IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 October 2025, accepted 5 November 2025, date of publication 5 January 2026, date of current version 9 January 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3650782

RESEARCH ARTICLE

TB-FusionNet: A Multi-Scale Feature Fusion Algorithm With Spatial and Channel Cross-Attention for Tuberculosis Detection

ZEYU DING[1,2], RAZALI YAAKOB[1], KOH TIENG WEI[3], AZREEN BIN AZMAN[1], (Member, IEEE), SITI NURULAIN BINTI MOHD RUM[1], NOR FADHLINA BINTI ZAKARIA[4], AND AZREE SHAHRIL AHMAD NAZRI[1]

[1]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor 43400, Malaysia
[2]Department of Computer Science, Changzhi University, Changzhi 046000, China
[3]Centre for Cyber-Physical Systems, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Malaysia
[4]Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang, Selangor 43400, Malaysia

Corresponding authors: Razali Yaakob (razaliy@upm.edu.my) and Koh Tieng Wei (koh.tiengwei@utp.edu.my)

This work was funded by the Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS/1/2020/ICT02/UPM/02/5), and the Article Processing Charge (APC) for this manuscript is fully supported by Universiti Teknologi PETRONAS (UTP).

ABSTRACT The lesions of tuberculosis (TB) in X-ray images are highly complex, exhibiting a variety of sizes, shapes, and structural variations. Single-scale features are insufficient to fully represent this diversity and complexity, thereby limiting the effectiveness of TB detection. As a result, multi-scale feature fusion has become a widely explored approach in the field of TB detection. However, current multi-scale feature fusion methods still have several limitations. First, the weight allocation in existing methods typically remains at the feature level without extending to local features and channels. This limitation prevents the model from precisely controlling the significance of local features and channels, resulting in coarse feature representations. Second, current methods neglect the contextual information between features at different levels, which further undermines the consistency of the fused features and leads to a lack of semantic coherence. To address the aforementioned issues, this study proposes TB-FusionNet, a multi-scale feature fusion algorithm based on channel and spatial cross-attention mechanisms, for tuberculosis classification. The algorithm first calculates the similarity between local features at different levels to precisely select low-level detail features that are closely related to high-level semantic features, thereby generating a more hierarchical feature representation. Next, the algorithm computes the dependencies between feature channels at different levels, allowing low-level channel features to be appropriately enhanced or suppressed under the guidance of high-level channels, effectively improving the semantic consistency between cross-level features. Through these operations, the model can dynamically adjust the weight distribution of local features and channels according to task requirements, thereby more flexibly adapting to the challenges of complex tasks. Experiments were conducted on the Shenzhen, Montgomery and HSAAS datasets in this study. The results demonstrate that the proposed method outperforms current state-of-the-art approaches, validating its effectiveness and robustness.

INDEX TERMS Tuberculosis, TB, deep learning, multi-scale feature fusion.

I. INTRODUCTION

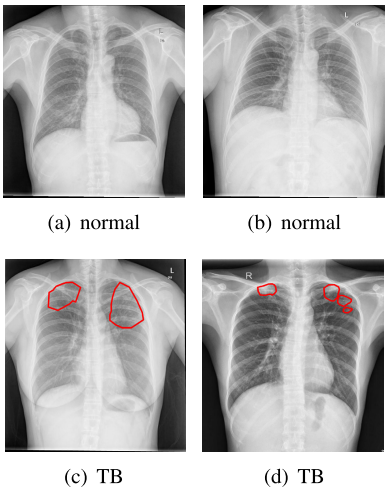
Tuberculosis (TB) remains one of the most common infectious diseases with the highest mortality rate worldwide, especially in areas where medical resources are scarce [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeeb Dey.

Early detection and treatment are essential for controlling TB. However, the diagnosis of TB still relies heavily on medical specialists, which is inefficient and highly subjective. For example, traditional TB diagnosis methods mainly involve sputum culture and microscopic examination. These methods are time-consuming, especially sputum culture, which typically requires several weeks to confirm the presence of Mycobacterium TB. Furthermore, TB diagnosis relies on radiologists' visual assessment of chest X-ray images. Variations in physicians' experience and expertise can result in inconsistent diagnostic outcomes.

In recent years, deep learning techniques have shown great advantages in the field of medical image processing, providing a new solution for the auxiliary diagnosis of TB, thus gaining the favour of researchers. For example, studies [2], [3] have achieved high-precision detection results on multiple public datasets of TB, greatly improving the efficiency of TB detection and demonstrating the potential value as an assistant to medical experts.

Figures 1(a) and 1(b) show normal X-ray images, while Figures 1(c) and 1(d) depict TB cases, with lesions annotated by expert physicians highlighted within the red circles [4]. As shown in Figure 1, TB lesions vary considerably in terms of size, shape, and location. In addition, they lack clear distinguishing features, making it difficult for observers without specialized training to identify the affected areas accurately. This heterogeneity poses substantial challenges for accurate TB detection. Multi-scale feature fusion is a promising approach to address this issue. Multiscale features contain rich information, with lower-level features typically capturing detailed information, such as lesion contours and textures. Higher-level features, on the other hand, include more semantic information, such as the lesion's type and location. Effectively integrating these different levels of information can significantly enhance the ability to identify lesions and improve the robustness of detecting complex and variable lesions. Many studies have also demonstrated the effectiveness of multi-scale feature fusion in medical image processing. For example, in detecting brain MRI lesions, research [5], [6] utilized multi-scale convolutional neural networks to optimize the extraction of multi-scale features, thereby enhancing the model's adaptability to tumor size, shape, and location, improving tumor recognition efficiency. In the detection of lung diseases, research [3], [7], [8] integrated multi-level information from chest X-ray images, enhancing the model's adaptability to various image qualities and complex structures, thereby improving its overall information perception.



(a) normal (b) normal

(c) TB (d) TB

FIGURE 1. Comparison between normal X-ray images and TB images.

Although multi-scale feature fusion methods have shown certain advantages,there are still many aspects that need improvement. Firstly, most of these methods primarily focus on integrating features across different levels without thoroughly exploring the local features within each individual level. Since each level of features contains important local information, assigning the same weight to all local features may lead to the neglect of key features, reducing the model's adaptability when handling complex and variable tasks. Moreover, current multi-scale fusion methods often fail to effectively utilize the contextual information between local features at different levels. Most existing approaches tend to independently extract key local features from each level and then combine them. However, this fusion strategy may not be highly efficient, as it neglects the dependencies between local features across levels. Consequently, it is unable to dynamically allocate the weights of critical features from each level in response to complex and varying task requirements. Finally, current multi-scale fusion methods also overlook the interaction between channels across different levels. Each channel reflects specific semantic information or attributes within the features. Without an effective mechanism to coordinate the relationships between these channels, the consistency of cross-level features may be lost. This limitation restricts the model's ability to dynamically enhance task-relevant key channels while suppressing irrelevant or noisy channels, ultimately reducing the efficiency of feature utilization.

To address the limitations of existing methods, this study proposes TB-FusionNet, a multi-scale feature fusion algorithm that integrates both spatial- and channel-level cross-attention mechanisms for TB detection. The main contributions of this work are as follows:

(1) Channel-level cross-attention: This study is the first to introduce a channel-level cross-attention mechanism for multi-scale feature fusion in TB detection, enabling adaptive enhancement or suppression of low-level channel features under the guidance of high-level semantic information.

(2) Integration with spatial cross-attention: A novel fusion of channel-level and spatial-level cross-attention mechanisms is designed to jointly capture dependencies among local features and contextual relationships across scales, leading to a more comprehensive and hierarchical feature representation.

(3) Improved diagnostic performance: By adaptively adjusting the weight distribution of local features and channels across levels, the proposed method achieves both global awareness and local focus, effectively handling complex lesion morphology, diverse spatial distributions, and scale variations. Extensive experiments on multiple public TB datasets demonstrate that TB-FusionNet consistently outperforms state-of-the-art approaches, validating its effectiveness and generalization ability.

II. RELATED WORKS

A. STUDIES OF MULTI-SCALE FEATURE FUSION IN MEDICAL IMAGE ANALYSIS

Multiscale feature fusion has demonstrated significant advantages in medical image analysis, leading to widespread attention in this field. Some studies directly combine multiscale features through concatenation or addition strategies. For example, Chandra et al. proposed a model for TB detection, in which new feature representations are generated through the addition of features at different scales [9]. The addition method demonstrated significant performance improvements. Similarly, Sarkar et al. employed a concatenation approach for features extracted from CNNs at different scales to detect various lung diseases [10]. This study achieved excellent results in a three-class classification task involving TB, COVID-19, and healthy cases. Although the study did not present binary classification results for tuberculosis and healthy cases, its strong performance in multi-class classification suggests that this feature fusion approach holds considerable potential for binary classification tasks in TB detection. Yazdan et al. proposed a multiscale convolutional neural network (MSCNN) that extracts features using multiscale convolutional kernels and directly combines them, improving the detection of brain lesions [5]. Zhang et al. developed a multiscale residual network (MSRN) that integrates multilayer feature fusion with residual structures, further enhancing the accuracy of benign and malignant lung nodule classification [11]. However, directly adding all features does not account for the differences in the importance of features across scales, which may lead to irrelevant details negatively affecting the results. To address this issue, some studies have proposed weighted fusion methods. Rahman et al. improved the accuracy of breast cancer classification by significantly enhancing feature fusion through a weighted combination of multi-scale features in BreastMultiNet [12]. Li et al. proposed a multi-layer residual feature fusion network (MLRFNet) [8] that combines a channel attention module with multilayer residual structures. The feature fusion component of this model achieves multi-level information integration by performing weighted fusion of residual feature vectors generated from feature maps at different scales. However, the weights assigned to features at each scale are limited to fixed values (0.5, 0.3, and 0.1), which makes the approach heavily reliant on empirical assumptions. Some studies have further proposed adaptive multi-scale feature fusion mechanisms that automatically assign weights to features from different scales. Fan et al. designed an adaptive weighted fusion strategy, which improved the model's performance in classifying diabetic retinopathy at different stages of severity [13]. Xu et al. proposed the EFPN network, which combines a feature pyramid structure and attention mechanism, fusing multi-scale features through the pyramid and adjusting the weights using a feature fusion attention module, thereby enhancing the performance of X-ray and MRI image detection [14]. Huo et al. introduced the HiFuse network, which combines Transformer and CNN architectures, using a multi-branch design to extract both local and global features, and performs weighted fusion through an adaptive hierarchical feature fusion module (HFF), significantly improving classification accuracy across various diseases [15]. Wu et al. proposed the AGGN network, which integrates multi-scale feature extraction, multi-modal information fusion, and attention mechanisms, improving the accuracy of glioma grading [16]. Although these methods can adaptively adjust the weights of features from different scales, they primarily focus on overall scale features and fail to deeply analyze the importance of local details within each scale, resulting in an insufficient integration of local feature information across all scales.

To better assign weights to local features at various scales, researchers have proposed several methods based on attention mechanisms. For example, Chen et al. proposed a method called CrossViT [17], which introduces features at different scales by using image patches of varying sizes as input. Cross-attention is then used to calculate the dependencies between these features, enabling more efficient multi-scale feature fusion and enhancing the hierarchical structure of the features. This approach requires significant computational resources. To address this issue, Wang et al. introduced a Cross-Scale Embedding Layer (CEL) and Long-Short Distance Attention (LSDA), further integrating multi-scale features into the self-attention mechanism [18]. This operation reduces computational overhead and enhances the efficiency of multi-scale feature fusion, thereby improving the model's performance in classification, detection, and segmentation tasks. The study [19] in CTRL-F employs a Multi-Level Feature Cross-Attention (MFCA) module, which fuses local and global features by exchanging information between different convolutional layers. This approach enables the model to achieve excellent robustness in classification tasks. Ates et al. [20] proposed Dual Cross-Attention, which adds a spatial and channel-based cross-attention module to the U-Net architecture. This enables the model to better understand the dependencies between the spatial and channel dimensions of features, enhancing the semantic consistency between the encoder and decoder. As a result, multi-scale features are more accurately represented in medical image segmentation, significantly improving segmentation accuracy.

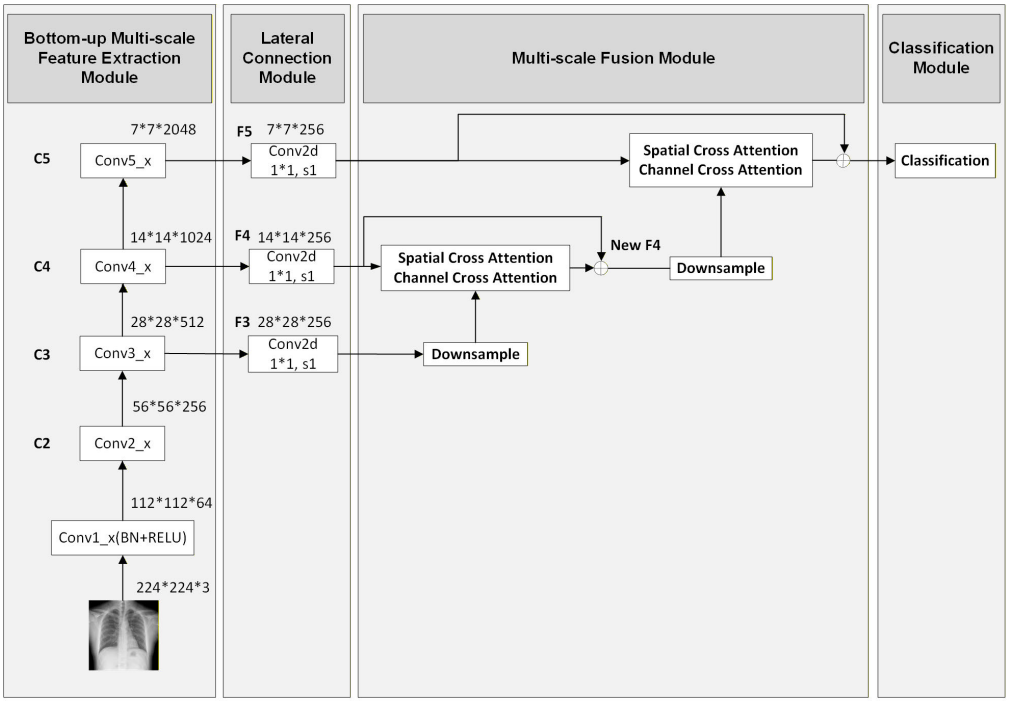


FIGURE 2. The overall architecture of the TB-FusionNet.

B. STATE-OF-THE-ART METHODS FOR TB DETECTION

A variety of state-of-the-art methods for TB detection have been developed and validated in recent years. For instance, the study [21] integrated shape features and edge texture features, which were then input into three different classifiers. A voting mechanism was applied to obtain the final TB detection result. This approach yielded good results on both the Shenzhen and Montgomery datasets. Another research [22] compared the performance of five pre-trained models for TB detection on chest X-ray images. The results indicated that AlexNet performed the best across most datasets. The approach [23] explored the performance of well-known deep convolutional network (DCN) architectures in TB detection under various anomaly conditions. The study found that shallow features, or features from earlier layers, consistently provided higher detection accuracy compared to deeper features. Based on this finding, the research employed an ensemble model for the TB classification task, which demonstrated significant performance improvement compared to single models. However, the model was only validated on the Shenzhen dataset for TB classification. Govindarajan and Swaminathan combined SURF features with traditional CNN-extracted features for TB detection and validated the approach on the Montgomery dataset [24]. The results indicated that this fusion method outperformed the approach using only CNN features. Another study by Ammar integrated the vision transformer with traditional CNNs to create a hybrid model, which significantly enhanced the performance of TB detection [25]. Win et al. combined handcrafted features with CNN-extracted features and employed machine learning algorithms to intelligently select the most relevant features, thus improving TB detection performance [26]. Research [27] identified that a significant portion of chest X-ray (CXR) images is dark, offering little useful information for diagnosis. Based on this, their method used a sophisticated segmentation network to extract the regions of interest from CXR images. These segmented images were then input into a model to perform TB classification. Additionally, the approach utilized explainable methods to visualize the TB-infected areas in the lungs, providing radiologists with diagnostic support. These methods will be compared with the approach proposed in this study in the following sections.

III. METHODS

This paper proposes a multi-scale feature fusion model based on spatial and channel cross-attention for TB classification, as shown in Figure 2. It consists of four components: a bottom-up multi-scale feature extraction module, a lateral connection module, a multi-scale fusion module, and a classification module.

A. BOTTOM-UP MULTI-SCALE FEATURE EXTRACTION MODULE

This section adopts the design approach of Feature Pyramid Networks (FPN) [28]. During the bottom-up multi-scale feature extraction process, ResNet50 is used as the backbone network. A TB X-ray color image with a size of 224×224 is fed into the network. The image is first processed by the Conv1 layer (7×7 convolution kernel with max pooling),

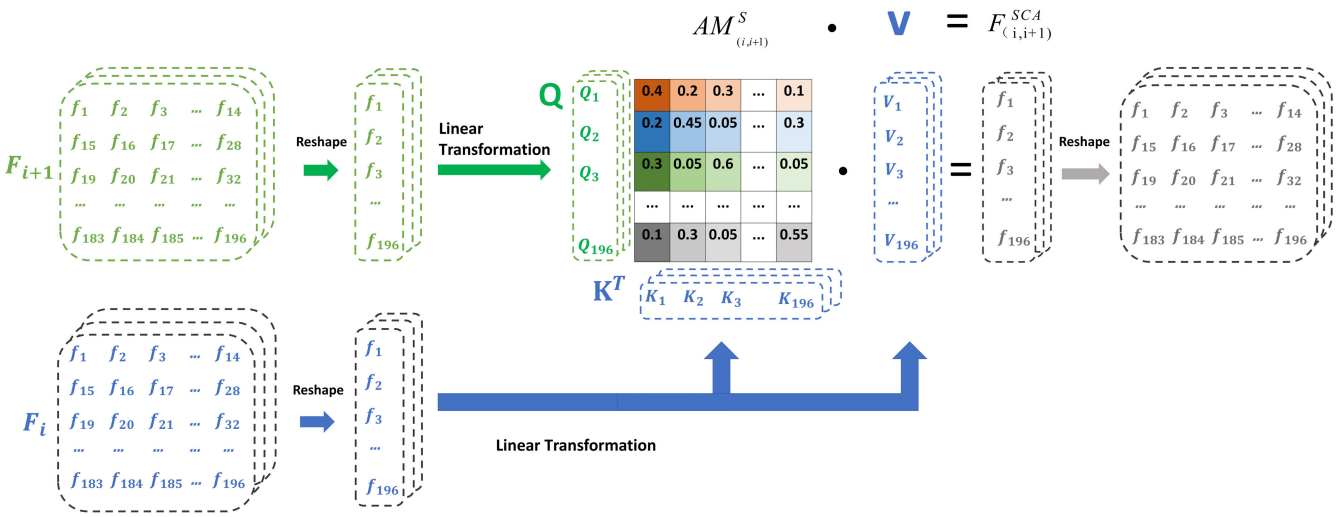


FIGURE 3. The process of spatial cross-attention mechanism.

reducing its size to 112 × 112 while increasing the number of channels to 64. Subsequently, the image passes through Conv2 to Conv5 layers, where feature maps are extracted step by step, with resolution gradually decreasing and the number of channels progressively increasing. The extracted feature maps are C2 (56 × 56 × 256), C3 (28 × 28 × 512), C4 (14 × 14 × 1024), and C5 (7 × 7 × 2048). Feature maps from different layers capture features at different scales. Lower-layer feature maps focus more on detailed information, while higher-layer feature maps emphasize semantic information.

B. LATERAL CONNECTIONS MODULE

The purpose of lateral connections is to align the channels of feature maps at different scales. Since the multi-scale features extracted in the bottom-up process have varying numbers of channels, it is necessary to ensure channel consistency during feature fusion. To achieve this, lateral connections use 1 × 1 convolution kernels to unify the number of channels in each feature map to 256. This ensures that feature maps from different scales remain consistent during fusion. After channel adjustment, the C3, C4, and C5 feature maps are transformed into F3, F4, and F5, respectively.

C. MULTI-SCALE FEATURE FUSION

1) OVERALL PROCESS OF MULTI-SCALE FEATURE FUSION

The fusion process is divided into two steps. First, the F3 and F4 features are fused to generate a new F4 feature. Then, the new F4 feature is fused with the F5 feature to obtain the final feature.Specifically, the fusion process of F3 and F4 features is as follows: First, F3 is downsampled to reduce its size, matching the spatial dimensions of F4 (14 × 14 × 256). Then, F3 and F4 are fused using cross-attention (The cross-attention mechanism includes spatial and channel cross-attention, which will be detailed in subsequent sections). The output of this fusion is added to the original F4 feature, resulting in the new F4 feature. The addition operation is inspired by the design philosophy of ResNet. It helps mitigate the vanishing gradient problem, and if the features learned by the cross-attention mechanism are not optimal, the network can still fall back on the original F4 features. (2) Fusion of new F4 and F5 features: The fusion process of new F4 and F5 features is similar to the fusion operation between F3 and F4, so it will not be described again.

2) SPATIAL CROSS-ATTENTION

After the low-level feature F_i is downsampled, its size matches that of the high-level feature F_{i+1}, denoted as R^{H*W*C}, where C is the number of channels, H is the height, and W is the width. Figure 3 shows the process of feature fusion between F_i and F_{i+1} using spatial cross-attention. For simplification in the figure, both H and W values are set to 14. First, the shapes of F_i and F_{i+1} are transformed into R^{N*C}, where $N = H * W$, and N represents the number of local features. Then, the higher-level feature F_{i+1} undergoes a linear transformation to obtain the vector Q_{i+1} ($Q_{i+1} = F_{i+1} \times W_Q$). Similarly, the lower-level feature F_i is processed in the same manner to generate two vectors, K_i and V_i ($K_i = F_i \times W_K$ and $V_i = F_i \times W_V$), where Q_{i+1}, K_i, and $V_i \in R^{N*C}$, W_K, W_Q and W_V are trainable weight matrices used to map the original features into a new space, enabling more efficient capture of relationships between features. Q_{i+1} is obtained by applying a linear transformation to F_{i+1}, and thus contains information from the high-level features. Similarly, K_i and V_i also contain information from the low-level features. Subsequently, Q_{i+1} is multiplied by the transpose of K_i, and a softmax operation is applied to obtain the attention matrix $AM^S_{(i,i+1)}$. This matrix reflects the correlation between the N local features from the low-level and the N local features from the high-level,

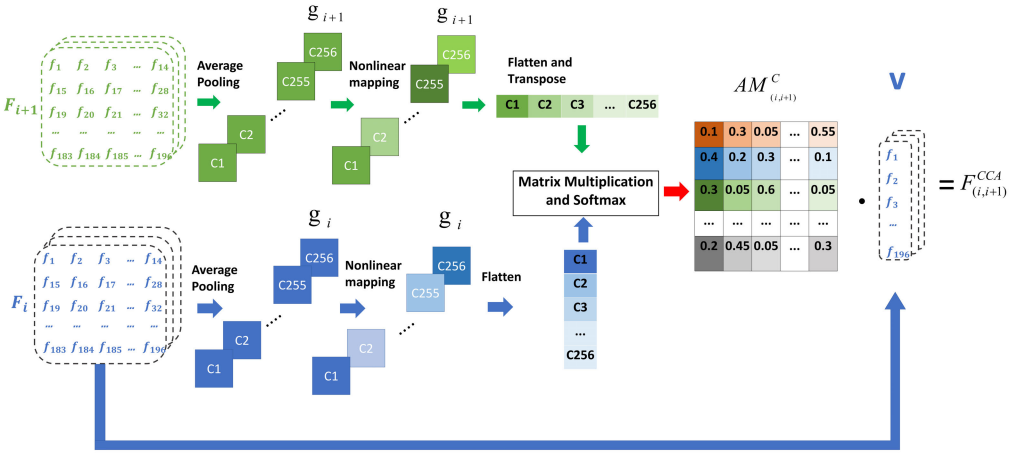


FIGURE 4. The process of channel cross-attention mechanism.

$AM^S_{(i,i+1)} \in R^{N*N}$. The core of the spatial cross-attention mechanism lies in the $AM^S_{(i,i+1)}$. A larger weight indicates a stronger correlation between the corresponding low-level and high-level features. The calculation is given by equation 1.

$$AM^S_{(i,i+1)} = softmax(Q_{i+1} \cdot K_i^T) \tag{1}$$

By multiplying the attention matrix $AM^S_{(i,i+1)}$ with V_i, the spatial cross-attention feature $F^{SCA}_{(i,i+1)}$ can be obtained, where the low-level features V_i are reweighted, thereby enhancing the components correlated with high-level features while suppressing the irrelevant ones. The equation is as follows:

$$F^{SCA}_{(i,i+1)} = AM^S_{(i,i+1)} * V_i \tag{2}$$

$F^{SCA}_{(i,i+1)} \in R^{N*C}$, and it needs to be reshaped into R^{H*W*C}.

By calculating the similarity and dependency between low-level detailed local features and high-level semantic local features, it is possible to effectively filter out the detailed features that are more relevant to the semantic information. This operation enables the model to more accurately integrate contextual information while optimizing interactions between features. It allows low-level detailed features to be reasonably enhanced or suppressed under the guidance of semantic features, ultimately forming a more comprehensive and hierarchical feature representation. This multi-dimensional information fusion approach enhances the model's ability to understand and analyze complex scenes, resulting in more comprehensive and accurate feature representations.

3) CHANNEL CROSS-ATTENTION

Figure 4 illustrates the process of fusing F_i and F_{i+1} using channel cross-attention. For simplicity, the values of H and W in the figure are both set to 14, and the value of channel is set to 256. The first step is to obtain global contextual information. Average pooling is applied to F_i and F_{i+1}, resulting in g_i and g_{i+1}, respectively, as shown in equations 3 and equation 4.

$$g_i = \frac{1}{H*W} \sum_{p=1}^{H} \sum_{q=1}^{W} F_i(p,q) \tag{3}$$

$$g_{i+1} = \frac{1}{H*W} \sum_{p=1}^{H} \sum_{q=1}^{W} F_{i+1}(p,q) \tag{4}$$

Here, p and q represent the pixel at the p-th row and q-th column. The dimensions of g_i and g_{i+1} are R^C.

The second step is to perform nonlinear mapping. Through the fully connected layers FC_1 and FC_2, along with the activation functions ReLU and σ, nonlinearity is added to g_i and g_{i+1}.

$$g_i = \sigma(\mathrm{FC}_2(\mathrm{ReLU}(\mathrm{FC}_1(g_i)))) \tag{5}$$

$$g_{i+1} = \sigma(\mathrm{FC}_2(\mathrm{ReLU}(\mathrm{FC}_1(g_{i+1})))) \tag{6}$$

In this process, FC_1 reduces the dimensionality of the features to $R^{C/8}$, the ReLU activation function introduces nonlinearity, FC_2 restores the dimensions to the original size R^C, and the σ function normalizes the weights.

The third step is the computation of attention map. The transpose of g_{i+1} is multiplied by g_i through matrix multiplication, followed by the application of a softmax layer. This results in the channel attention map $AM^C_{(i,i+1)} \in R^{C*C}$, which reflects the correlation between the channels.

$$AM^C_{(i,i+1)} = softmax(\mathrm{g_{i+1}} \cdot \mathrm{g_i}) \tag{7}$$

The final step is to obtain the channel cross-attention features $F^{CCA}_{(i,i+1)}$. F_i is reshaped into R^{N*C}, and then multiplied with the attention map $AM^C_{(i,i+1)}$ to obtain the channel cross-attention features.

$$F^{CCA}_{(i,i+1)} = F_i \cdot AM^C_{(i,i+1)} \tag{8}$$

Here, $F^{CCA}_{(i,i+1)} \in R^{N*C}$, and it is finally reshaped into R^{H*W*C}.

Channel features represent information extracted by different filters. The relative importance of channels varies depending on the specific task. By computing dependencies between low-level and high-level channels, contextual information can be integrated more precisely. This enhances channel feature interactions and effectively filters low-level channels that are more relevant to high-level semantic channel information, allowing the model to dynamically allocate channel weights based on task requirements. Therefore, channel cross-attention enhances the model's ability to understand and dynamically adjust the relationships between channels, allowing for better fusion of multi-scale features.

The channel cross-attention mechanism presented in this study significantly differs from the channel attention mechanism in CBAM [29]. CBAM merely extracts features via maximum and average pooling and subsequently fuses them. This process does not involve multi-scale feature extraction, nor does it calculate the dependencies between features at different scales. In contrast, while our proposed channel cross-attention mechanism also employs average pooling to extract channel features, its essence lies in computing the dependencies among channel features across different scales and selectively enhancing or suppressing low-level channel features based on these dependencies.

4) FUSION OF SPATIAL AND CHANNEL ATTENTION FEATURES

The final new high-level feature F_{i+1}^{new} is the sum of the original F_{i+1}, the spatial cross-attention feature $F_{(i,i+1)}^{SCA}$ and the channel cross-attention feature $F_{(i,i+1)}^{CCA}$. The equations is as follows:

$$F_{i+1}^{new} = F_{i+1} + F_{(i,i+1)}^{SCA} + F_{(i,i+1)}^{CCA} \tag{9}$$

To alleviate the high memory consumption incurred by computing dependencies across multi-scale features, the spatial and channel attention modules are executed sequentially and subsequently fused at the final stage. Addition is employed as the feature fusion strategy owing to its ability to preserve the hierarchical organization of features, enabling low-level details and high-level semantics to be effectively combined for richer information in classification tasks. Moreover, addition is simple and computationally efficient, helping maintain overall model efficiency. In contrast, multiplicative fusion may cause information loss when feature values approach zero, while concatenation greatly increases dimensionality, leading to heavier computation and potential redundancy. Considering these factors, addition was selected as the fusion method.

D. CLASSIFICATION

After obtaining the final fused features, a fully connected layer followed by a Softmax function is used to map the features to the classification result, which indicates the sample contains TB or is normal. During training, the model is optimized using the weighted cross-entropy loss function.

$$Loss = -\sum_{i=1}^{C} w_i \, y_i \, \log(\hat{y}_i) \tag{10}$$

In equation 10, y denotes the one-hot encoded ground-truth label and $\hat{y}_i$ corresponds to the predicted probability distribution output by the model. C denotes the total number of classes, and w_i denotes the weight assigned to class, which mitigates the effect of data imbalance, and its computation is given in equation 11.

$$w_i = \frac{N}{C \cdot n_i}, \quad i = 1, 2, \ldots, C \tag{11}$$

Here, n_i represents the number of samples in class i, and N refers to the total number of samples in the training set.

TABLE 1. Information of all datasets used in this study.

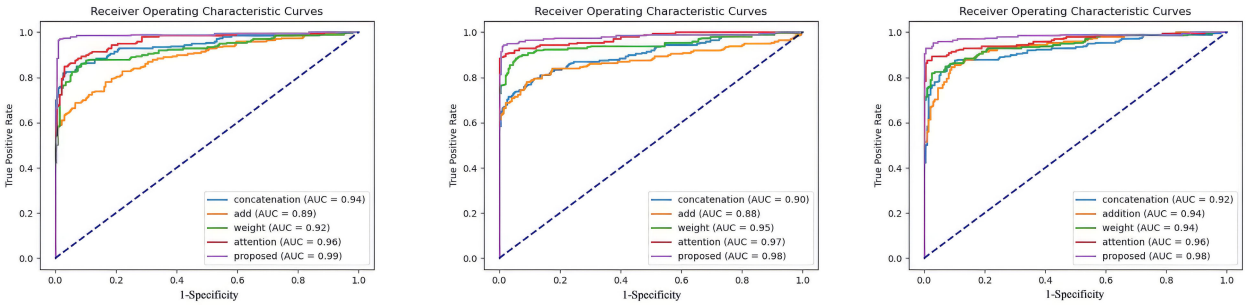
Dataset Name	Sample Size and type	TB Type
Shenzhen	662 cases (326 normal, 336 PTB)	PTB
Montgomery	138 cases (80 normal, 58 PTB)	PTB
HSAAS	447 cases (257 normal, 146 PTB, 44 EPTB)	PTB + EPTB

IV. EXPERIMENT

A. DATASETS

This study utilizes two public TB datasets (Shenzhen [30], Montgomery [30]) and one private TB dataset (HSAAS). The details of these datasets are summarized in Table 1. The Shenzhen dataset, collected at Shenzhen No.3 People's Hospital, contains 662 frontal chest X-rays (326 normal and 336 with TB). The images are high-resolution, mainly in PNG format, and include both adult and pediatric cases with PA or AP views. They cover a wide range of TB manifestations such as nodules, infiltrates, and apical thickening, making the dataset suitable for TB detection and classification research. The Montgomery dataset, provided in collaboration with Montgomery County in the United States, consists of 138 chest X-rays (80 normal and 58 with TB). All images were captured in PA view with large image sizes, ensuring sufficient detail for clinical and computational analysis. Together, these datasets provide complementary characteristics in terms of scale, demographics, and imaging conditions, and have become widely used benchmarks for evaluating computer-aided diagnosis methods in TB research.

The HSAAS dataset was acquired from HOSPITAL SULTAN ABDUL AZIZ SHAH (HSAAS). Annotation of each image was performed by experienced physicians, guided by sputum smear acid-fast bacillus microscopy findings and corroborated with radiographic manifestations on chest X-rays, thereby ensuring precise labeling and providing a reliable foundation for TB detection research. HSAAS consists of 447 images in total, including 257 X-rays of healthy individuals, 146 images of pulmonary tuberculosis (PTB),



(a) Shenzhen (b) Montgomery (c) HSAAS

FIGURE 5. ROC comparison of various multi-scale feature fusion methods.

and 44 images of extrapulmonary tuberculosis (EPTB). To ensure fairness in the binary classification experiments, only 404 images (healthy and PTB) were selected. For experiments detecting EPTB, all 447 samples were used.

The X-ray images across all datasets are of relatively high resolution, and their preprocessing procedures will be described in the next section (Experiment Details). To ensure the robustness and reliability of the experimental results, this study employed a 10-fold cross-validation method on each dataset. Specifically, the dataset is randomly divided into 10 subsets. In each iteration, one subset is used as the test set, while the remaining 9 subsets are used to train the model. This process is repeated 10 times, ensuring that each subset is used as the test set once. Finally, the performance metrics from all 10 experiments are averaged to obtain the model's final evaluation results.

B. EXPERIMENTAL DETAILS

The experiments in this paper were conducted on a computer equipped with an Intel(R) Core(TM) i7-7700K processor, 16GB of memory, and an NVIDIA GeForce RTX 3060 12GB graphics card. The operating system is Ubuntu 20.04.3 LTS, and the experimental environment is based on PyTorch version 1.12.1 and Python version 3.9.

This paper employs the FPN [28] network architecture, with ResNet-50 as the backbone, and initializes the model parameters by loading pre-trained weights. The images are first resized to 256×256 pixels, followed by random rotation and random horizontal flipping for data augmentation. Then, the images are center-cropped to a final input size of 224×224 pixels. In the final preprocessing step, the data were standardized to a zero-mean and unit-variance distribution, a transformation that enhances the stability of model optimization and promotes faster convergence during training. To address class imbalance during training, a weighted cross-entropy loss function was employed. The class weights were determined based on the sample distribution of each dataset. For the binary classification tasks, the weights were as follows: Shenzhen ($w_{normal} = 1.015$, $w_{TB} = 0.985$), Montgomery ($w_{normal} = 0.863$, $w_{TB} = 1.190$), and HSAAS ($w_{normal} = 0.784$, $w_{TB} = 1.380$). For the multi-class classification task, which was conducted only on the HSAAS dataset, the corresponding weights were $w_{normal} = 0.579$, $w_{PTB} = 1.020$, and $w_{EPTB} = 3.386$. The model was optimized using the Adam optimizer, with an initial learning rate of 1e-4, trained for 20 epochs, and the learning rate was reduced by half every 5 epochs.

C. PERFORMANCE EVALUATION METRICS

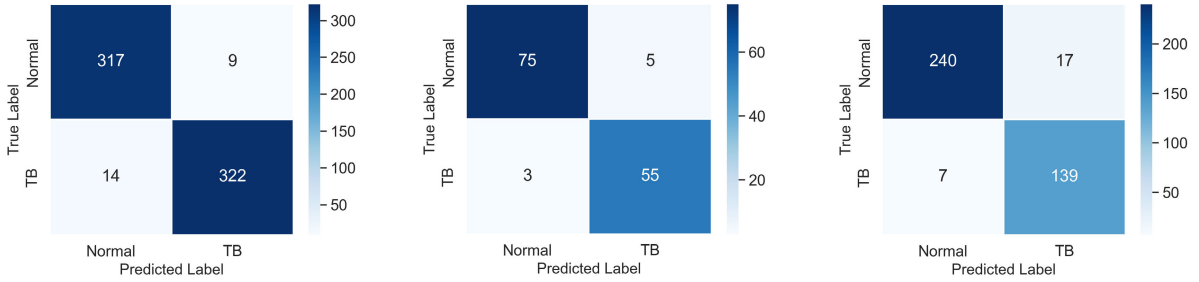
In evaluating the classification performance of the model, this study used several metrics, including the confusion matrix, accuracy, sensitivity, specificity, precision, F1 score, and receiver operating characteristic (ROC) curve. These metrics provide a comprehensive assessment of the model's performance, ensuring an accurate evaluation under various conditions.

V. RESULTS AND ANALYSIS

A. COMPARISON OF VARIOUS FEATURE FUSION METHODS

To compare the effectiveness of different multi-scale feature fusion methods, this study implemented several approaches and evaluated their performance. These methods include feature concatenation [10], addition [9], weighted fusion [13], attention mechanism [17], and TB-FusionNet. Since the fusion strategies in research [9] and [10] are relatively simple, only their feature fusion methods were adopted during the reproduction process, while all other components maintained the structure proposed in this work. In contrast, the weighted fusion method was reproduced by strictly adhering to the model structure of the original study [13]. As [17] has made its source code publicly available, this work directly utilized that code, replacing the dataset with the dataset employed in the present study to perform the corresponding experiments.

After conducting 10-fold experiments, the performance of each model is summarized in Table 2, Table 3 and Table 4. In the header of table, Acc represents accuracy, Sen represents sensitivity, Spe represents specificity, and F1 represents the F1 score. The numbers in the table represent percentages. In the first row of the table, Non



(a) Shenzhen (b) Montgomery (c) HSAAS

FIGURE 6. Confusion matrices of the TB-FusionNet model.

TABLE 2. Performance of various feature fusion methods on Shenzhen dataset.

Method	Acc	Sen	Spe	Pre	F1	AUC
Non	87.4	88.5	86.3	86.3	87.4	90.0
Concatenation	89.7	89.0	90.4	90.0	89.5	92.0
Addition	87.6	88.6	86.6	86.5	87.6	89.0
Weighted fusion	90.3	91.2	89.4	89.3	90.3	94.0
Attention	94.8	95.5	94.1	94.0	94.8	96.0
TB-FusionNet	**96.5**	**97.2**	**95.8**	**95.8**	**96.5**	**99.0**

TABLE 3. Performance of various feature fusion methods on Montgomery dataset.

Method	Acc	Sen	Spe	Pre	F1	AUC
Non	86.4	87.2	85.8	81.7	84.3	88.0
Concatenation	89.8	90.3	89.4	86.1	88.2	90.0
Addition	86.9	85.9	87.6	83.4	84.6	88.0
Weighted fusion	91.9	92.4	91.5	88.8	90.6	95.0
Attention	92.8	93.6	92.2	89.7	91.6	97.0
TB-FusionNet	**94.2**	**94.8**	**93.8**	**91.7**	**93.3**	**98.0**

indicates that the multi-scale feature fusion method was not used, and only the highest-level feature C5 was utilized for classification. From these three tables, it can be observed that: (1) All multi-scale feature fusion methods show an improvement in performance compared to the method using a single-scale feature (Non). (2) The performance of the addition method shows a limited improvement over the Non-method, with the average accuracy across the three datasets increasing by only 0.57%. The concatenation and weighted fusion methods further enhance performance, with average accuracy improvements of 2.8% and 3.8%, respectively. The attention-based method demonstrates a significant improvement in performance, with an average accuracy increase of 5.7%, which strongly highlights the effectiveness of the attention mechanism in multi-scale feature fusion. (3) The proposed TB-FusionNet outperformed other methods across all metrics. Figure 5 shows the ROC curves and AUC values of all multi-scale feature fusion methods, where our model achieved the highest AUC. Figure 6 presents the

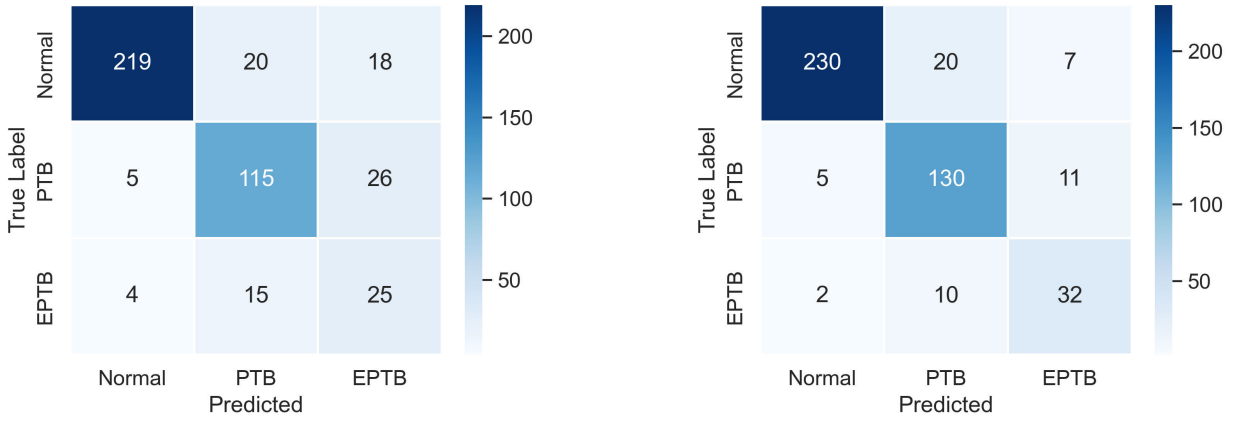
TABLE 4. Performance of various feature fusion methods on HSAAS dataset.

Method	Acc	Sen	Spe	Pre	F1	AUC
Non	88.8	89.2	88.6	81.6	85.2	92.0
Concatenation	91.5	91.0	91.8	86.3	88.6	94.0
Addition	89.8	90.4	88.5	83.0	86.5	92.0
Weighted fusion	91.8	91.5	92.0	86.6	89.0	94.0
Attention	92.1	91.0	92.7	87.7	89.3	96.0
TB-FusionNet	**94.1**	**95.2**	**93.5**	**89.2**	**92.1**	**98.0**

confusion matrix of TB-FusionNet. On the Shenzhen dataset with 662 samples, the model correctly classified 317 healthy and 322 TB cases, with only 9 false positives and 14 false negatives, achieving an accuracy of 96.5%, a sensitivity of 97.2%, and a specificity of 95.8%. On the Montgomery dataset with 138 samples, the model achieved an accuracy of 94.2% and a sensitivity of 94.8%, with 75 true negatives, 55 true positives, 5 false positives, and 3 false negatives. On the HSAAS dataset with 403 samples, the model achieved an accuracy of 94.1% and a sensitivity of 95.2%, with 240 true negatives, 139 true positives, 17 false positives, and 7 false negatives. These results demonstrate the outstanding performance and reliability of the proposed model in the task of TB classification.

B. COMPARISON WITH STATE-OF-THE-ART METHODS

This study also compared the proposed method with other state-of-the-art approaches on the Shenzhen and Montgomery datasets, as shown in Table 5 and Table 6. The results indicate that the TB-FusionNet outperforms other methods in most metrics. However, it lags behind the method [25] in terms of specificity on the Montgomery dataset. In TB screening, high specificity reduces false positives, while high sensitivity reduces the risk of missed diagnoses. However, these two metrics typically have a trade-off, where improving sensitivity often comes at the cost of reduced specificity, and vice versa. Given the highly contagious nature of TB and its potential health hazards, the consequences of missed



(a) Confusion matrix of multi-classification using single-scale features

(b) Confusion matrix of multi-classification using TB-FusionNet

FIGURE 7. Confusion Matrix for Multi-classification With or Without Using TB-FusionNet.

TABLE 5. Performance comparison with other state-of-the-art methods on Shenzhen dataset.

Method	Acc	Sen	Spe	Pre	F1	AUC
Santosh et al. [21]	91.0	90.0	91.0	-	-	96.0
Sivaramakrishnan et al. [22]	85.5	-	-	-	-	92.6
Islam et al. [23]	90.0	88.0	92.0	-	-	94.0
Ammar et al. [25]	96.0	96.0	96.0	96.0	96.0	-
Xie et al. [31]	90.2	85.4	95.1	-	-	94.1
Win et al. [26]	95.5	-	-	-	95.4	99.5
Nafisah et al. [27]	92.4	92.3	92.5	92.3	92.3	98.0
TB-FusionNet	**96.5**	**97.2**	**95.8**	**95.8**	**96.5**	**99.0**

TABLE 6. Performance comparison with other state-of-the-art methods on Montgomery dataset.

Method	Acc	Sen	Spe	Pre	F1	AUC
Santosh et al. [21]	83.0	86.0	81.0	-	-	90.0
Sivaramakrishnan et al. [22]	75.8	-	-	-	-	83.3
Govindarajan et al. [24]	87.8	87.7	85.9	-	-	94.0
Xie et al. [31]	92.6	93.1	92.3	-	-	97.7
Ammar et al. [25]	94.0	93.0	**95.0**	-	-	-
Win et al. [26]	92.7	-	-	-	93.3	99.5
Nafisah et al. [27]	88.9	93.3	83.3	87.5	90.3	92.7
TB-FusionNet	**94.2**	**94.8**	93.8	**91.7**	**93.3**	**98.0**

diagnoses are more severe than false positives. Therefore, sacrificing some specificity is acceptable in TB screening.

C. MULTI-CLASSIFICATION (NORMAL, PTB, AND EPTB)

This study conducted a multi-classification (Normal, PTB, and EPTB) experiment on the HSAAS dataset to evaluate TB-FusionNet's performance. The confusion matrix obtained when using the single-scale features for multi-classification is shown in Figure 7 (a). The average accuracy is 86.9%,

TABLE 7. Performance of multi-classification using single-scale features.

Metric	Normal	PTB	EPTB	Average
Acc	89.5	85.2	86.0	86.9
Pre	96.1	76.7	36.2	69.7
Sen	85.2	78.7	56.8	73.6
F1	90.4	77.7	44.4	70.8
Spe	95.3	88.4	89.1	90.9

TABLE 8. Performance of multi-classification using TB-FusionNet.

Metric	Normal	PTB	EPTB	Average
Acc	92.4	89.8	93.4	91.9
Pre	97.0	81.2	64.0	80.7
Sen	89.5	89.0	72.7	83.7
F1	93.1	84.9	68.0	82.0
Spe	96.3	90.0	95.5	93.9

and the other metrics are shown in the TABLE 7. When TB-FusionNet was applied, its confusion matrix is shown in Figure 7 (b). The average accuracy achieved was 91.9%, which represents an improvement of 5.0% compared to using only single-scale features. Other metrics are provided in Table 8. The results indicate that TB-FusionNet also contributes to improved performance in multi-class classification, demonstrating its effectiveness in distinguishing between PTB and EPTB.

D. ABLATION STUDY

This study conducted ablation experiments to evaluate the contribution of the spatial and channel modules to the cross-attention mechanism. The experimental results are shown in TABLE 9, TABLE 10 and TABLE 11. From the results, it can be observed that both the spatial and channel modules contribute to performance improvement.

TABLE 9. Ablation experiment results on Shenzhen dataset.

Method	Acc	Sen	Spe	Pre	F1	AUC
Non	87.4	88.5	86.3	86.3	87.4	90.0
Spatial	95.3	95.9	94.7	94.6	95.3	97.0
Channel	93.8	94.0	93.6	93.4	93.7	96.0
Spatial+Channel	**96.5**	**97.2**	**95.8**	**95.8**	**96.5**	**99.0**

TABLE 10. Ablation experiment results on Montgomery dataset.

Method	Acc	Sen	Spe	Pre	F1	AUC
Non	86.4	87.2	85.8	81.7	84.3	88.0
Spatial	93.8	94.2	93.5	91.3	92.7	97.0
Channel	93.1	93.5	92.8	90.4	91.9	96.0
Spatial+Channel	**94.2**	**94.8**	**93.8**	**91.7**	**93.3**	**98.0**

TABLE 11. Ablation experiment results on HSAAS dataset.

Method	Acc	Sen	Spe	Pre	F1	AUC
Non	88.8	89.2	88.6	81.6	85.2	92.0
Spatial	93.3	93.7	93.1	88.5	91.0	97.0
Channel	92.4	92.8	92.2	87.1	89.8	96.0
Spatial+Channel	**94.1**	**95.2**	**93.5**	**89.2**	**92.1**	**98.0**

The spatial module improves the average accuracy across the three datasets by 6.6%, while the channel module results in an accuracy increase of 5.6%. When the spatial and channel cross-attention modules are combined, the overall performance reaches its best, with an average accuracy improvement of 7.4%. The performance reaches its highest when the spatial and channel modules are combined.

VI. DISCUSSION

Multi-scale features contain rich semantic and detailed information, and the fusion strategy affects the efficiency of information combination, ultimately determining the performance of tuberculosis classification. In this study, various methods for multi-scale feature fusion were experimentally compared, and an attempt was made to explain the underlying principles. Combining multi-scale features in a straightforward manner, such as addition or concatenation, can enhance the richness of the information. Compared to using single-scale features alone, this method can improve TB detection performance. However, it does not make efficient use of the available information. If the model assigns equal weights to features at all levels, it may suppress critical features and introduce redundancy, which can degrade the model's performance. Allocating weights based on the varying importance of features at different scales is a potential solution. Experimental results also demonstrate that the weighted fusion of multi-scale features outperforms simple feature fusion methods. However, this weighted summation fusion approach operates only at the level of complete features, without addressing local regions or channels within each feature layer. The attention mechanism can address this limitation by enhancing features from key regions while suppressing those from less significant areas, thereby further improving the effectiveness of TB detection. TB-FusionNet demonstrates superior utilization of contextual information during multi-scale feature fusion. By computing both spatial and channel dependencies between features at different scales, it dynamically incorporates features from key regions and important channels of lower layers into higher-level features. This enhances feature correlation and overall perceptual capability, enabling our proposed model to surpass other state-of-the-art models. Experimental results demonstrate that TB-FusionNet achieved significant advantages in both accuracy and robustness, providing a more powerful tool for TB detection.

Although the TB-FusionNet demonstrates significant effectiveness in TB detection tasks, it still has certain limitations. The cross-attention mechanism employed in the model requires computing the spatial and channel dependencies of features at different levels, which demands substantial computational resources. This limitation restricts the model's applicability in resource-constrained regions. Future work should focus on further optimization and the exploration of more efficient algorithms. Furthermore, TB-FusionNet has only been validated on relatively small-scale datasets, without verification of its generalization capability on larger, more diverse datasets. Future work will involve collecting larger and more diverse datasets that encompass samples from different types and sources, to evaluate the model's adaptability across various scenarios. Additionally, while the attention mechanism can focus on key regions, it may exhibit instability when processing images with subtle or complex lesion features, potentially introducing interference. Future improvements will consider implementing more refined attention mechanisms and incorporating techniques such as image enhancement to improve the model's robustness.

VII. CONCLUSION

This study proposes TB-FusionNet, a multi-scale feature fusion algorithm based on channel and spatial cross-attention mechanism, for TB detection. The model employs a cross-attention mechanism that effectively utilizes the contextual information from multi-scale features. By calculating the dependencies between different levels of spatial and channel dimensions, it dynamically assigns appropriate weights, enabling higher-level features to better integrate with the corresponding low-level detailed features. This enhances the model's ability to represent TB-related features. Experimental results show that TB-FusionNet outperforms traditional multi-scale feature fusion approaches, including concatenation, addition, weighted fusion, and attention mechanisms. Additionally, this method surpasses the current state-of-the-art methods in TB detection. The method achieved an accuracy of 96.5% on the Shenzhen dataset, 94.2% on the Montgomery dataset, and 94.1% on the HSAAS dataset., setting new performance benchmarks in the field. However, this method also has some limitations. For example, its computational complexity is relatively high,

and it may exhibit some instability when handling data with significant distribution differences. Future research could explore areas such as reducing model complexity, improving cross-domain adaptability, and investigating multi-task joint learning. In conclusion, this study provides a novel and effective solution for AI-assisted TB detection, contributing technical support to global TB prevention and control efforts. It also lays the foundation for further research on the application of deep learning in medical image processing.

REFERENCES

[1] S. Bagcchi, "Who's global tuberculosis report 2022," *Lancet Microbe*, vol. 4, no. 1, p. 20, 2023.
[2] V. Sharma, Nillmani, S. K. Gupta, and K. K. Shukla, "Deep learning models for tuberculosis detection and infected region visualization in chest X-ray images," *Intell. Med.*, vol. 4, no. 2, pp. 104–113, May 2024.
[3] V. Ravi, V. Acharya, and M. Alazab, "A multichannel EfficientNet deep learning-based stacking ensemble approach for lung disease detection using chest X-ray images," *Cluster Comput.*, vol. 26, no. 2, pp. 1181–1203, Apr. 2023.
[4] F. Yang, P. X. Lu, M. Deng, Y. X. J. Wáng, S. Rajaraman, Z. Xue, L. R. Folio, S. K. Antani, and S. Jaeger, "Annotations of lung abnormalities in the Shenzhen chest X-ray dataset for computer-aided screening of pulmonary diseases," *Data*, vol. 7, no. 7, p. 95, Jul. 2022.
[5] S. A. Yazdan, R. Ahmad, N. Iqbal, A. Rizwan, A. N. Khan, and D.-H. Kim, "An efficient multi-scale convolutional neural network based multi-class brain MRI classification for SaMD," *Tomography*, vol. 8, no. 4, pp. 1905–1927, Jul. 2022.
[6] T. Agrawal, P. Choudhary, A. Shankar, P. Singh, and M. Diwakar, "MultiFeNet: Multi-scale feature scaling in deep neural network for the brain tumour classification in MRI images," *Int. J. Imag. Syst. Technol.*, vol. 34, no. 1, p. 22956, Jan. 2024.
[7] Z. Liang, H. Lu, R. Zhou, Y. Yao, and W. Zhu, "CMFuse: Correlation-based multi-scale feature fusion network for the detection of COVID-19 from chest X-ray images," *Multimedia Tools Appl.*, vol. 83, no. 16, pp. 49285–49300, Oct. 2023.
[8] Q. Li, Y. Lai, M. J. Adamu, L. Qu, J. Nie, and W. Nie, "Multi-level residual feature fusion network for thoracic disease classification in chest X-ray images," *IEEE Access*, vol. 11, pp. 40988–41002, 2023.
[9] L. Chandra, M. Atulkar, and P. Tripathi, "Fusion of local and global texture descriptors for improved tuberculosis detection in chest X-ray images," in *Proc. 15th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Jun. 2024, pp. 1–7.
[10] O. Sarkar, M. R. Islam, M. K. Syfullah, M. T. Islam, M. F. Ahamed, M. Ahsan, and J. Haider, "Multi-scale CNN: An explainable AI-integrated unique deep learning framework for lung-affected disease classification," *Technologies*, vol. 11, no. 5, p. 134, Sep. 2023.
[11] G. Zhang, D. Zhu, X. Liu, M. Chen, L. Itti, Y. Luo, and J. Lu, "Multi-scale pulmonary nodule classification with deep feature fusion via residual network," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 11, pp. 14829–14840, Nov. 2023.
[12] M. M. Rahman, M. S. I. Khan, and H. M. H. Babu, "BreastMultiNet: A multi-scale feature fusion method using deep neural network to detect breast cancer," *Array*, vol. 16, Dec. 2022, Art. no. 100256.
[13] R. Fan, Y. Liu, and R. Zhang, "Multi-scale feature fusion with adaptive weighting for diabetic retinopathy severity classification," *Electronics*, vol. 10, no. 12, p. 1369, Jun. 2021.
[14] Z. Xu, X. Zhang, H. Zhang, Y. Liu, Y. Zhan, and T. Lukasiewicz, "EFPN: Effective medical image detection using feature pyramid fusion enhancement," *Comput. Biol. Med.*, vol. 163, Sep. 2023, Art. no. 107149.
[15] X. Huo, G. Sun, S. Tian, Y. Wang, L. Yu, J. Long, W. Zhang, and A. Li, "HiFuse: Hierarchical multi-scale feature fusion network for medical image classification," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105534.
[16] P. Wu, Z. Wang, B. Zheng, H. Li, F. E. Alsaadi, and N. Zeng, "AGGN: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion," *Comput. Biol. Med.*, vol. 152, Jan. 2023, Art. no. 106457.
[17] C.-F.-R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 347–366.
[18] W. Wang, W. Chen, Q. Qiu, L. Chen, B. Wu, B. Lin, X. He, and W. Liu, "CrossFormer++: A versatile vision transformer hinging on cross-scale attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3123–3136, May 2024.
[19] H. S. EL-Assiouti, H. El-Saadawy, M. N. Al-Berry, and M. F. Tolba, "CTRL-F: Pairing convolution with transformer for image classification via multi-level feature cross-attention and representation learning fusion," 2024, *arXiv:2407.06673*.
[20] G. C. Ates, P. Mohan, and E. Celik, "Dual cross-attention for medical image segmentation," *Eng. Appl. Artif. Intell.*, vol. 126, Nov. 2023, Art. no. 107139.
[21] K. C. Santosh and S. Antani, "Automated chest X-ray screening: Can lung region symmetry help detect pulmonary abnormalities?" *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1168–1177, May 2018.
[22] S. Rajaraman, S. Antani, S. Candemir, Z. Xue, G. R. Thoma, P. O. Alderson, J. Abuya, and M. Kohli, "Comparing deep learning models for population screening using chest radiography," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575. Bellingham, WA, USA: SPIE, 2018, p. 49.
[23] M. T. Islam, M. A. Aowal, A. T. Minhaz, and K. Ashraf, "Abnormality detection and localization in chest X-rays using deep convolutional neural networks," 2017, *arXiv:1705.09850*.
[24] S. Govindarajan and R. Swaminathan, "Analysis of tuberculosis in chest radiographs for computerized diagnosis using bag of keypoint features," *J. Med. Syst.*, vol. 43, no. 4, p. 87, Apr. 2019.
[25] L. B. Ammar, K. Gasmi, and I. B. Ltaifa, "ViT-TB: Ensemble learning based ViT model for tuberculosis recognition," *Cybern. Syst.*, vol. 55, no. 3, pp. 634–653, Apr. 2024.
[26] K. Y. Win, N. Maneerat, K. Hamamoto, and S. Sreng, "Hybrid learning of hand-crafted and deep-activated features using particle swarm optimization and optimized support vector machine for tuberculosis screening," *Appl. Sci.*, vol. 10, no. 17, p. 5749, Aug. 2020.
[27] S. I. Nafisah and G. Muhammad, "Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence," *Neural Comput. Appl.*, vol. 36, no. 1, pp. 111–131, Jan. 2024.
[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[29] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11211, Sep. 2018, pp. 3–19.
[30] S. Jaeger, S. Candemir, S. Antani, Y. Wáng, P.-X. Lu, and G. R. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantum Imag. Med. surgery*, vol. 4, no. 6, pp. 475–7, 2014.
[31] Y. Xie, Z. Wu, X. Han, H. Wang, Y. Wu, L. Cui, J. Feng, Z. Zhu, and Z. Chen, "Computer-aided system for the detection of multicategory pulmonary tuberculosis in radiographs," *J. Healthcare Eng.*, vol. 2020, pp. 1–12, Aug. 2020.

ZEYU DING received the bachelor's degree in electronic information engineering from Hefei University of Technology, China, in 2013, and the master's degree in control engineering from the University of Science and Technology of China, in 2016. He is currently pursuing the Ph.D. degree in intelligent systems with Universiti Putra Malaysia. His research interests include artificial intelligence and medical image processing.

RAZALI YAAKOB received the bachelor's and master's degrees in computer science from Universiti Putra Malaysia, in 1996 and 1999, respectively, and the Ph.D. degree from the University of Nottingham, U.K., in 2008. He is currently an Associate Professor with the Faculty of Computer Science and IT, Universiti Putra Malaysia. His areas of research interests include artificial neural networks, pattern recognition, and evolutionary computation in game playing. He is a member of the Faculty of Intelligent Computing Group.

KOH TIENG WEI received the Bachelor of Computer Science (Hons.), Master of Science, and Ph.D. degrees in software engineering from Universiti Putra Malaysia (UPM), in 2004, 2007, and 2012, respectively. He is currently an Associate Professor with the Department of Computing, Faculty of Science, Management and Computing, Universiti Teknologi PETRONAS (UTP), where he is also the Head of the Centre for Cyber-Physical Systems. He has published extensively in academic journals and international conferences, and has led multiple research projects funded by industry and government agencies. His research interests include software engineering, intelligent automation systems, smart manufacturing technologies, data-driven process optimization, and sustainable industrial solutions.

AZREEN BIN AZMAN (Member, IEEE) received the Diploma degree in software engineering from the Institute of Telecommunication and Information Technology, in 1997, the Bachelor of Information Technology degree in information systems engineering from Multimedia University, Malaysia, in 1999, and the Ph.D. degree in computing science, specializing in information retrieval from the University of Glasgow, Scotland, in September 2007. Before joining his Ph.D. degree, he was with industry for few years. He is currently an Associate Professor with University Putra Malaysia. His current research interests include information retrieval, text mining, natural language processing, and intelligent systems. He serves as a Committee Member for Malaysian Society of Information Retrieval and Knowledge Management (PECAMP) and Malaysian Information Technology Society (MITS).

SITI NURULAIN BINTI MOHD RUM received the bachelor's degree from Universiti Teknologi Malaysia (UTM) and the master's and Ph.D. degrees in computer science from the University of Malaya (UM), in 2012 and 2017, respectively. She was an IT practitioner for 15 years. She is currently a Senior Lecturer with the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM). She has published a number of academic journals and presented her research work at international conferences. Her research interests include artificial intelligence, database processing, data science analytic, semantic web, and social media analytics.

NOR FADHLINA BINTI ZAKARIA received the Master of Medicine and Doctor of Medicine degrees from Universiti Kebangsaan Malaysia, in 2010 and 2014, respectively. Her research interests include internal medicine and nephrology.

AZREE SHAHRIL AHMAD NAZRI received the Ph.D. degree in mathematics from the University of Cambridge, U.K. He is currently a Senior Lecturer with University Putra Malaysia. His research interest includes the development of artificial general intelligence.