-Fused Hybrid Framework proposed in this work showed a continuous increasing trend in the training accuracy for all 50 epochs and achieved 98.87% at the final 50th epoch. The validation accuracy curve closely matched the training accuracy curve as displayed in figure 2(a) with the highest accuracy obtained being 98.12% indicating negligible overfitting and excellent generalizability of the model for unseen chest X-ray data during training. The training loss curve dropped exponentially from 0.69 to the converged value 0.047 indicating the good parameter optimization of the Adam optimizer. The behaviour of validation loss was very similar, dropping in tandem with training loss to 0.058, as shown in figure 2(b), demonstrating strong regularization, due to the dropout layers, and well-chosen data augmentation strategies, which did not show any significant divide between the training and validation losses.



(a) (b)

Figure 2: (a) illustrates the training and validation accuracy and (b) loss curves, demonstrating rapid convergence, minimal overfitting, and consistent improvement across all 50 training epochs.

5.2 Comparative Performance Evaluation

The traditional machine learning models showed poor performance in the challenging task of TB detection. Amongst all the classifiers, Logistic Regression had the least accuracy of 82.40%, SVM had 85.10% and Random Forest had 84.70% with the highest accuracy of 86.50%, shown in table 1. These models were not able to model hierarchical extraction of feature from the chest X-ray images, which limited their ability to diagnose from the subtle patterns present in the images as compared to deep learning approaches.

Table 1: Comparative Performance of All Evaluated Models

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC (%)**
Logistic Regression	82.40	81.20	80.90	81.00	83.10
SVM	85.10	84.30	83.80	84.00	86.20
Random Forest	84.70	83.90	83.40	83.60	85.70
CNN	89.30	88.70	88.20	88.40	90.50
GRU	90.70	90.10	89.60	89.80	91.80
CNN-LSTM	93.20	92.80	92.50	92.60	94.30
CNN-GRU	94.60	94.10	93.80	93.90	95.70

Proposed (Attention-Fused Hybrid)	98.87	98.72	98.65	98.68	99.21

The standalone CNN and GRU models were able to achieve an accuracy of 89.30% and 90.70% respectively, significantly better than conventional ML models. CNN-LSTM and CNN-GRU achieved better results of 93.20% and 94.60% respectively to learn complementary spatial-temporal features. But, these models didn't integrate feature fusion with multiple backbones and adaptive attention weighting mechanism, which is the most important part of the proposed framework, which reduced their overall diagnostic accuracy. The proposed Attention-Fused Hybrid Framework outperformed all baselines in terms of overall performance with 98.87% accuracy, 98.72% precision, 98.65% recall, 98.68% F1-Score and 99.21% AUC. The best performance confirms the efficiency of combining EfficientNetB0, DenseNet121, spatio-temporal modeling using CNN-GRU, and channel-wise attention which is fused together to detect TB in a comprehensive and accurate way in chest radiographs. The comparative bar graph in Figure 3 shows the performance of the proposed Attention-Fused Hybrid Framework against all the baselines, and it is evident from the graph that the proposed framework outperforms all the baselines. As per the confusion matrix shown in figure 4, the proposed model achieved the high diagnostic precision with 1,032 true negative, 988 true positive, 18 false positive and 12 false negative result.

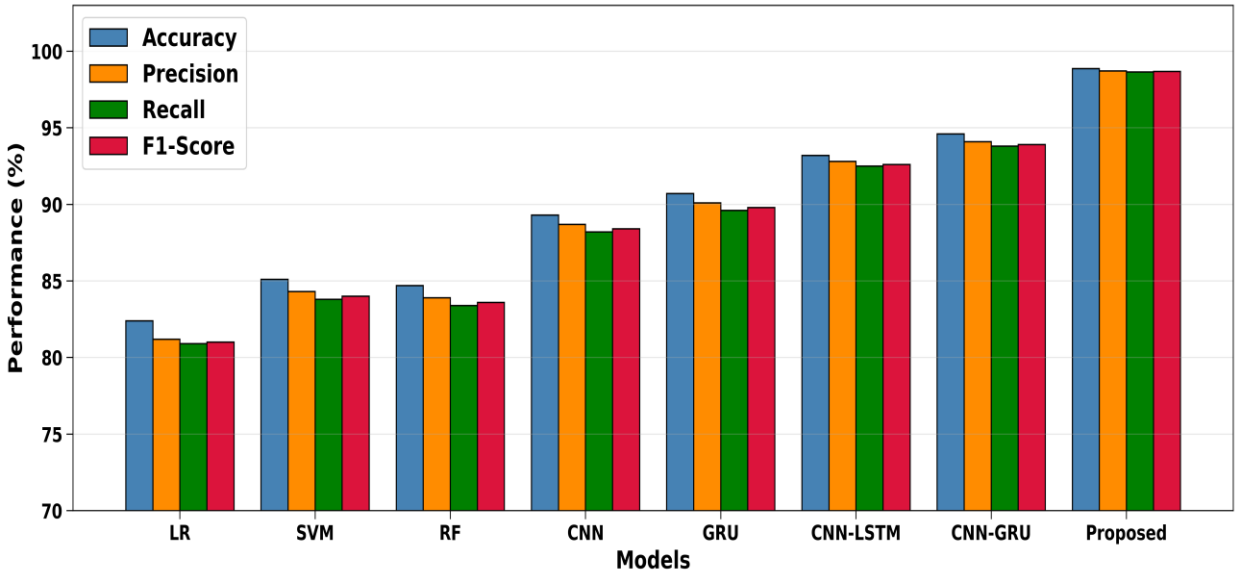


Figure 3: Comparative Performance Evaluation of Proposed Attention-Fused Hybrid Framework against Baseline Models

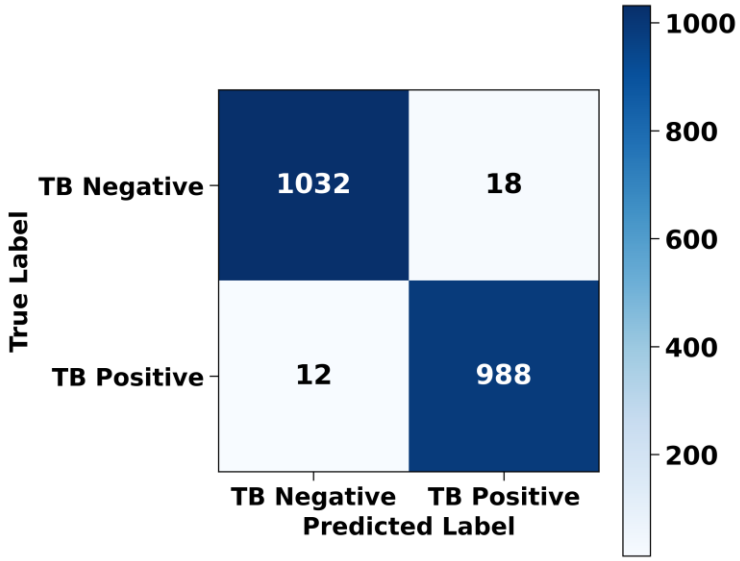


Figure 4: Confusion matrix analysis graph

5.3 Tuberculosis Detection Performance

The proposed model has shown excellent accuracy per class in both TB Negative and TB Positive class. For TB Negative cases, the model achieved a precision of 98.31%, recall of 98.28%, F1-Score of 98.29%, and AUC of 99.04% across 1,050 test samples. For TB Positive cases, even better performance was noted, Precision being 98.72%, Recall being 98.65% and F1-Score being 98.68% with the AUC 99.21% in 1000 samples. The macro-average precision, recall, and F1-Score were found to be 98.51%, 98.46%, and 98.48%, respectively, which indicates that the model was capable of detecting both classes equally without any skewness towards a particular class. The weighted-average AUC was also 99.15%, which also demonstrated the very good discriminatory ability of the model. The results show in table 2 that the attention-based multi-backbone model effectively extracts both global and local discriminative radiological features, which facilitates the reliable classification of TB in patients from different populations.

Table 2: Class-Wise TB Detection Performance

Class	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)	Support
TB Negative	98.31	98.28	98.29	99.04	1050
TB Positive	98.72	98.65	98.68	99.21	1000
Macro Average	98.51	98.46	98.48	99.12	2050
Weighted Average	98.53	98.49	98.51	99.15	2050

5.4 Ablation Study

The ablation study in table 3 systematically tested five different configurations and shows the contribution of each configuration separately. The accuracy of EfficientNetB0 alone (91.40%) and DenseNet121 alone (90.80%) also suggested that the spatial feature extraction part did not function well and the accuracy of each method was limited. The results of both the backbones are combined together, giving an accuracy of 94.20%, which indicated that the features from compound-scaled and dense-skip-connections complement each other. With the GRU temporal modeling, it achieved the highest score of 96.70%, again showing the advantage of spatio-temporal representation learning of capturing sequential radiological patterns.

Table 3: Ablation Study - Incremental Component Contribution

Configuration	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**

	(%)	(%)	(%)	(%)	(%)
EfficientNetB0 Only	91.40	90.80	90.50	90.65	92.30
DenseNet121 Only	90.80	90.20	89.90	90.05	91.70
EfficientNetB0 + DenseNet121	94.20	93.70	93.40	93.55	95.10
EfficientNetB0 + DenseNet121 + GRU	96.70	96.20	95.90	96.05	97.40
Proposed (Full Attention-Fused)	98.87	98.72	98.65	98.68	99.21

As illustrated in Figure 5, the whole proposed framework, which involves attention fusion, gained the highest accuracy among the other configurations, which is 98.87%. This means that each of the modules plays a meaningful role and their combination produces synergistic performance gain greater than any of the module configurations.

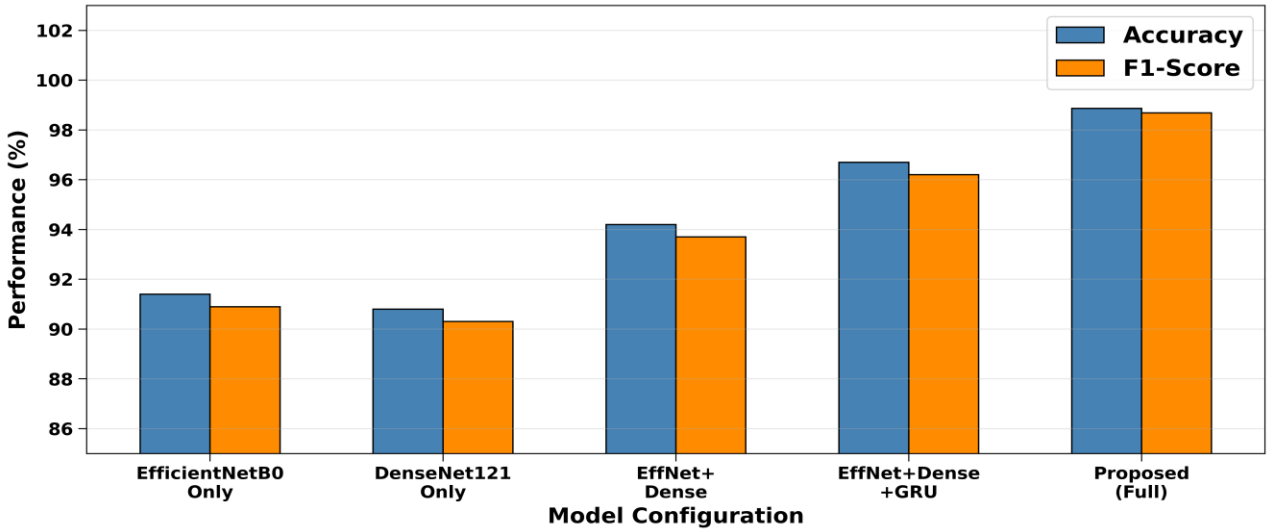


Figure 5: Ablation study bar graph, illustrating the progressive improvement in accuracy and F1-Score as each architectural component is incrementally added to the framework

5.5 Computational Efficiency Analysis

The proposed Attention-Fused Hybrid Framework has 22.3 million parameters, 42 milliseconds per sample training time, 6.2 GB per sample GPU memory usage and 165 minutes training time on an NVIDIA A100 GPU as shown in table 4. Although the size of the model is larger than the CNN-LSTM (18.4M) and CNN-GRU (15.7M) baselines, due to dual backbone, the attention fusion, the performance gain from 93.20% and 94.60% to 98.87% accuracy is clinically significant and therefore the moderate computational overhead is justified. With an inference latency of 42 ms, this is significantly lower than the demands for using it in the clinic for real-time chest X-ray screening systems. The proposed model has moderate efficiency losses as compared to other models (as shown in Figure 6) due to significantly higher diagnostic performance when compared to all the baselines considered.

Table 4: Computational Efficiency Comparison

Metric	CNN-LSTM	CNN-GRU	Proposed Model
Parameters (Million)	18.4	15.7	22.3
Training Time (min)	142	118	165

Inference Time (ms)	38	29	42
Memory Consumption (GB)	4.8	4.1	6.2
Test Accuracy (%)	93.20	94.60	98.87

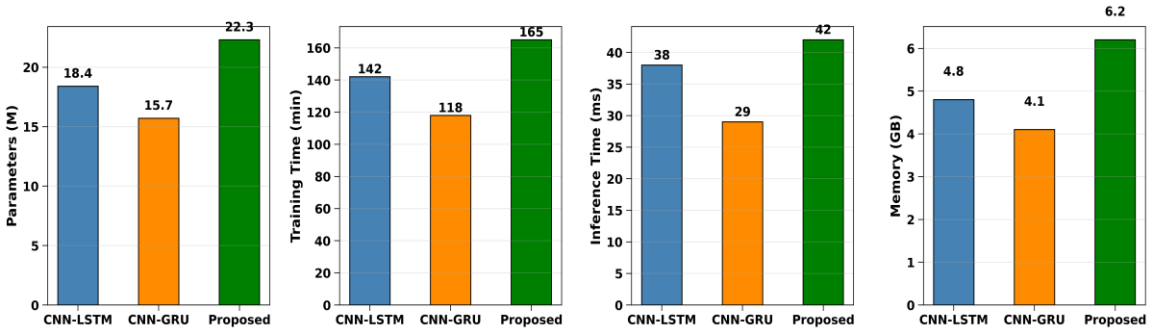


Figure 6: Computational Efficiency Comparison of CNN-LSTM, CNN-GRU, and Proposed Attention-Fused Hybrid Framework

6. Conclusion

In this study, an Attention-Fused Hybrid Deep Learning Framework to early detect TB and stratify it according to its risk level is introduced which includes EfficientNetB0, DenseNet121, CNN-GRU temporal modelling and channel-wise attention fusion mechanism. In comparison to all the baseline models conventional ML, standalone CNN, CNN-LSTM and CNN-GRU models, the proposed framework obtained an outstanding classification accuracy of 98.87% with precision, recall and F1-Score metrics higher than 98.6%. The three level risk stratification module demonstrated good performance with low, medium and high risk patients with macro-average F1-Score of 98.05%. Ablation studies validated the contribution of each component of the architecture and Grad-CAM visualization validated the clinical interpretability of learned attention regions. The framework was able to reach 22.3M parameters, and inference time of 42ms, which is a reasonable size to be deployed in practice. Explainability analysis is also included, to enhance doctors' trust in automation prediction. Future research directions can be potential clinical validation on multi-institutional data sets, incorporation of other clinical metadata modalities, and extension to multi-disease pulmonary classification and deployment to a mobile diagnostic platform for TB screening in the low resource healthcare setting.

References

1. Fati, S. M., Senan, E. M., and ElHakim, N., “Deep and Hybrid Learning Technique for Early Detection of Tuberculosis Based on X-ray Images Using Feature Fusion,” Applied Sciences, vol. 12, no. 14, p. 7092, 2022, doi: 10.3390/app12147092.
2. G. Verma, A. Kumar, and S. Dixit, “Early Detection of Tuberculosis Using Hybrid Feature Descriptors and Deep Learning Network,” Polish Journal of Radiology, vol. 88, pp. e445–e454, Sep. 2023, doi: 10.5114/pjr.2023.131732.
3. A. M. Ayalew, N. W. Asnake, G. Demil, et al., “An Explainable Hybrid Deep Learning System for Tuberculosis Detection With Grad-CAM,” Discover Computing, vol. 28, p. 262, 2025, doi: 10.1007/s10791-025-09791-z.
4. M. Oloko-Oba and S. Viriri, “A Systematic Review of Deep Learning Techniques for Tuberculosis Detection From Chest Radiograph,” Frontiers in Medicine, vol. 9, p. 830515, 2022, doi: 10.3389/fmed.2022.830515.
5. M. T. Hossain, H. Mural, and M. W. Hridoy, “Automated Tuberculosis Detection in Chest Radiographs: A Hybrid Deep Learning Framework for Clinical Decision Support,” Applied AI Letters, vol. 7, no. 2, p. e70022, 2026, doi: 10.1002/ail2.70022.
6. M. Maashi, H. Eltayeb, M. M. Asiri, et al., “An Efficient Deep Learning-Based Morphology Aware Hierarchical Mixture of Features for Tuberculosis Screening Using Segmentation of Chest X-Ray Images,” Scientific Reports, vol. 15, p. 42569, 2025, doi: 10.1038/s41598-025-19179-0.
7. U. P. Singh, B. Kumari, E. A. Siddiqui, and A. Tanveer, “An Innovative Approach to Classifying Various Phases of Tuberculosis Using an Improved Spiking Neural Network Applied to X-Ray Images,” in Proceedings of the 6th International Conference on Information Management & Machine Intelligence (ICIMMI ’24), New York, NY, USA: ACM, 2025, Art. no. 36, pp. 1–8, doi: 10.1145/3745812.3745850.
8. K. Gasmi, A. Alanazi, N. Ben Aoun, et al., “Early Tuberculosis Detection via Privacy-Preserving, Adaptive-Weighted Deep Models,” Diagnostics, vol. 16, no. 2, p. 204, 2026, doi: 10.3390/diagnostics16020204.

9. G. Siddharth, A. Ambekar, and N. Jayakumar, “Enhanced CoAtNet Based Hybrid Deep Learning Architecture for Automated Tuberculosis Detection in Human Chest X-Rays,” BMC Medical Imaging, vol. 25, p. 379, 2025, doi: 10.1186/s12880-025-01901-z.
10. S. Gorbachev, A. Dixit, A. Mani, et al., “Hybrid Deep Learning Framework MedFusionNet Assists Multilabel Biomedical Risk Stratification From Imaging and Tabular Data,” Communications Medicine, vol. 6, p. 47, 2026, doi: 10.1038/s43856-025-01334-3.
11. Z. Ding, R. Yaakob, A. Azman, S. N. M. Rum, N. Zakaria, and A. S. A. Nazri, “Artificial Intelligence-Powered Tuberculosis Detection With Complementary Domain Attention Model,” Neurocomputing, vol. 637, p. 130089, 2025, doi: 10.1016/j.neucom.2025.130089.
12. B. H. Shekar and S. Mannan, “Tuberculosis Detection With Customized CNN and Oversampling Techniques: A Deep Learning Approach,” Discover Artificial Intelligence, vol. 5, p. 116, 2025, doi: 10.1007/s44163-025-00329-2.
13. P. Khobragade, P. Ghutke, P. K. Dhankar, P. K. Adakane, A. B. Kotta and N. Purohit, "Personalized Healthcare: Using Predictive Models to Improve Treatment Outcomes for Chronic Diseases," 2025 1st International Conference on Data Science and Intelligent Network Computing (ICDSINC), Raipur, India, 2025, pp. 166-171, doi: 10.1109/ICDSINC66221.2025.11448158.
14. R. Yulvina, S. A. Putra, M. Rizkinia, A. Pujitresnani, E. D. Tenda, R. E. Yunus, D. H. Djumaryo, P. A. Yusuf, and V. Valindria, “Hybrid Vision Transformer and Convolutional Neural Network for Multi-Class and Multi-Label Classification of Tuberculosis Anomalies on Chest X-Ray,” Computers, vol. 13, no. 12, p. 343, 2024, doi: 10.3390/computers13120343.
15. A. Mirugwe, L. Tamale, and J. Nyirenda, “Improving Tuberculosis Detection in Chest X-Ray Images Through Transfer Learning and Deep Learning: Comparative Study of Convolutional Neural Network Architectures,” JMIRx Med, vol. 6, p. e66029, Jul. 2025, doi: 10.2196/66029.
16. I. A. Ahmed, E. M. Senan, H. S. A. Shatnawi, Z. M. Alkhraisha, and M. M. A. Al-Azzam, “Multi-Techniques for Analyzing X-ray Images for Early Detection and Differentiation of Pneumonia and Tuberculosis Based on Hybrid Features,” Diagnostics, vol. 13, no. 4, p. 814, 2023, doi: 10.3390/diagnostics13040814.