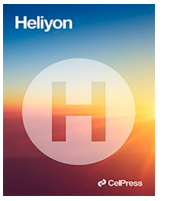
Heliyon 10 (2024) e38088

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Research on improved YOLOv8s model for detecting mycobacterium tuberculosis

Hao Chen [a,*], Wenye Gu [b], Haifei Zhang [a], Yuwei Yang [a], Lanmei Qian [a]

[a] School of Information Engineering, Nantong Institute of Technology, Nantong, 226002, China
[b] Affiliated Hospital of Nantong University, Nantong, 226007, China

ARTICLE INFO

Keywords:
M. tuberculosis
YOLOv8s
Multi-scale feature fusion
CA

ABSTRACT

Accurate identification of Mycobacterium tuberculosis (M. tuberculosis) is a critical step in the diagnosis of tuberculosis. Existing object detection methods struggle with the challenges posed by the varied morphology and size of M. tuberculosis in sputum smear images, which makes precise targeting difficult. To solve these problems, an improved YOLOv8s model is proposed. Specifically, an additional detection head is added to focus on small target information. Second, a multi-scale feature fusion module is introduced to adapt the model to different sizes of M. tuberculosis. In addition, a convolutional layer is added to the Coordinate Attention (CA) module to extract more advanced semantic features. Finally, a self-attention mechanism is added after the CA module to enhance the model's ability to accurately understand and localize the varied morphology of M. tuberculosis. Our model performed well with an average precision of 85.7 % when tested on a publicly available dataset. This clearly demonstrates the effectiveness of our proposed model in M. tuberculosis detection.

1. Introduction

Tuberculosis is a chronic infectious disease caused by Mycobacterium Tuberculosis (M. tuberculosis) [1], which can invade various organs of the human body. When the human body is infected with M. tuberculosis, the disease may be triggered by a decrease in resistance. Because tuberculosis poses a serious threat to human health, accurate diagnosis is critical.

In the field of tuberculosis detection, two methods are commonly used: X-rays observation of the lungs [2,3] and sputum smear for detection of M. tuberculosis [4]. X-rays have an important application in tuberculosis screening. However, although X-rays can detect lung abnormalities such as tuberculosis focus, they cannot directly visualize M. tuberculosis. In contrast, sputum smear images clearly show the morphology and distribution of M. tuberculosis and provide a direct means of detecting the bacteria. Therefore, to detect M. tuberculosis more accurately, this paper favors the use of sputum smear images as the detection object.

Sputum smear microscopy has become a common tool for the detection of M. tuberculosis due to its noninvasive nature and low cost [5]. However, M. tuberculosis images on sputum smears are small and morphologically variable, making accurate identification and localization very difficult. The traditional acid-fast staining method [6] is fast and accurate, but it requires professional experience and is costly.

To overcome the limitations of traditional detection methods, machine learning methods have been gradually introduced into the field of automated detection of M. tuberculosis. These methods can automatically learn features from the data and have achieved some

* Corresponding author.
E-mail address: chenhao@ntit.edu.cn (H. Chen).

https://doi.org/10.1016/j.heliyon.2024.e38088
Received 8 January 2024; Received in revised form 16 September 2024; Accepted 17 September 2024
Available online 18 September 2024

success in this field [7]. However, machine learning tends to show low robustness and accuracy when dealing with the varied morphology and size of M. tuberculosis.

The rapid development of deep learning technology has led many researchers to focus on its application in the automatic detection of M. tuberculosis. Relevant studies have shown that the detection performance of deep learning methods outperforms that of machine learning methods [8,9]. Currently, advanced transformer-based methods offer powerful global modeling capabilities, opening new possibilities for pathology image analysis [10]. However, transformer models for tuberculosis detection face high computational demands and suboptimal object localization, which can pose challenges in practical use.

Convolutional neural networks (CNN) have been the cornerstone of the target detection field. Faster R-CNN [11], Single Shot MultiBox Detector (SSD) [12], and You Only Look Once (YOLO) [13] are CNN-based target detection models. Among them, the YOLO model has a relatively simple structure and fast processing speed, which makes it advantageous in M. tuberculosis detection [14]. The YOLO model has gone through several iterations, with each version continuously optimized to improve performance [15–17]. However, the morphological diversity of M. tuberculosis places higher demands on the detection algorithms. Additionally, M. tuberculosis targets are usually small, which also poses a challenge for YOLO detection. Because these small targets occupy only a few pixels in an image, they are easily ignored. To address these challenges, the researchers optimized in the YOLO model. For example, a multi-scale feature fusion mechanism can improve the detection of small targets like M. tuberculosis by capturing different scale information [18]. However, the morphological variations of M. tuberculosis are extremely complex. Efficiently acquiring its multi-scale features remains a challenge. Additionally, avoiding the loss of targets is an issue that requires further exploration in current research.

Given the excellent performance of YOLOv8s in target detection tasks [19], it is chosen as the basic network for detecting M. tuberculosis in sputum smear images. This paper combines attention mechanisms and a multi-scale feature fusion module. This enables the model to adapt to M. tuberculosis of varied morphology and size. The following are the most important contributions to this paper.

(1) The original large target detection head is deleted and a small target detection head is added to better deal with the features of small targets. Additionally, to better adapt to different sizes of M. tuberculosis, a multi-scale feature detection method is introduced to provide richer contextual information for the model.
(2) An improved attention mechanism is proposed to address the problem of unclear spatial and positional information of M. tuberculosis. By adding a convolutional layer to the Coordinate Attention (CA), the model can analyze various parts of the image in greater detail. Additionally, the introduction of the self-attention mechanism further enhances the model's expressive power. This allows the model to understand and localize the varied morphology of M. tuberculosis.
(3) To assess the effectiveness of our proposed model, experiments are conducted on public datasets. The experimental results show that our proposed model achieves a significant improvement in detection accuracy compared to existing advanced algorithms.

The rest of the paper is organized as follows: Section 2 describes related work. Section 3 describes the design of the model structure proposed in this paper. Experimental results and discussion are given in Section 4. Section 5 provides the conclusion.

2. Related work

Table 1 illustrates a summary of related work. The traditional acid-fast staining methods have demonstrated rapidity and accuracy in the detection of M. tuberculosis, but their main disadvantages are the need for specialized operational experience and the high cost. To address this problem, machine learning methods are gradually being introduced for the automated detection of M. tuberculosis. Ayas et al. [20] proposed an improved Random Forest (RF) method for automatic classification of M. tuberculosis. Costa et al. [21] selected 30 features from four different color spaces and then used a rule-based filter to separate the bacilli, and finally used Support Vector Machine (SVM) for classification. Mithra et al. [22] used a threshold method to obtain segmentation results and then extracted key features such as length and density. Finally, they used an improved decision tree to classify M. tuberculosis. Xu et al. [23] used Logistic Regression (LR), RF, and SVM to classify M. tuberculosis and evaluated their classification effectiveness. All of these methods demonstrate the effectiveness of machine learning methods for automatic classification of M. tuberculosis. However, when dealing with the varied morphology and size of M. tuberculosis, machine learning methods tend to have low robustness and accuracy.

In recent years, the development of deep learning has provided new ideas for the detection of M. tuberculosis. Among them, transformer-based technology is widely used in the analysis of pathology images. For example, Huang et al. [10] utilized a transformer with unlearned parameter attention to achieve interpretable grading of laryngeal squamous cell carcinoma. Wang et al. [24]

Table 1
Related work summary.

Method	Reference	Advantage	Limitation
Traditional	[6]	Fast and accurate.	Relying on professional expertise, high cost.
Machine learning	[20–23]	Automated feature extraction for effective classification.	Low robustness when dealing with morphologically variable M. tuberculosis.
Deep learning	[10,24,25]	Used for pathological image classification, addressing class imbalance issues, and providing interpretable grading.	High computational cost, suboptimal object localization.
	[26–29]	Attention mechanism and multi-scale feature modules improve small	Challenges remain when dealing with M.
	[30–33]	object detection accuracy.	tuberculosis in its various forms.

effectively solved the problem of category imbalance and high similarity in lung adenocarcinoma classification. Huang et al. [25] integrated multiple adversarial multimodal learning to improve the grading performance of laryngeal histopathology images. Although they perform well in some tasks, their application in M. tuberculosis detection is limited due to high computational costs and suboptimal object localization. YOLO-based models, on the other hand, are widely used for the detection of M. tuberculosis due to their ability to process medical images. An et al. [26] successfully detected M. tuberculosis from sputum smear images using the YOLO algorithm, which significantly improved the diagnostic accuracy and reduced the possibility of misdiagnosis. However, because the visual representation of M. tuberculosis varies greatly at different scales. Guo et al. [27] added a multi-sensory field module to YOLO, which achieved multi-scale feature fusion and effectively enhanced the ability to extract information from the deep feature layer. Identifying M. tuberculosis target regions is a major challenge when processing sputum smear images due to the varied morphology of M. tuberculosis. It has been shown that the introduction of an attention mechanism enables the model to better focus on critical regions in the image [28–30]. Adding an attention mechanism module to YOLO has been shown to effectively improve the detection accuracy of tuberculosis [31]. Lv et al. [32] added a convolutional block attention module to the feature pyramid network in YOLO to improve the ability to extract M. tuberculosis features. Li et al. [33] proposed an improved YOLO model to address the characteristics of complex image backgrounds and small M. tuberculosis targets. The model combines the self-attention mechanism and multi-scale feature fusion method, which effectively improves the accuracy of target detection. However, the irregular distribution of M. tuberculosis in sputum smear images leads to unclear spatial structure and location information.

In summary, this paper proposes a YOLOv8s-based detection method for M. tuberculosis. The method mainly focuses on the two problems of M. tuberculosis with varied morphology and size. By introducing multi-scale fusion and attention mechanism techniques, it aims to improve the accuracy of M. tuberculosis detection.

3. Method

Fig. 1 shows the network structure of the model designed in this paper. ECCSA represents enhanced convolutional coordinate self-attention, CBS represents convolutional-batchnorm-silu, CSP represents cross stage partial, and SPPF represents spatial pyramid pooling fast. MSFF, which stands for multi-scale feature fusion, also encompasses the addition of a small target detection head and the

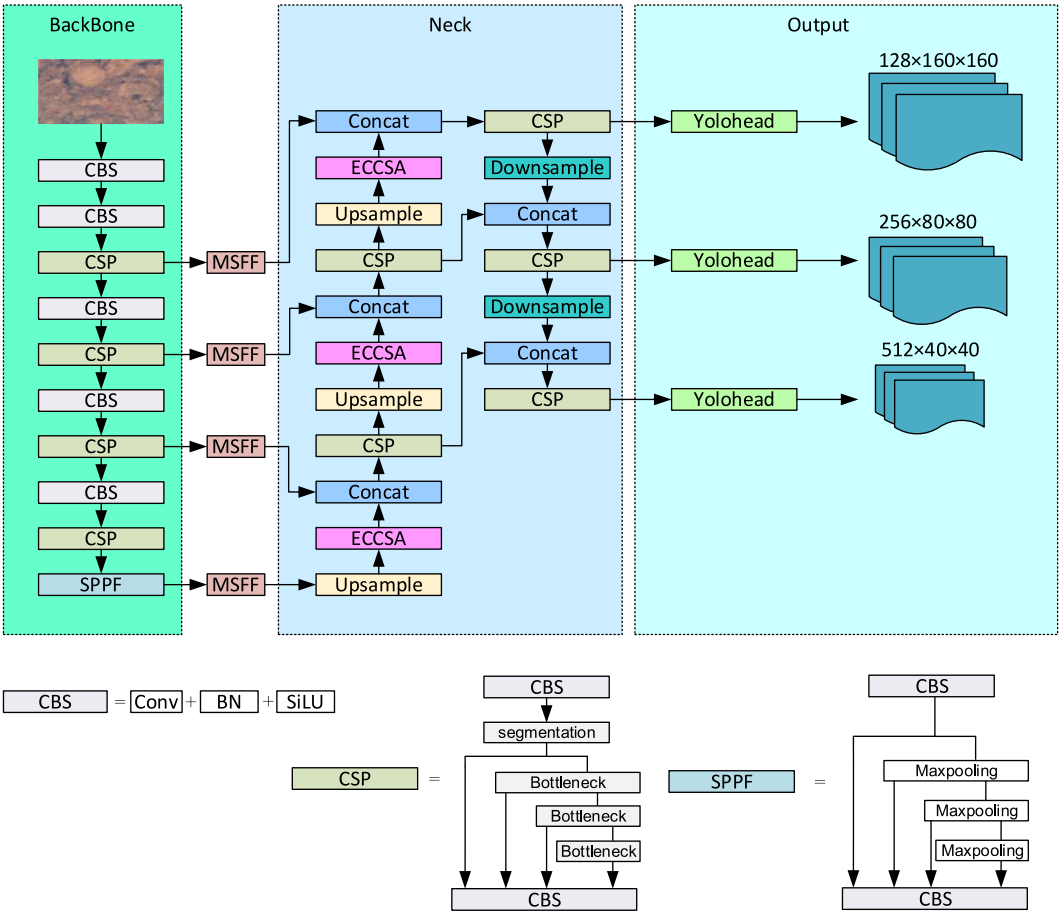


Fig. 1. Model network structure.

removal of a large target detection head. In subsequent ablation experiments, these components are collectively referred to as MSFF. Firstly, the input image is fed into the backbone network to extract features. At this stage, four feature layers are extracted. To further obtain the multi-scale features of M. tuberculosis, these four feature layers are processed by the multi-scale feature detection module. Next, the Neck enhances feature extraction for the model. To enhance the representative of the model, an attention mechanism is introduced after each up sampling. This mechanism enhances the accuracy of the model in locating the position of M. tuberculosis. Finally, three enhanced effective feature layers are obtained in the output section of the model. The Head is responsible for predicting the location and category of M. tuberculosis.

3.1. MSFF

In M. tuberculosis detection, the targets are usually small. Shallower feature maps tend to contain more detailed information about the target, which is crucial for small targets. As a result, the YOLOv8s model is adjusted to account for the specificity of M. tuberculosis detection. The original large target detection head may not be suitable for such small target detection. Therefore, the original large target detection head is removed, and a small target detection head is added to better capture the features of M. tuberculosis. Furthermore, the feature map with a scale of 20 × 20, while providing a large receptive field, may be too large for detecting small targets. As a result, we decided to remove the feature map of this size.

The varied size of M. tuberculosis makes accurate detection challenging. To address this challenge, a multi-scale feature fusion module is introduced after the backbone network. Fig. 2 illustrates the structure of multi-scale feature fusion. The module is designed with three convolution kernels of different sizes: 1x1, 3x3, and 5x5. These convolution kernels extract features at different scales from the input image, thus enhancing the model's ability to recognize M. tuberculosis at different scales. Suppose f'_{in} is the input feature map, and the outputs obtained after three convolutional layers are f_{t1}, f_{t2}, and f_{t3}. f_{t1}, f_{t2}, and f_{t3} can be obtained from Equation (1), Equation (2), and Equation (3).

$$f_{t1} = \mathrm{ReLU}\left(\mathrm{Conv}\left(f'_{in}, r=1\right)\right) \tag{1}$$

$$f_{t2} = \mathrm{ReLU}\left(\mathrm{Conv}\left(f'_{in}, r=3\right)\right) \tag{2}$$

$$f_{t3} = \mathrm{ReLU}\left(\mathrm{Conv}\left(f'_{in}, r=5\right)\right) \tag{3}$$

where ReLU represents Rectified Linear Unit, and Conv represents the convolution operation. Features from different scales of convolutional layers are then spliced in the channel dimension. This splicing helps the model to combine features from various scales, thus improving the accuracy of detection. The feature map f_0 obtained after splicing can be represented by Equation (4).

$$f_0 = f_{t1} + f_{t2} + f_{t3} \tag{4}$$

However, due to the use of different sized convolutional kernels, the output feature maps may have different channel counts. To solve this problem, a 1x1 convolutional layer is introduced for channel number adjustment. The resulting feature map f'_{out} can be expressed by Equation (5).

$$f'_{out} = \mathrm{Conv}\left(f'_{in}, r=1\right) \tag{5}$$

In the YOLOv8s model, the backbone feature extraction network is primarily responsible for extracting basic features. MSFF serves as a powerful complement to the backbone network, further enhancing the detection performance of the model.

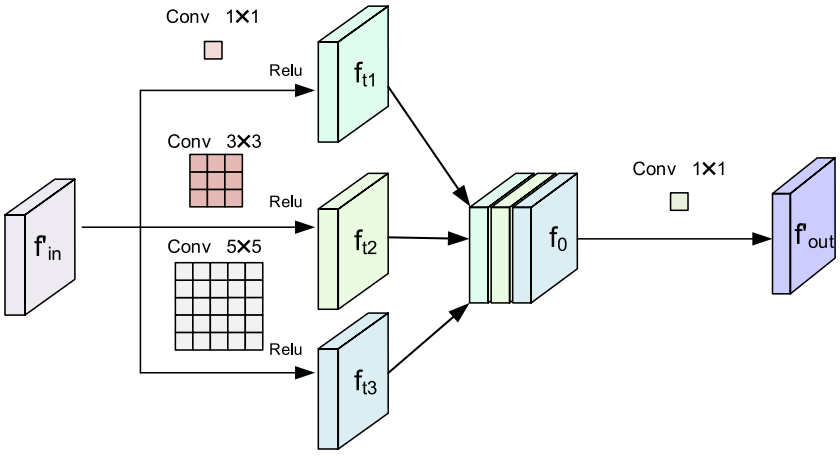


Fig. 2. Multi-scale feature fusion structure.

3.2. ECCSA

In M. tuberculosis detection, it is essential to accurately understand and localize the varied morphology of M. tuberculosis. To achieve this goal, it is critical to concentrate on channel and position information. CA captures inter-channel dependencies and also considers positional information related to orientation. However, when dealing with small-scale datasets, the model's ability to learn location features may be limited. To address this problem, a convolutional layer is added to the CA to allow more detailed image analysis. In addition, by introducing a self-attention mechanism after CA, the expressive power of the model is significantly enhanced. This enables the model to locate the position of M. tuberculosis more accurately, thus significantly improving the accuracy of detection.

Fig. 3 illustrates the structure of the improved attention mechanism module. The module contains two main parts: the CA and the self-attention mechanism module. Suppose the input feature map is F, whose shape is [C,H,W], where C denotes the number of channels of the feature map, H denotes the height of the feature map, and W denotes the width of the feature map. Then, the formula for averaging along the height is expressed by Equation (6).

$$\mathrm{avg}_H = \frac{1}{H} \sum_{j=1}^{H} F_{i,j,k} \quad \text{for } i = 1, 2, \cdots, C \quad \text{and } k = 1, 2, \cdots, W \tag{6}$$

The formula for averaging along the width is expressed by Equation (7).

$$\mathrm{avg}_W = \frac{1}{W} \sum_{k=1}^{W} F_{i,j,k} \quad \text{for } i = 1, 2, \cdots, C \quad \text{and } j = 1, 2, \cdots, H \tag{7}$$

After the above operations, the dimensions of the feature maps obtained are $[C, 1, W]$ and $[C, H, 1]$. The two parallel stages are then merged to obtain a feature layer with dimensions $[C, 1, H + W]$. Next, the features are obtained after convolution, normalization, and activation functions. Afterward, they are separated into two feature dimensions, resulting in two feature layers, F_H and F_W. Sigmoid is used to obtain the attention in wide and high dimensions, respectively. The original features and the features F' after the convolution of the other branch are multiplied, resulting in the final output feature F_S. F_S is calculated as in Equation (8).

$$F_S = F \times \mathrm{Sigmoid}(F') \times \mathrm{Sigmoid}(F_H) \times \mathrm{Sigmoid}(F_W) \tag{8}$$

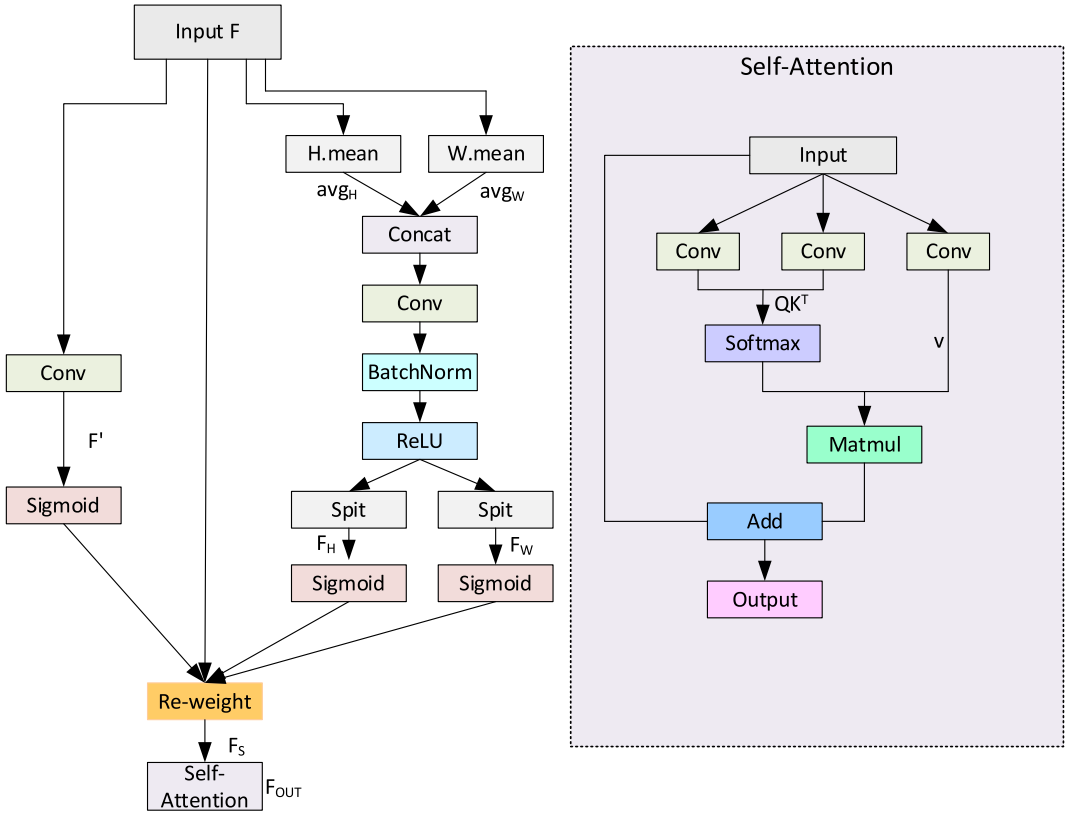


Fig. 3. Eccsa structure.

The expressive power of the model can be improved by introducing the self-attention mechanism after CA. The self-attention mechanism module generates attention weights by calculating the similarity between the query and the key. It then performs a weighted summation based on these weights. This allows the model to better understand the local features of the input data and improves the CA representation capability. Firstly, the input data is processed through three convolutional layers to obtain the query matrix Q, key matrix K, and value matrix V. Then, calculate the dot product between the query and the key, and divide it by the Euclidean norm to obtain the attention weight. Next, apply the Softmax function to the attention weights to ensure that the sum of weights is 1. The self-attention mechanism [30] is calculated as shown in Equation (9).

$$\mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \tag{9}$$

where d_k represents the dimension of K. The final output is calculated by Equation (10).

$$\mathrm{F_{out}} = \mathrm{F_S} + \mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) \tag{10}$$

4. Experiment and discussion

4.1. Dataset

The dataset [32] consists of sputum samples, and all images are related to tuberculosis. It contains 1265 sputum images and 3734 bounding boxes of M. tuberculosis. The training set used in this paper covers 1024 samples, the validation set contains 140 samples, and the test set contains 101 samples. Fig. 4 shows examples of the sputum sample image.

4.2. Experimental environment and parameter settings

The experimental environment is configured as follows: the programming language is python, the operating system is Ubuntu 18.04, the GPU is NVIDIA RTX3050TI, the memory is 8G, and the CUDA version is 11.7.

Table 2 shows the main parameters of model training. These parameters significantly influence the performance of the model. By adjusting the parameters, the detection accuracy of the model can be improved.

4.3. Evaluation metrics

To comprehensively assess the performance of the M. tuberculosis detection model, Average Precision (AP) [17] and frames per second (FPS) are used as evaluation metrics. FPS is used to evaluate the detection speed of the model. The higher the FPS, the faster the model processes. AP is calculated as shown in Equation (11):

$$\mathrm{AP} = \sum_{\mathrm{n}} (\mathrm{R_n} - \mathrm{R_{n-1}})\mathrm{P_n} \tag{11}$$

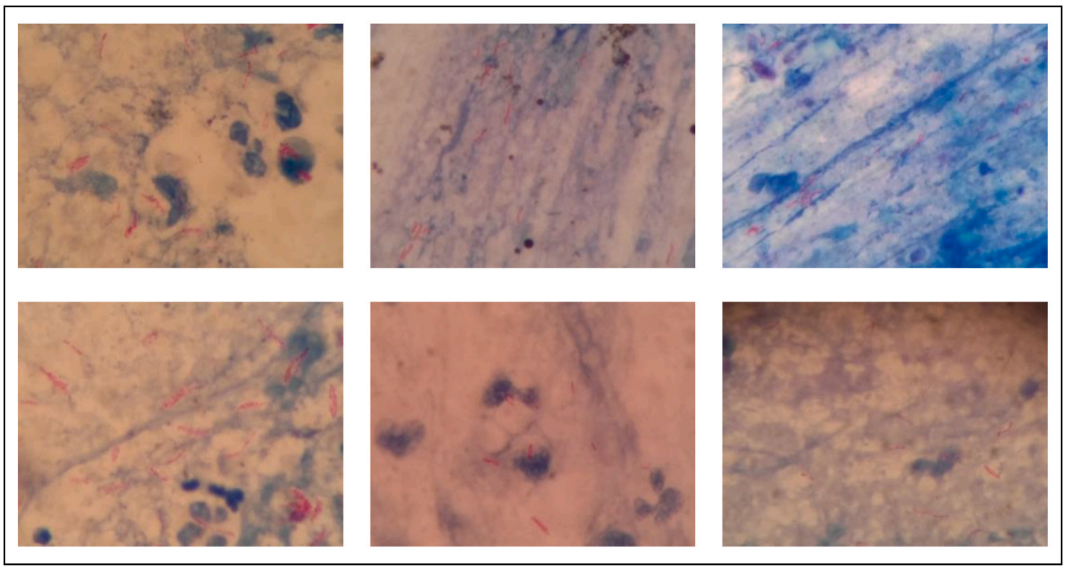


Fig. 4. Examples of sputum sample image.

Table 2
Parameter settings.

Parameter	Value
Input size	640*640
Unfreeze epoch	200
Unfreeze batch size	8
Freeze epoch	50
Freeze batch size	16
Optimizer	SGD
Momentum	0.937

where R_n and P_n are recall and precision at threshold $n = 0.5$.

4.4. Experimental results

4.4.1. Comparative experiments

Fig. 5 shows the total loss curve for model training. The training and validation losses can assist us in understanding the training progress of the model. From Fig. 5, it is clear that as the number of epochs increases, the loss value gradually decreases. As the training progresses, the ability of the model to fit the training and validation sets increases, resulting in a gradual decrease in the total loss value. When Epoch is 150, the loss value gradually stability. This indicates that the model has achieved excellent performance at this point.

Fig. 6 shows the detection results of the improved YOLOv8s model on the dataset. The performance of the model can be visually evaluated, and it can be observed that it successfully detects almost all M. tuberculosis. However, the varied morphology and size of M. tuberculosis pose challenges for target localization. Nevertheless, the introduction of MSFF and ECCSA allows the model to effectively address these challenges. However, although all the M. tuberculosis in the images is detected, the confidence level for individual detections is only 60 %–70 %, which could be further improved. Overall, the improved YOLOv8s model demonstrated good performance for the detection results.

To further validate the performance of the improved Yolov8s model, experiments compare the model with SSD [12], YOLOv3 [15], YOLOv5s [17], Improved YOLOv4 [27] and Improved YOLOv5s [32]. The comparative experiments allow us to assess the performance of our model in the M. tuberculosis target detection task more thoroughly.

Fig. 7 shows the trend plot of AP for each model in the comparison experiments. The AP of our model achieves the highest score, indicating its excellent performance in the target detection task. Table 3 shows the results of the comparative experiments on the validation set. In terms of accuracy, our model achieved the highest detection accuracy of 85.7 %. Compared to SSD, AP improved by 25.5 %. Compared to YOLOv3, AP improved by 23.8 %. Compared to YOLOv5s, AP improved by 6.7 %. Compared to Improved YOLOv4, AP has increased by 7.1 %. Compared to Improved YOLOv5s, AP has improved by 4.3 %. In terms of FPS, the SSD performs best at 84.0, while our model's FPS is relatively low at 26.9. Nonetheless, given the practical demands of M. tuberculosis detection, the high accuracy of our model may make this compromise in processing speed acceptable.

The performance of the various comparison models for M. tuberculosis detection is illustrated in Fig. 8 (a)-8 (f). The SSD model resulted in missing more M. tuberculosis targets due to its relatively low detection accuracy. False detections of M. tuberculosis appear in Fig. 8 (a) and 8 (b), and 8 (c). Additionally, some omissions occur during the detection process in Fig. 8 (d) and 8 (e). In contrast, in

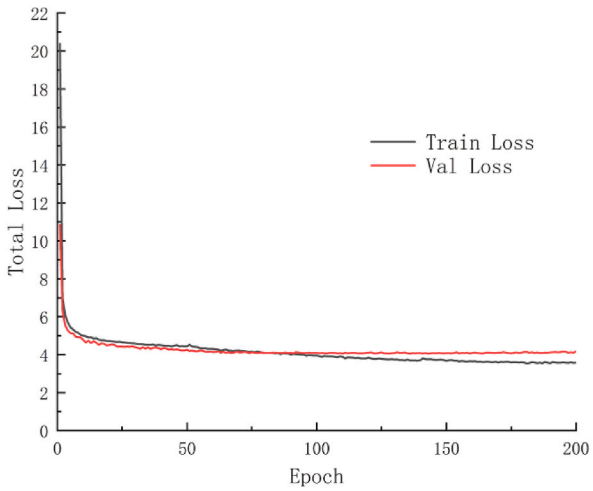


Fig. 5. Total training loss curve of the model.

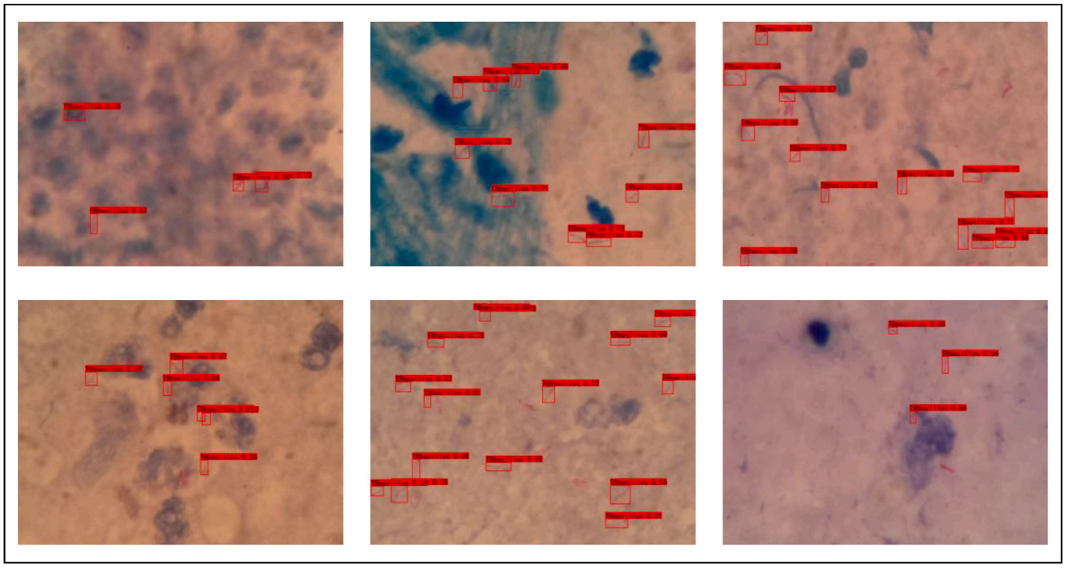


Fig. 6. Detection results of the proposed model.

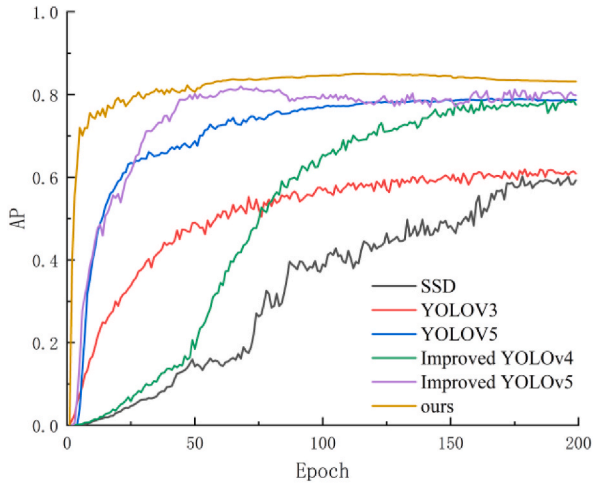


Fig. 7. Trend plot of AP for each model in the comparison experiments.

Table 3
Contrast results on the dataset.

Model	AP (%)	FPS
SSD	60.2	84.0
YOLOv3	61.9	36.8
YOLOv5s	79	49.2
Improved YOLOv4	78.6	29.6
Improved YOLOv5s	81.4	45.7
Ours	85.7	26.9

most cases, the YOLOv8s-based model can detect M. tuberculosis in samples. This is mainly attributed to the introduction of MSFF and ECCSA, which help the model better identify and localize M. tuberculosis bacilli of varied morphology. Overall, our model exhibits better target detection performance in M. tuberculosis detection.

4.4.2. Ablation experiments

Ablation experiments are used to assess the impact of individual components in the model on overall performance. By gradually

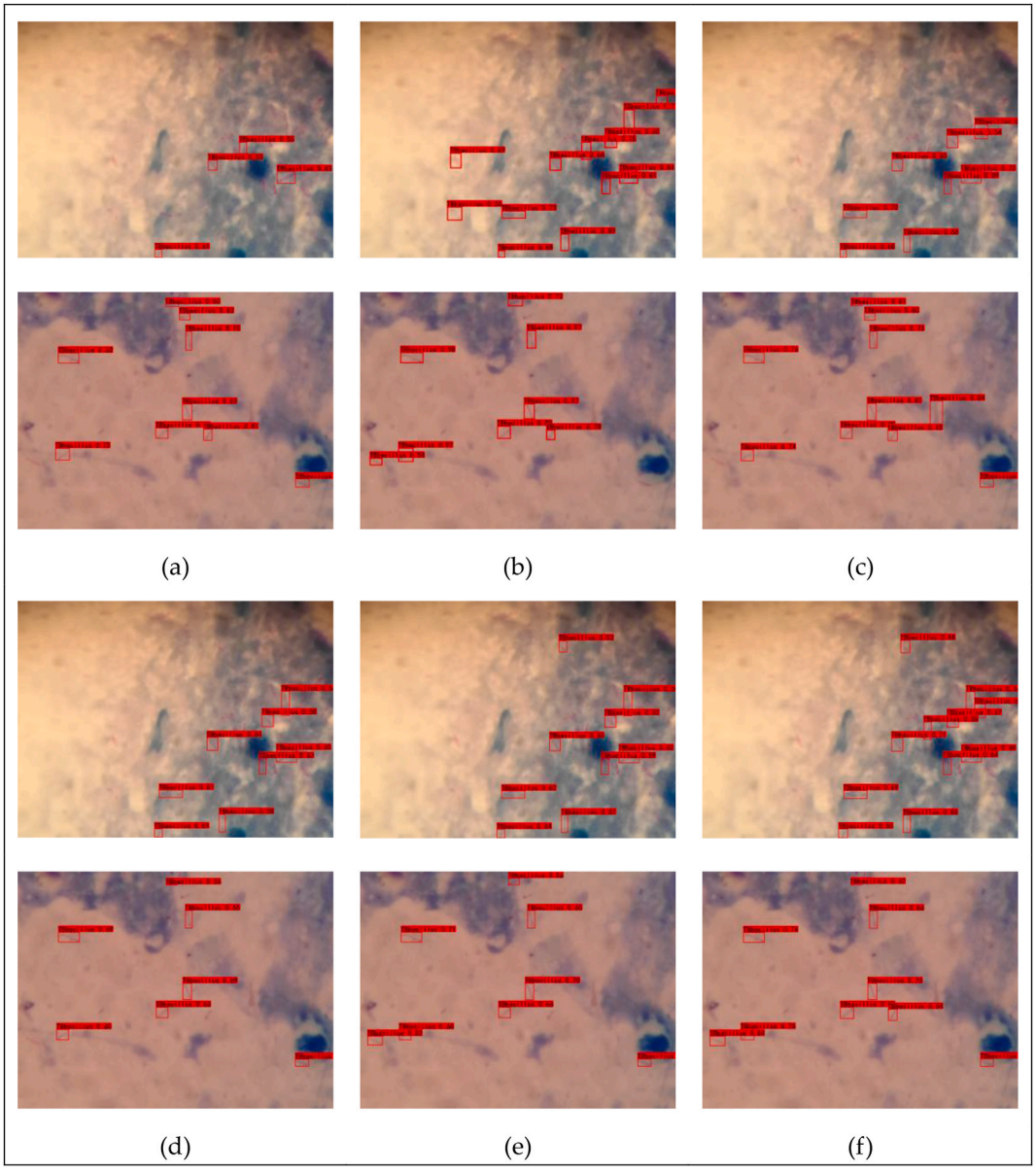


Fig. 8. Detection results in the comparison experiments. (a) SSD. (b) YOLOv3. (c) YOLOv5s. (d) Improved YOLOv4. (e) Improved YOLOv5s. (f) Ours.

removing MSFF and ECCSA, the contribution of each module to the model performance improvement can be quantified.

Fig. 9 presents the AP of each model in the ablation experiments. The impact of various modules on the model's detection performance can be clearly seen. It is noteworthy that the highest AP is achieved when the model introduces both MSFF and ECCSA. This indicates that our model performs well in the target detection task.

Table 4 demonstrates the results of the ablation experiments on the validation set. AP is 82.8 % when using YOLOv8s for detection. Introducing MSFF increases the AP to 84.4 %, an improvement of 1.6 % compared to YOLOv8s. Introducing ECCSA results in an AP of 83.6 %, an improvement of 0.8 % compared to YOLOv8s. Introducing both MSFF and ECCSA achieves an AP of 85.7 %. On the other hand, the introduction of MSFF or ECCSA in a decrease in FPS. Compared to YOLOv8s, the FPS of our model decreased by 9.8, but the AP increased by 2.9 %.

Three different attention mechanisms are chosen for comparison: coordinate attention, spatial attention, and ECCSA. These attention mechanisms are added after up sampling in the feature enhancement extraction network. Table 5 demonstrates comparative results. The results show that the introduction of the attention mechanisms improves the AP, verifying their effectiveness. Among them, ECCSA performs the best, with the highest AP. This indicates that it is more effective in guiding the model to focus on key regions. Consequently, it improves detection performance compared to other attention mechanisms.

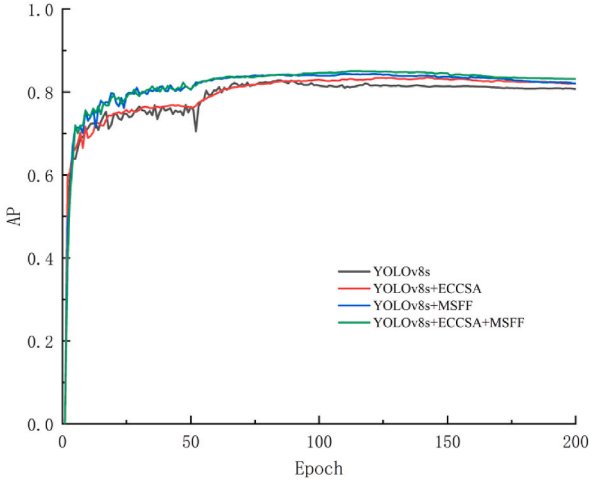


Fig. 9. Trend plot of AP in the ablation experiments.

Table 4
Ablation results on the dataset.

Model	MSFF	ECCSA	AP (%)	FPS
YOLOv8s			82.8	36.7
	✓		84.4	29.4
		✓	83.6	33.1
	✓	✓	85.7	26.9

The detection results in the ablation experiments are shown in Fig. 10. In Fig. 10 (a) and 10 (b), it is found that using the YOLOv8s model, target loss occurs in the detection of morphologically diverse M. tuberculosis. Fig. 10 (c) shows that our model can detect the target better. This indicates that our model has a significant advantage over the YOLOv8s model for the M. tuberculosis detection.

4.5. Discussion

The SSD model is mainly designed for the detection of fixed-size targets. Therefore, its effectiveness may be limited when detecting M. tuberculosis with large scale variations, resulting in missed detections. Compared to YOLOv8s, the YOLOv3 and YOLOv5s models are slightly less capable of feature extraction. Improved YOLOv4 enhances the feature extraction capability with the multi-sensory field module, which enables it to deal with M. tuberculosis with different morphologies. Improved YOLOv5s uses the bridge attention mechanism with a bidirectional feature pyramid network. This combination captures both shallow features of the target and incorporates deeper semantic information, thus improving the accuracy of target recognition. However, M. tuberculosis targets tend to have varied morphology and size, which makes the recognition ability of the above models still challenged. To address these issues, we propose an improvement scheme that includes MSFF and ECCSA. The addition of these modules significantly improves the model's ability to detect M. tuberculosis with varied morphology and size.

In M. tuberculosis detection, the YOLOv8s model has been used with some accuracy. However, due to the varied morphology and size of M. tuberculosis, the model has shown limitations in handling these challenges. By incorporating MSFF into the YOLOv8s model, the model's detection capability is significantly improved. The module allowed the model to better capture M. tuberculosis features at various scales, improving the accuracy of detection of scale-varying targets. Incorporating a small target detection head allows the model to detect smaller targets. Additionally, with the introduction of ECCSA, the model can better understand and localize the varied morphology of M. tuberculosis. The attention mechanism helps the model to focus on more critical regions and reduce the interference of background noise, thus improving the accuracy of target localization. Finally, when combined with MSFF and ECCSA, the model's performance is optimal.

The FPS of our model is reduced in the experiments when compared to other methods, which is primarily due to the model's more complex network structure. Nonetheless, this strategy resulted in a significant improvement in AP, fully demonstrating the excellent performance of our model in detecting M. tuberculosis. It is worth noting that in the practical application of M. tuberculosis detection, the accuracy of detection is often more important than the speed. This is because accurate detection results are directly related to the timely diagnosis of the disease. Therefore, based on practical application requirements, the trade-off between a decrease in FPS and a significant increase in AP in our method is a reasonable consideration.

In summary, this paper proposes a new target detection model that is effective in the detection of M. tuberculosis. However, our studies still have some limitations. First, we only used a single dataset for evaluation in our experiments, which may affect the

Table 5
Comparative results of different attention mechanisms.

Model	Coordinate attention	Spatial attention	ECCSA	AP (%)
YOLOv8s	✓			83.0
		✓		82.9
			✓	83.6

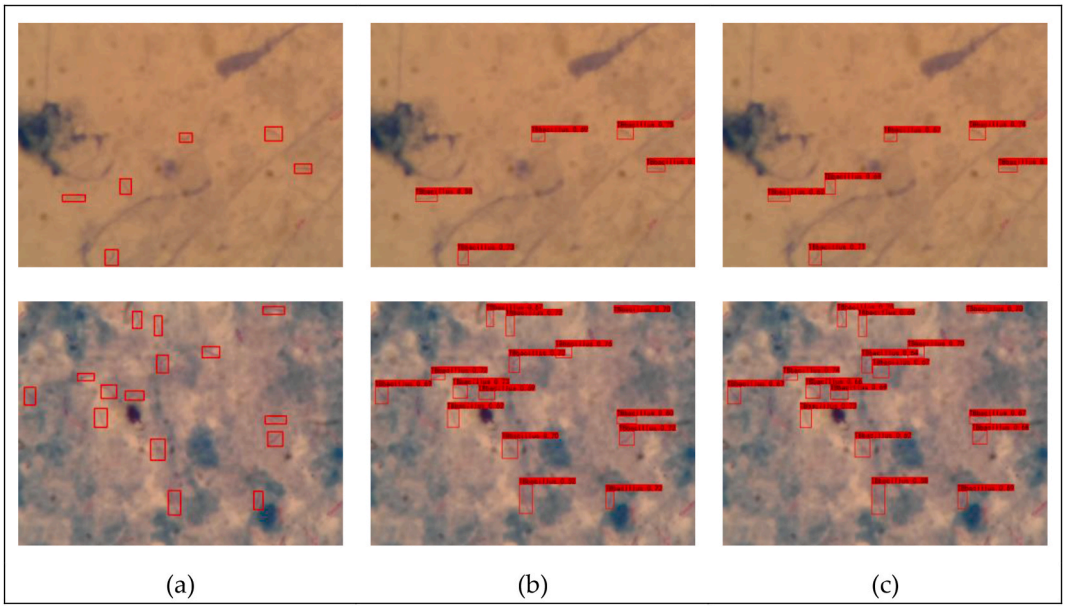


Fig. 10. Detection results in the ablation experiments. (a) Test Sample. (b) YOLOv8s. (c) Ours.

generalizability of the results. Second, the model did not have a high level of confidence in detecting M. tuberculosis, which may have impacted the accuracy of the results.

5. Conclusion

To address the impact of varied morphology and size on target detection accuracy in M. tuberculosis detection, an improved YOLOv8s model is proposed. The original large target detection head is removed, and a new head that focuses on small target detection is added. Additionally, a multi-scale feature detection module is proposed to ensure accurate detection of multi-scale M. tuberculosis in the scene. Furthermore, the attention mechanism is improved to allow the model to better focus on the spatial and positional information of M. tuberculosis, enhancing the localization accuracy for its varied morphology. The experimental results are satisfactory, and our model achieves an AP of 85.7 %.

In the future, we will conduct tests on multiple datasets to verify the generalization ability of the model. Additionally, we will introduce more contextual information into the model design to improve the accuracy of target detection.

Data availability statement

The dataset used in this study is publicly available on [AI Studio] and can be accessed at https://aistudio.baidu.com/datasetdetail/83968/0.

Ethical approval

This study does not require ethical approval as it uses a publicly available dataset that does not involve human participants or animals.

CRediT authorship contribution statement

Hao Chen: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Wenye Gu:** Resources, Investigation, Formal analysis, Data curation. **Haifei Zhang:** Resources, Methodology,

Investigation. **Yuwei Yang:** Validation, Formal analysis. **Lanmei Qian:** Validation, Supervision.

Declaration of competing interest

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

Acknowledgments

This research was funded by Natural Science Foundation of University in Jiangsu Province (Grant 23KJD520011), Research topic on educational informatization in Jiangsu Higher Education Institutions (Grant 2023JSETKT126), and Directive projects of Nantong municipal science and technology plan (Grant MS2023061). The authors are thankful to all the personnel who either provided technical support or helped with data collection. We also acknowledge all the reviewers for their useful comments and suggestions.

References

[1] A. Natarajan, P.M. Beena, A.V. Devnikar, S. Mali, A systemic review on tuberculosis, Indian J. Tubercul. 67 (3) (2020) 295–311.
[2] L. An, K. Peng, X. Yang, P. Huang, Y. Luo, P. Feng, B. Wei, E-TBNet: light deep neural network for automatic detection of tuberculosis with X-ray DR imaging, Sensors 22 (3) (2022) 821.
[3] E. Kotei, R. Thirunavukarasu, A comprehensive review on advancement in deep learning techniques for automatic detection of tuberculosis from chest X-ray images, Arch. Comput. Methods Eng. 31 (1) (2024) 455–474.
[4] R. Dinkele, S. Gessner, A. McKerry, B. Leonard, R. Seldon, A.S. Koch, D.F. Warner, Capture and visualization of live Mycobacterium tuberculosis bacilli from tuberculosis patient bioaerosols, PLoS Pathog. 17 (2) (2021) e1009262.
[5] M. Zachariou, O. Arandjelović, D.J. Sloan, Automated methods for tuberculosis detection/diagnosis: a literature review, BioMedInformatics 3 (3) (2023) 724–751.
[6] Y. Bai, K. Liu, Y. Chen, H. Zhao, Y. Wang, X. Liu, L. Zheng, Disseminated infection of Nocardia farcinica in an immunocompetent adult: mistaken for tuberculosis bacilli in acid-fast staining of bronchoalveolar lavage fluid, J. Cytol. 38 (2) (2021) 106–108.
[7] H. Yousefi, F. Mohammadi, N. Mirian, N. Amini, Tuberculosis bacilli identification: a novel feature extraction approach via statistical shape and color models, in: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2020, December, pp. 366–371.
[8] M. El-Melegy, D. Mohamed, T. ElMelegy, M. Abdelrahman, Identification of tuberculosis bacilli in ZN-stained sputum smear images: a deep learning approach, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 1131–1137.
[9] R.S. Chithra, P. Jagatheeswari, Severity detection and infection level identification of tuberculosis using deep learning, Int. J. Imag. Syst. Technol. 30 (4) (2020) 994–1011.
[10] P. Huang, H. Xiao, P. He, C. Li, X. Guo, S. Tian, J. Qin, LA-ViT: a network with transformers constrained by learned-parameter-free attention for interpretable grading in a new laryngeal histopathology image dataset, IEEE Journal of Biomedical and Health Informatics (2024).
[11] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015).
[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, Ssd: single shot multibox detector, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer International Publishing, 2016, pp. 21–37.
[13] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[14] S. Aulia, A.B. Suksmono, T.R. Mengko, B. Alisjahbana, A novel digitized microscopic images of ZN-stained sputum smear and its classification based on IUATLD grades, IEEE Access (2024).
[15] J. Redmon, A. Farhadi, Yolov3: An incremental improvement (2018) *arXiv preprint arXiv:1804.02767*.
[16] A. Bochkovskiy, C.Y. Wang, H.Y.M. Liao, Yolov4: Optimal speed and accuracy of object detection (2020) *arXiv preprint arXiv:2004.10934*.
[17] X. Zhu, S. Lyu, X. Wang, Q. Zhao, TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2778–2788.
[18] T.F.M. Carvalho, V.L.A. Santos, J.C.F. Silva, L.J. de Assis Figueredo, S.S. de Miranda, R. de Oliveira Duarte, F.G. Guimarães, A systematic review and repeatability study on the use of deep learning for classifying and detecting tuberculosis bacilli in microscopic images, Prog. Biophys. Mol. Biol. 180 (2023) 1–18.
[19] M. Parveen Rahamathulla, W.R. Sam Emmanuel, A. Bindhu, M. Mustaq Ahmed, YOLOv8's advancements in tuberculosis identification from chest images, Frontiers in Big Data 7 (2024) 1401981.
[20] S. Ayas, M. Ekinci, Random forest-based tuberculosis bacteria classification in images of ZN-stained sputum smear samples, Signal, Image and Video Processing 8 (2014) 49–61.
[21] F. Costa, C F F, P.C. Levy, C.D.M. Xavier, L.B.M. Fujimoto, M.G.F. Costa, Automatic identification of tuberculosis mycobacterium, Research on biomedical engineering 31 (2015) 33–43.
[22] K.S. Mithra, W.S. Emmanuel, FHDT: fuzzy and Hyco-entropy-based decision tree classifier for tuberculosis diagnosis from sputum images, Sādhanā 43 (2018) 1–15.
[23] C. Xu, D. Zhou, Y. Zhai, Y. Liu, Automatic segmentation and classification of mycobacterium tuberculosis with conventional light microscopy, in: MIPPR 2015: Parallel Processing of Images and Optimization; and Medical Imaging Processing, vol. 9814, SPIE, 2015, December, pp. 42–47.
[24] Y. Wang, F. Luo, X. Yang, Q. Wang, Y. Sun, S. Tian, H. Xiao, The Swin-Transformer network based on focal loss is used to identify images of pathological subtypes of lung adenocarcinoma with high similarity and class imbalance, J. Cancer Res. Clin. Oncol. 149 (11) (2023) 8581–8592.
[25] P. Huang, C. Li, P. He, H. Xiao, Y. Ping, P. Feng, J. Qin, MamlFormer: priori-experience guiding transformer network via manifold adversarial multi-modal learning for laryngeal histopathological grading, Inf. Fusion 108 (2024) 102333.
[26] L. An, K. Peng, X. Yang, P. Feng, P. Huang, Automated detection of tuberculosis bacilli using deep neural networks with sputum smear images, in: 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), IEEE, 2022, August, pp. 1040–1045.
[27] Z. Guo, J. Wang, J. Wang, J. Yuan, Lightweight YOLOv4 with multiple receptive fields for detection of pulmonary tuberculosis, Comput. Intell. Neurosci. (2022), 2022.
[28] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, Cbam: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[29] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713–13722.
[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[31] H. Xie, M. Lu, J. Liu, B. Xu, X. Shi, C. Zhang, J. Cong, Secondary pulmonary tuberculosis lesions detection based on improved YOLOv5 networks, in: International Conference on Swarm Intelligence, Springer Nature Switzerland, Cham, 2023, July, pp. 220–231.
[32] B. Lv, H. Lan, Improved YOLOv5-based detection model for Mycobacterium, in: 2023 IEEE 7th Information Technology and Mechatronics Engineering Conference (ITOEC), vol. 7, IEEE, 2023, September, pp. 1360–1364.
[33] Y. Li, C. Zhou, Z. Zhao, L. Li, Research on detection method of Tubercle Bacilli based on the improved YOLOv5, Phys. Med. Biol. 68 (10) (2023) 105008.