Multitask Learning Based on YOLOv11 for Automatic Detection and Segmentation of Tuberculosis Bacilli in Sputum Smear Microscopy

No Author Given

No Institute Given

Abstract. Tuberculosis, caused by the Mycobacterium tuberculosis bacterium, remains one of the leading causes of death from infectious diseases worldwide, particularly in developing countries such as Brazil. In this study, we evaluated the performance of the YOLOv11 model for identifying Mycobacterium tuberculosis in sputum bacilloscopy images, employing different pre-processing approaches aimed at accelerating the diagnostic process. Four datasets were assessed: (a) RGB, (b) LUV, (c) CLAHE, and (d) LUV + CLAHE. The results indicated that the model trained with the LUV color space achieved the best performance, reaching 97.60% accuracy, 95.60% recall, and 98.30% mAP in detection, as well as 84.60% accuracy, 82.80% recall, and 84.80% mAP in segmentation. These findings demonstrate the potential of the model to automate the detection and segmentation of disease-causing bacteria, thereby streamlining the diagnostic workflow.

Keywords: Tuberculosis · YOLOv11 · Sputum Smear Microscopy.

1 Introduction

Tuberculosis (TB) is an infectious disease caused by Mycobacterium tuberculosis (MT). This condition primarily affects the lungs but can also compromise other parts of the body, such as the kidneys, bones, nervous system, and lymph nodes. Transmission occurs mainly through the air when infected individuals cough or sneeze, releasing small droplets that contain the infectious agent. Historical records indicate the presence of tuberculosis for centuries, with evidence of infection found in Egyptian mummies, and its study already documented in Ancient Greece [4]. In recent years, TB has once again become the leading cause of death from infectious diseases worldwide, surpassing COVID-19, which held this position for three years. Therefore, tuberculosis remains a serious global public health issue [12].

According to the 2023 Global Tuberculosis Report by the World Health Organization (WHO), 1.25 million deaths were caused by tuberculosis, of which 161,000 occurred in people living with HIV [12]. In Brazil, the number of TB cases remains alarming. According to data from the Ministério da Saúde (MS), approximately 80,000 new cases of infection and around 5,500 TB-related deaths are reported annually [10]. Each year, the disease becomes an increasingly serious concern due to the rising number of deaths. Most cases occur in countries with low and middle Gross National Income (GNI) [12].

Despite technological advances that have enabled the development of new approaches for the diagnosis and treatment of tuberculosis, many of these methods are still not widely adopted due to their high production and implementation costs [6]. Available diagnostic methods include: microscopic examination, chest radiography, skin testing, MT culture, and interferon-gamma release assays. However, the most widely used technique remains sputum bacilloscopy, which involves the microscopic analysis of patients' sputum samples. The identification and counting of Mycobacterium tuberculosis are performed by trained technicians who examine the slides under a microscope [1]. Although this is the most commonly used procedure, its efficiency can still be improved, as it relies on manual analysis. Moreover, it is a time-consuming test that demands patience and an experienced professional, particularly in the early stages of the disease when detection becomes more challenging [8].

Given the aforementioned context and recent technological advances, several innovative methods have been developed to support the fight against tuberculosis and enhance its treatment, such as the application of Computer Vision based on Deep Learning. The use of this technology in the medical field has expanded over the years. In the context of TB diagnosis, these methods can be employed to automate and improve the detection of MT in sputum samples, making the process more accurate and efficient. Furthermore, techniques involving color space transformations of clinical images have shown positive impacts on the performance of neural networks. These approaches contribute significantly to the early identification of the disease, enabling more effective treatment.

In this context, the main distinguishing feature of our method lies in the use of the latest YOLOv11 architecture for the detection and segmentation of MT, combined with a detailed study of the impact of different image pre-processing techniques. This approach enables a more accurate assessment of the effectiveness of these techniques in automating tuberculosis diagnosis.

2 Related Work

In recent years, several approaches have been proposed for the automatic classification and segmentation of bacilli using machine learning techniques and image pre-processing. In this section, we present relevant studies that address methods similar to ours, discussing their advancements and limitations.

Gomide et al. [1] proposed the development of an automated system for the detection and counting of acid-fast bacilli (AFB) in TB diagnosis, using the YOLOv3 architecture. The study includes the creation of an image dataset and experiments with various configurations. It also involves a comparison between manual slide readings and a Deep Learning algorithm, using 340 prepared slides. Preliminary findings indicate that the artificial intelligence-based approach is more accurate than previous methods.

Reis [7] presented a Deep Learning-based approach for tuberculosis detection through the analysis of sputum smear microscopy images. The study employs the MLP-Mixer to classify image patches as containing or not containing MT, and YOLOv7 to localize and delimit its regions. The results were promising, achieving a mAP of 74.08%, precision of 81.36%, recall of 71.36%, and F1-score of 76.46%.

The study by S. dos Santos et al. [9] addresses the detection of MT using the EfficientDet architecture with various backbones and four different color representations. Through a 5-fold cross-validation, the results showed an IoU of 0.523, recall of 0.925, precision of 0.694, and an F1-score of 0.774.

In their study, Lopez [3] proposed an automatic detection method for MT using Convolutional Neural Networks (CNNs) and bacilloscopic images in RGB, R-G, and Grayscale formats. The best patch classification results were obtained using R-G and RGB images, achieving 99% accuracy with models composed of two and three convolutional layers. For full-image detection, the best performance was achieved with RGB images, reaching a precision of 56.82%, recall of 86.15%, and an F1-score of 68.47%.

Xiong et al. [13] investigated the clinical effectiveness of an artificial intelligence (AI)-assisted method for detecting tuberculosis MT in Ziehl-Neelsen-stained samples. The proposed model, named TB-AI, was developed using Convolutional Neural Networks (CNNs) to recognize the bacilli. Compared to pathologist-based diagnoses, TB-AI achieved a sensitivity of 97.94% and a specificity of 83.65%.

Table 1 presents a summary of the metrics from the previously discussed works.

Table 1. Summary of related works.

Reference	Method	Dataset	Key Metrics
Gomide et al. [1]	YOLOv3 for detection and counting of AFB	340 prepared slides	Higher accuracy compared to manual methods
Reis [7]	MLP-Mixer for classification and YOLOv7 for detection	Sputum smear microscopy	mAP: 74.08%, Precision: 81.36%, Recall: 71.36%, F1-score: 76.46%
S dos Santos et al [9]	EfficientDet with multiple backbones	Sputum smear microscopy	IoU: 0.523, Recall: 0.925, Precision: 0.694, F1-score: 0.774
López [3]	CNNs for automatic detection	Sputum smear (RGB, R-G, Grayscale)	Precision: 56.82%, Recall: 86.15%, F1-score: 68.47% (full images)
Xiong et al. [13]	CNNs for bacilli detection (TB-AI)	Ziehl-Neelsen stained samples	Sensitivity: 97.94%, Specificity: 83.65%

The literature review shows that several studies address MT detection for TB diagnosis, with some employing the YOLO architecture. However, the versions explored so far, such as YOLOv3 and YOLOv7, present limitations when compared to more recent architectures, particularly in terms of segmentation capability and refined extraction of the bacilli's morphological features.

Unlike previous studies, this work employs YOLOv11, which offers advancements in feature extraction and object detection, enabling more accurate segmentation of MT in microscopic images. Moreover, while earlier approaches focus solely on detection, this study incorporates a segmentation pipeline, allowing for the precise localization of infected regions.

Another relevant aspect is the image pre-processing. While previous studies predominantly use RGB representations or grayscale, this work investigates the impact of the LUV color space and the CLAHE histogram equalization technique. The combination of these strategies enhances the contrast of MT and reduces lighting interference, thereby improving the model's performance.

3 Materials and Methods

This section presents all the steps involved in the detection and segmentation of bacilli, as well as the selection of the best color channel for training. The methods include the acquisition and organization of the dataset, image pre-processing for different color channels to generate new datasets, followed by model training with the datasets, and finally, evaluation of the results. Figure 1 provides a summary of these steps.

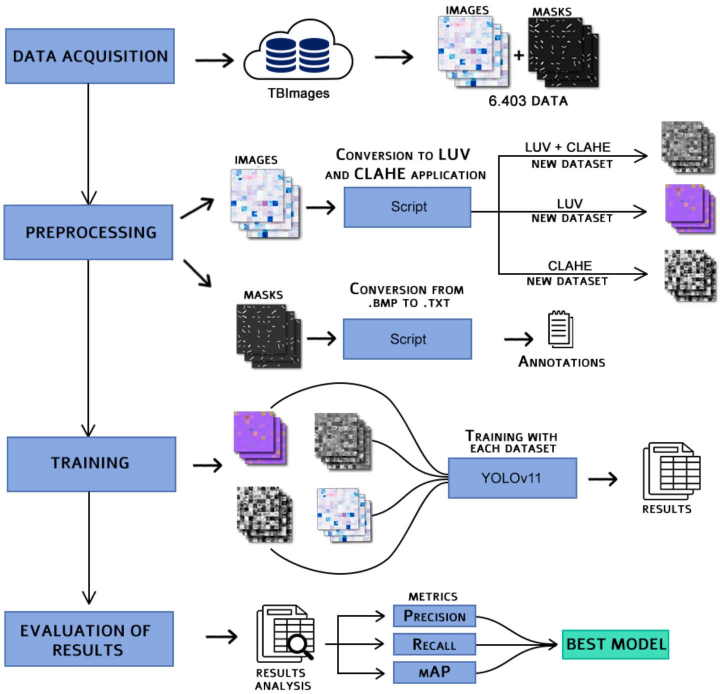


Fig. 1. Proposed methodological flowchart

3.1 Data Acquisition

This work uses MT images provided by the Pattern Recognition and Optimization Research Group at the Federal University of Amazonas (UFAM) [11]. The dataset employed in the research consists of a total of 6,403 images, of which 2,852 were extracted from the digital fields of slides S01 to S15, while the remaining 3,551 correspond to binary mask images. The masks are black-and-white representations, where white pixels indicate the presence of MT and black pixels correspond to the background of the image.

All images have a resolution of 400 × 400 pixels, including both the original slides and their corresponding masks. The files are stored in the BMP (Bitmap) format, a widely used raster image format for uncompressed digital visual data. The images are arranged in mosaics and contain regions with both positive and negative staining. Each sample in the dataset consists of a set of 100 sub-images. Figure 2 shows a representative example of an image from the dataset along with its corresponding mask.

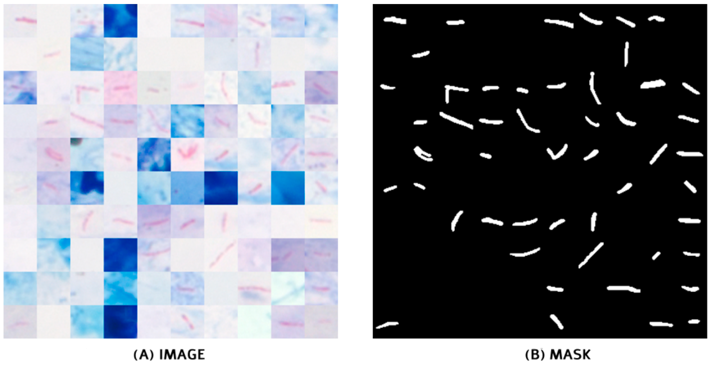


Fig. 2. Example of an image and its corresponding mask (Source: UFAM dataset [11])

3.2 Preprocessing

Before processing, we identified a discrepancy in the number of files in the training and validation sets. In the training set, there were 3,051 masks but only 2,540 images, resulting in an excess of 511 unmatched masks. The validation set exhibited an even more pronounced imbalance, with 250 masks for only 62 images. In contrast, the test set was balanced, containing 250 files of each type. To address these inconsistencies, we applied data filtering to ensure each image had a corresponding mask. After adjustment, the training set contained 2,032 images and masks, while the validation set was reorganized to include 62 image-mask pairs. Thus, the final dataset was structured as follows: approximately 87% of the images were allocated to the training set, 10% to the test set, and 3% to the validation set.

To enhance images quality and improve model performance, we applied various preprocessing techniques to the original dataset. First, the images were converted to the LUV color space, which is more suitable for contrast enhancement and texture analysis. This color space was selected due to its ability to separate chrominance from luminance, potentially aiding the segmentation of microscopic structures. We also employed Contrast Limited Adaptive Histogram Equalization (CLAHE), a technique widely used in medical imaging to improve contrast while reducing the effects of non-uniform illumination. Finally, we combined LUV conversion with CLAHE to leverage the benefits of both methods. At the end of this process, we obtained four distinct datasets: (a) images in the RGB color space, (b) images in the LUV color space, (c) RGB images processed CLAHE, and (d) images with LUV + CLAHE. Figure 3 shows an example of the same image across these different datasets.

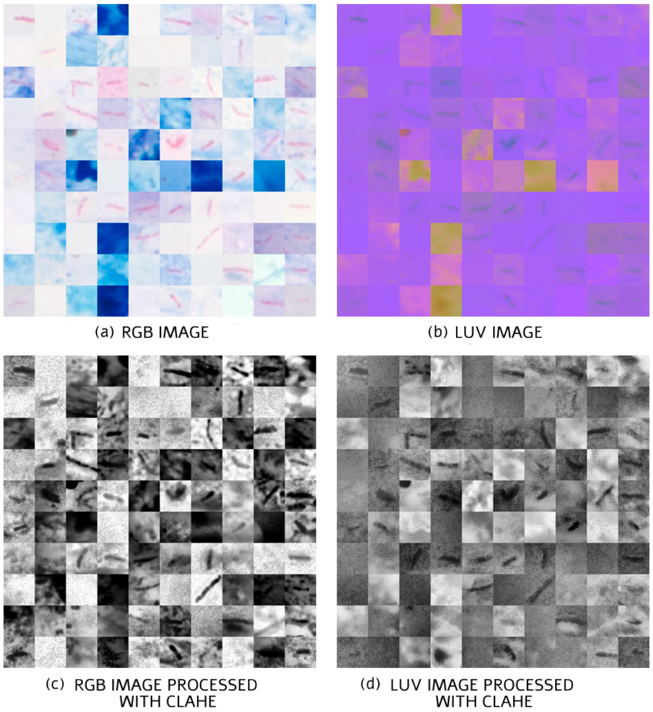

(a) RGB IMAGE
(b) LUV IMAGE
(c) RGB IMAGE PROCESSED WITH CLAHE
(d) LUV IMAGE PROCESSED WITH CLAHE

Fig. 3. Example of image in different datasets

To ensure compatibility with YOLOv11, which uses TXT-format annotations, we converted the originally BMP-stored masks. For this purpose, we developed a Python script to automate the conversion process, extracting mask information and generating the corresponding annotation files. This procedure ensured each mask was accurately represented in coordinates suitable for MT detection, streamlining the neural network training process.

3.3 Training with YOLOv11 Model

In this work, we employed a Convolutional Neural Network (CNN)-based architecture for MT detection and segmentation. Among the various available approaches, we selected YOLO (You Only Look Once), a widely recognized architecture known for its balance between speed and accuracy, making it an efficient choice for object detection and segmentation applications [5]. Furthermore, its latest version, YOLOv11, demonstrates significant improvements over previous versions, particularly in terms of enhanced accuracy, speed, and computational efficiency.

YOLOv11 improves its architecture through three key components: the *backbone*, *neck*, and *head*. In the backbone, the C2F module has been replaced with C3K2, which enhances feature extraction efficiency without increasing computational complexity. The neck incorporates the novel C2PSA module, which serves as an attention mechanism to emphasize the most relevant image information. For the head component, YOLOv11 adopts an anchor-decoupled design where regression uses standard convolutions while classification employs more efficient convolutions, simultaneously reducing computational costs and improving accuracy [2].

The training hyperparameters were empirically adjusted to optimize the model's performance. The initial learning rate was set to 3×10^{-4}, a value suitable for ensuring stable convergence without abrupt fluctuations, and was combined with a warm-up strategy during the first epoch to prevent unstable updates at the beginning of training. The batch size was fixed at 16, balancing computational efficiency and gradient propagation stability. The number of epochs was set to 200, providing sufficient time for model convergence without overfitting the data. To monitor training behavior, a loss evolution chart was generated over the epochs, allowing verification of optimization stability and the identification of any signs of overfitting or underfitting.

For images preprocessing, we utilized an 11th-generation Intel Core i5 processor with 16GB RAM and a SATA SSD for data storage. For model training, we employed an NVIDIA GeForce RTX 2080 Ti Graphics Processing Unit (GPU) to accelerate and optimize the training process.

3.4 Evaluation

Model performance was evaluated using standard statistical metrics, including Precision (Pre), Recall (Rec), and Mean Average Precision (mAP). These metrics were employed to assess both bounding box generation and segmentation mask quality, enabling identification of the best-performing model. The YOLO framework automatically computes these metrics at the end of each training cycle, providing quantitative and objective measurements of model effectiveness across different detection and segmentation aspects.

4 Results and Discussion

This section presents the results obtained after training each dataset for MT detection and segmentation. The primary objective is to analyze the evaluation metrics discussed in Section 3.4, comparing the performance across different datasets. The results were analyzed considering four dataset variations: (a) RGB, (b) LUV, (c) RGB + CLAHE, and (d) LUV + CLAHE. This approach enables an assessment of the impact of each preprocessing technique on the model's performance.

4.1 Performance Evaluation of Detection and Segmentation

Table 2 presents the results of MT detection on sputum bacilloscopy images. During training, the techniques described in Section 3.2 were applied to determine which color representation best fits the study's methodology. The results indicate that the LUV color space stood out by achieving the highest metrics, with a Precision of 97.60%, Recall of 95.60%, and a mAP of 98.30%.

Table 2. Bounding box metrics

Dataset	Precision	Recall	mAP
RGB	97.30%	95.30%	97.90%
LUV	97.60%	95.60%	98.30%
CLAHE	84.00%	85.40%	89.00%
LUV + CLAHE	89.30%	81.10%	88.20%

The results indicate that the model trained with the LUV color space achieved the highest precision, Recall and mAP, outperforming the other approaches. This can be attributed to the fact that the LUV color space separates luminance from chromaticity, thereby reducing illumination interference and enhancing the differentiation of the morphological structures of the bacilli. Although the dataset in the RGB color space yielded slightly lower metrics, the minimal difference suggests that conversion to LUV improves detection accuracy without compromising the model's sensitivity. On the other hand, the application of CLAHE, either in isolation or combined with LUV, resulted in inferior performance, possibly due to the increased contrast granularity, which may introduce noise that confuses the neural network during bacilli identification.

Table 3 presents the results of MT segmentation. It is observed that the dataset using the LUV color space achieved the best overall performance, reaching a Precision of 84.60%, a Recall of 82.80%, and a mAP of 84.80%, outperforming the values obtained with the RGB-based dataset. On the other hand, the application of the CLAHE technique, either in isolation or combined with LUV, resulted in inferior performance, with significant drops in segmentation metrics.

Table 3. Segmentation metrics

Dataset	Precision	Recall	mAP
RGB	83.50%	81.50%	83.70%
LUV	84.60%	82.80%	84.80%
CLAHE	71.50%	72.70%	74.50%
LUV + CLAHE	70.80%	64.30%	65.70%

In segmentation, the model trained with the LUV color space maintained the highest metrics, reinforcing the hypothesis that the separation of luminance and chromaticity enhances the definition of bacilli contours, enabling more accurate segmentation. However, the CLAHE technique had a negative impact, both in isolation and when combined with LUV, leading to a significant reduction in all metrics. This behavior can be explained by the fact that histogram equalization tends to amplify local contrast variations, which, in medical images, may accentuate artifacts or patterns that are irrelevant to the model, thereby hindering the correct segmentation of the bacilli.

4.2 Comparison with Other Results

The comparative analysis of the results obtained in this study with those from other works in the literature reveals both advancements and ongoing challenges in MT detection and segmentation. Table 4 highlights the main points of comparison between the results presented in this study and those of three relevant studies in the field of MT detection in bacilloscopy images.

Table 4. Comparative analysis of MT detection methods

Study	Precision	Recall	F1-Score	mAP
Reis [7]	81.36%	71.36%	76.46%	74.08%
S dos Santos et al. [9]	69.40%	92.50%	77.40%	–
López [3]	56.82%	86.15%	68.47%	–
Our Approach (LUV)	**97.60%**	**95.60%**	–	**98.30%**

The comparison with previous studies shows that the proposed approach significantly outperformed models based on earlier versions of YOLO, traditional CNNs, and architectures such as EfficientDet. The use of YOLOv11, with optimizations in the *backbone*, *neck* and *head*, enabled improvements in feature extraction and segmentation, resulting in superior metrics. Furthermore, the selection of the LUV color space and the exclusion of techniques such as CLAHE reinforce that specific preprocessing adjustments can significantly impact model performance. These results underscore the importance of evaluating not only the network architecture but also the image transformations applied, in order to maximize the robustness of detection and segmentation.

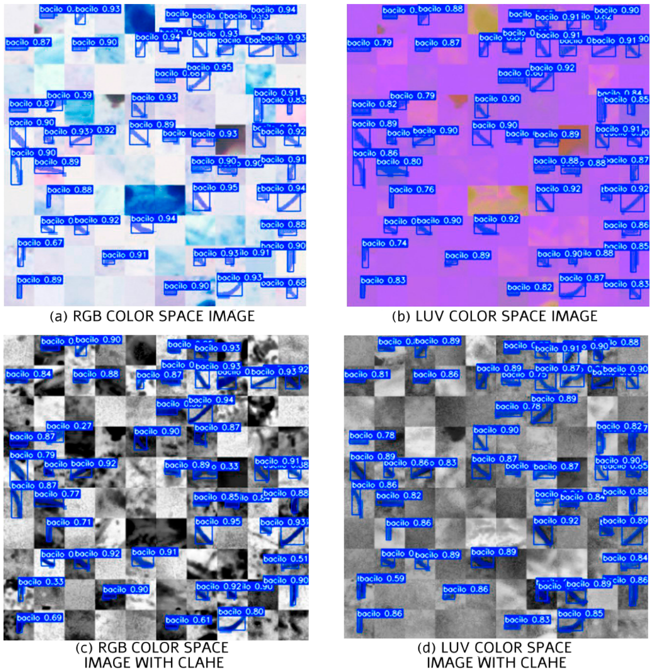
4.3 Case Study

In this case study, a visual assessment of MT predictions and segmentations was conducted to validate the models' accuracy in identifying small and low-opacity MT. The analysis was performed using an image from the dataset characterized by small-sized and low-visibility MT, which could lead to prediction inaccuracies. The results were visually compared across the four datasets used in training: (a) RGB, (b) LUV, (c) RGB + CLAHE, and (d) LUV + CLAHE.

Figure 4 presents the MT prediction and segmentation results obtained from different datasets. The model trained with the dataset in the LUV color space achieved the best performance, detecting 51 MT out of a total of 52. In comparison, the model trained with the RGB dataset also detected 51 bacilli, demonstrating similar performance. The model trained with the LUV + CLAHE dataset detected the same number as the previous models, while the model using only the CLAHE technique showed the worst performance, detecting 46 bacilli.

In addition to analyzing the number of detections made by each model, the prediction confidence scores were also examined to assess the reliability of each approach in detecting small or low-visibility MT. The model trained with the LUV color channel showed a slight reduction in confidence, with values not exceeding 90%. Similarly, in the segmentation task, some MT were not fully identified. The model based on the RGB color channel yielded comparable results, although some predictions had very low confidence scores and difficulties in segmenting the bacilli. The model trained with the LUV + CLAHE dataset exhibited a slight decrease in confidence values compared to the previous models, in addition to missing some MT and producing duplicate detections of the same bacillus. The model based solely on the CLAHE technique faced significant challenges in both detection and segmentation, with extremely low confidence scores.

These findings demonstrate that, although the quantitative metrics indicate overall good performance, the qualitative analysis reveals limitations in identifying MT with more challenging characteristics. This issue may directly affect the practical applicability of the model, particularly in clinical scenarios where accurate detection of small or low-visibility MT is essential for reliable diagnosis. Therefore, the results suggest the need for improvements in image preprocessing and model training, including strategies such as data augmentation focused on these specific cases, adjustments to the neural network architecture, or the use of advanced segmentation techniques.


(a) RGB COLOR SPACE IMAGE

(b) LUV COLOR SPACE IMAGE

(c) RGB COLOR SPACE IMAGE WITH CLAHE

(d) LUV COLOR SPACE IMAGE WITH CLAHE

Fig. 4. Visual comparison of predictions and segmentations in different datasets

5 Conclusion

The results of this study demonstrated that the YOLOv11 model achieved promising performance in the detection and segmentation of MT in sputum bacilloscopy images. The findings confirm that conversion to the LUV color space enhances the detection and segmentation of bacilli, outperforming traditional approaches. The ability of this color space to separate chrominance from luminance contributed to a more precise and robust identification of MT, minimizing interference caused by lighting variations and improving the distinction between regions of interest and the background. It was observed that the RGB color space dataset yielded results similar to those obtained with LUV, although slightly lower. In contrast, the datasets processed with CLAHE and LUV + CLAHE did not demonstrate the same level of effectiveness in detection and segmentation, presenting inferior metrics compared to the other approaches.

In addition to the advancements demonstrated in this study, there remains room for improvement and future investigation. Optimizing the model for the prediction of low-visibility and small-sized MT could further enhance the robustness of detection and segmentation performance. Thus, this work represents a significant step toward the application of convolutional neural networks in supporting TB diagnosis, paving the way for further research and advancements in the field of artificial intelligence applied to healthcare.

References

1. Gomide, J.V.B., Augusto, C.J., Leal, É.A., Tarabal, J.P., de Castro Barroso, N.V., Soares, M.A.C., Lima, B.P.: Detecção e contagem automáticas de bacilos álcool ácido resistentes para o diagnóstico da tuberculose. Código 31: revista de informação, comunicação e interfaces **1**(1) (2023)
2. Huang, J., Wang, K., Hou, Y., Wang, J.: Lw-yolo11: A lightweight arbitrary-oriented ship detection method based on improved yolo11. Sensors **25**(1), 65 (2024)
3. López, Y.P.: Detecção do mycobacterium tuberculosis em imagens de baciloscopia de campo claro utilizando redes neurais convolutivas (2018)
4. MARTINS, V.D.O., DE MIRANDA, C.V.: Diagnóstico e tratamento medicamentoso em casos de tuberculose pulmonar: revisão de literatura. Revista saúde multidisciplinar **7**(1) (2020)
5. de Oliveira, M.D., Guedes, E.B.: Detecção de lixo em áreas costeiras: Uma aplicação de segmentação com r-cnns da família yolo. In: Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais (WCAMA). pp. 11–20. SBC (2024)
6. Pinto, V.J.: Detecção do bacilo da tuberculose através do processamento e análise de imagem microscópica. Ph.D. thesis, Mestrado em Sistemas de Informação e Gestão do Conhecimento (2018)
7. Reis, F.J.D.S.: Uma abordagem automatizada para detecção e classificação do bacilo da tuberculose (2022)
8. Rodrigues, F.M., Reis, F.J., Veloso, M.A., Diniz, J.O., Veloso, R.R., Antonio Filho, O.: Metodologia automática para detecção de bacilos de tuberculose utilizando retinanet e modelos de cores. In: Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS). pp. 334–345. SBC (2022)
9. S dos Santos, P.R., de C Brito, V., de Carvalho Filho, A.O.: Kochdet: Arquitetura profunda baseada em bifpn para detecção de bacilos de kock. Revista de Sistemas e Computação-RSC **13**(1) (2023)
10. da Saúde Brasil, M.: Tuberculose (2023), https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/t/tuberculose
11. Serrão, M.K.M., Costa, M.G.F., Fujimoto, L.B.M., Ogusku, M.M., Costa Filho, C.F.F.: Automatic bright-field smear microscopy for diagnosis of pulmonary tuberculosis. Computers in Medicine and Biology (2024), accepted for publication, Feb. 15, 2024
12. WHO, W.H.O.: Tuberculosis (2025), https://www.who.int/news-room/fact-sheets/detail/tuberculosis
13. Xiong, Y., Ba, X., Hou, A., Zhang, K., Chen, L., Li, T.: Automatic detection of mycobacterium tuberculosis using artificial intelligence. Journal of thoracic disease **10**(3), 1936 (2018)