

Multimodal deep learning from sputum image segmentation to classify *Mycobacterium tuberculosis* using IUATLD assessment

Nia Saurina^{1,2}, Nur Chamidah^{3,4}, Riries Rulaningtyas⁵, Aryati Aryati⁶

¹Doctoral Study Program of Mathematics and Natural Science, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

²Department of Informatics, Faculty of Engineering, Universitas Wijaya Kusuma Surabaya, Surabaya, Indonesia

³Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

⁴Research Group of Statistical Modeling in Life Science, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

⁵Department of Physics, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

⁶Department of Clinical Pathology, Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia

Article Info

Article history:

Received Sep 3, 2024

Revised Nov 5, 2024

Accepted Nov 19, 2024

Keywords:

Image segmentation

International Union Against

Tuberculosis and Lung Disease

Multimodal deep learning

Sputum

Tuberculosis

ABSTRACT

Tuberculosis (TB) continues to be a major global health issue, especially in areas with limited resources where diagnostic tools are often insufficient. Traditional TB detection methods are slow and lack sensitivity, particularly for early-stage or low bacterial load cases. This study introduces a new multimodal deep learning model that integrates sputum image segmentation across RGB, hue, saturation, and value (HSV), and CIELAB color channels, using the YOLOv8 model for real-time detection and segmentation. The model uses the International Union Against Tuberculosis and Lung Disease (IUATLD) grading scale for accurate *Mycobacterium tuberculosis* (MTB) classification. Our approach shows high accuracy (92.24%) and precise forecasting (mean absolute percent error (MAPE) of 0.23%), greatly enhancing diagnostic speed and reliability. This research offers a novel method for classifying MTB using a multimodal deep learning model that integrates sputum image segmentation across RGB, HSV, and CIELAB color channels. By using the YOLOv8 model for real-time bounding box detection and segmentation, and the IUATLD grading scale for classification, our method achieves high accuracy and precision in identifying TB bacteria. Our findings indicate that this multimodal deep learning approach significantly improves diagnostic accuracy and speed, providing a reliable tool for early TB detection.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Nur Chamidah

Doctoral Study Program of Mathematics and Natural Science, Faculty of Science and Technology

Universitas Airlangga

Mulyorejo, Surabaya, East Java 60115, Indonesia

Email: nur-c@fst.unair.ac.id

1. INTRODUCTION

Tuberculosis (TB), caused by the bacterium *Mycobacterium tuberculosis* (MTB), is an infectious disease that spreads through the air from infected individuals. Nearly a quarter of the global population is infected with this bacterium, with approximately 89% of TB cases occurring in adults and 11% in children. TB remains the leading cause of death after HIV/AIDS [1]. In 2020, an estimated 9.9 million people worldwide were affected by TB. Basic Health Research (Riskesdas) identified Papua, Banten, and West Java as the Indonesian provinces with the highest TB prevalence, at 0.77%, 0.76%, and 0.63%, respectively in

Ministry of Health Republic Indonesia. According to the Decree of the Minister of Health of the Republic of Indonesia No. 364/MENKES/SK/V/2009, TB diagnosis can be performed through microscopic sputum examination. This technique, widely used in most primary health centers (PHCs) in Indonesia, provides faster results compared to other tests [2]. The sputum sample is stained using the Ziehl-Neelsen (ZN) method [3], which turns MTB bacteria red against a blue background, making them clearly visible under a microscope [4].

You only look once (YOLO) is a novel object detection method that allows for the prediction of objects and their locations in an image at a glance [5]. YOLO creates bounding boxes and uses feature extraction, but its selective search is confined to specific locations within the image [6]. YOLO is an object recognition and localization algorithm based on deep learning neural networks [7]. Gao and Qian [8] combined computed tomography (CT) technology with a high-precision classification model for five types of pulmonary TB (PTB) using convolutional neural networks (CNN) and support vector machines. Ma *et al.* [9] developed an automatic detection model for active PTB using U-Net. CNN-based detection models are crucial for AI-assisted diagnosis [10]. The lightweight YOLOv4 model, named M_IcrograPhs-MYcobacterium and called MIP-MY, is used for TB detection [11]. YOLOv5 is a single-stage object recognition algorithm [12]. An enhanced YOLOv5 strategy for TB classification based on whole CT slices incorporates three additional modules: the convolutional block attention module (CBAM) architecture, SCYLLA-IoU (SIoU) loss function, and data augmentation [13]. Segmentation involves dividing an image into non-overlapping regions that are homogeneous according to a criterion, covering the entire image [14]. In this context, a binary object is any set of pixels in a binary image corresponding to a bacillus or bacilli grouping [15]. A multimodal deep learning classifier leverages information from multiple data modalities [16], aiming to utilize their heterogeneous nature through intermediate fusion [17]. Zhang *et al.* [18] proposed dual-wing harmoniums to learn a joint representation of image and text modalities. Zhen and Yeung [19] introduced a probabilistic generative approach called multimodal latent binary embedding. Limited data resources are a significant concern in the first category [20]. Modalities can increase parameter counts, resulting in high training accuracy but low test accuracy for multimodal systems [21].

This study focuses on the automatic identification of MTB in sputum images, utilizing color channel detection and multimodal deep learning. Chamidah *et al.* [22] presents a method for automatically counting MTB in sputum images, which is more efficient than manual counting by pathologists. It uses a nonparametric Poisson regression model with a local linear estimator. Hema and Kannan [23] highlights the effectiveness of the hue, saturation, and value (HSV) color space for image segmentation, which is superior to other color models for extracting key foreground elements from color images. An interactive Python-based graphical user interface (GUI) tool was developed, allowing users to adjust HSV interactively for optimal segmentation results. Torres-González *et al.* [24] developed StainView, a fast and reliable method for mapping stains on building facades using image classification in HSV and CIELab color spaces. It enables the automatic location and mapping of critical areas with high efficiency, reducing inspection time and human error. Rao *et al.* [25] introduces a new algorithm for segmenting fine details in the CIELAB color space, considering human vision's contrast sensitivity, aiming to improve the precision and applicability of color difference metrics in image processing. Radu *et al.* [26] introduces deep autoencoder models that can reconstruct both modalities when given only one, demonstrating robustness to missing modalities and effective multimodal fusion. Shaban and Yousefi [27] introduces a novel deep learning architecture that emphasizes learning both intra-modality and cross-modality relations, which had not been previously applied to wearable modeling tasks. The research includes a novel architecture that integrates convolutional layers within the modality-specific architecture, enhancing the ability to capture complex patterns from multimodal data. Another study on multimodal learning from [28] presented a model based on a Bayesian nonparametric model to learn the underlying semantically meaningful and abstract features of multimodal data. Multimodal deep learning has been used by [29], integrating unimodal CNNs for music and video into multimodal structures using a late fusion strategy, enhancing the accuracy of emotion classification. It provides a detailed comparative analysis of various unimodal and multimodal CNN architectures, identifying the best models for emotion classification. Choi *et al.* [30] introduces a novel approach to multimodal classification by incorporating a second objective using variational inference. These innovations aim to address the challenges of overfitting and limited data in multimodal systems. Minyilu *et al.* [31] introduces a multimodal diagnostic model for predicting PTB, combining traditional diagnostic methods with deep learning-based automated detection algorithms (DLADs) to improve the accuracy and efficiency of TB diagnosis. Examination and assessment of the acid-fast bacillus (AFB) smear sputum were carried out by expert officers in each health facility designated as a health facility for TB service programs by the government. Ramirez-Hidalgo *et al.* [32] introduces a new scoring system, the Pulmonary Tuberculosis Sputum Score (PTBSCore), to predict the length of the infectious period in patients with PTB. This score is based on clinical, radiological, and analytical parameters.

Regrettably, there has been no research on image segmentation of MTB in sputum using multimodal deep learning across three different color channels and the International Union Against Tuberculosis and

Lung Disease (IUATLD) assessment. Sputum examination results were interpreted using the IUATLD grading scale, categorizing subjects into 1+, 2+, and 3+ groups based on the number of AFB observed [33]. The staining was examined with a 1000x magnification microscope by applying immersion oil to the sample. BTA bacteria appear brick pink, while non-BTA bacteria appear blue. The number of AFB was read according to the IUATLD scale [34], as shown in Table 1.

Table 1. Scale of IUATLD [34]

Readings under the microscope	Reporting of results
BTA was not found in 100 laps of vision	Negative
1-9 BTA in 100 fields of view	Scanty
10-99 BTA in 100 fields of view	+1 (positive 1)
1-10 BTA in 1 field of view	+2 (positive 2)
>10 BTA in 1 field of view	+3 (positive 2)

Although there have been significant advancements in TB detection, several challenges remain. This research aims to develop a new multimodal model for the early diagnosis and accurate prediction of TB by using sputum image segmentation across three color channels: RGB, HSV, and CIELAB. TB continues to be a major global health issue, especially in developing countries where timely and accurate diagnosis is essential for effective treatment and control. Traditional diagnostic methods often lack speed and accuracy, prompting the need for advanced techniques. This study utilizes the latest advancements in multimodal deep learning, specifically the YOLOv8 model, to improve the real-time detection and segmentation of sputum images. By integrating features from multiple color channels, the study aims to extract more distinctive features, thereby enhancing diagnostic accuracy. The novelty of this approach lies in its multimodal integration and the application of deep learning to a traditionally manual process, potentially revolutionizing TB diagnostics. The model classifies results using the IUATLD scale, indicating negative and positive (1+, 2+, 3+) outcomes.

The major contribution of this work as follows:

- A novel multimodal model is proposed for early diagnosis and accurate prediction of Mycobacterium TB patients based on sputum images segmentation from three different canal which is RGB, HSV and CIELAB.
- In the proposed model the sputum images segmentation will be detected using bounding box with YOLOv8 to produce image and coordinate from each canal.
- The model integrates, sputum images segmentation model from three different canal model to extract discriminatory features such as texture and shape features from the sputum images segmentation using multimodal deep learning techniques.
- The proposed model uses IUATLD assessment to counting the total of MTB's sputum to classify negative and positive (1+, 2+, 3+) of TB.

The remaining portions of this article are broken down into the following sections: the second section digs into the details of our designed methodology. Section 3 explores and analyzes the experimental findings and discussion about substantial and intellectual contribution. Finally, the conclusions and future works are found in section 4.

2. METHOD

In this study, data collection involved obtaining 1,265 microscopic images of TB from the Department of Clinical Pathology at the Faculty of Medicine, Universitas Airlangga. The TB classification process consists of multiple stages, as illustrated in Figure 1, beginning with the input of image data. The dataset was labeled by identifying TB bacteria locations within each image using a bounding box approach through YOLO. All images were resized from 1632×1442 to 640×480 to standardize them for the multimodal deep learning model. These resized images were then processed through three color channels- RGB, HSV, and CIELAB to capture a broader range of features. The YOLOv8 model was used for classification, enabling both detection and segmentation of TB bacteria in the images. Model accuracy was assessed using mean average precision (mAP) criteria, and the IUATLD evaluation with mean absolute percent error (MAPE) criteria was applied to count MTB, providing classifications as negative, 1+, 2+, or 3+.

2.1. Labelling data

During the data labeling stage, images are resized to standardize their dimensions from 1632×1442 to 640×480, followed by labeling using YOLO's bounding box method. After generating the YOLO-labeled dataset, color segmentation is applied. The RGB color image is first converted to grayscale and then

transformed into a binary image. Using deep learning-based segmentation methods enhances accuracy and reduces processing time [35]. Once segmentation is completed, the processed images are fed into a multimodal deep learning model for further analysis.

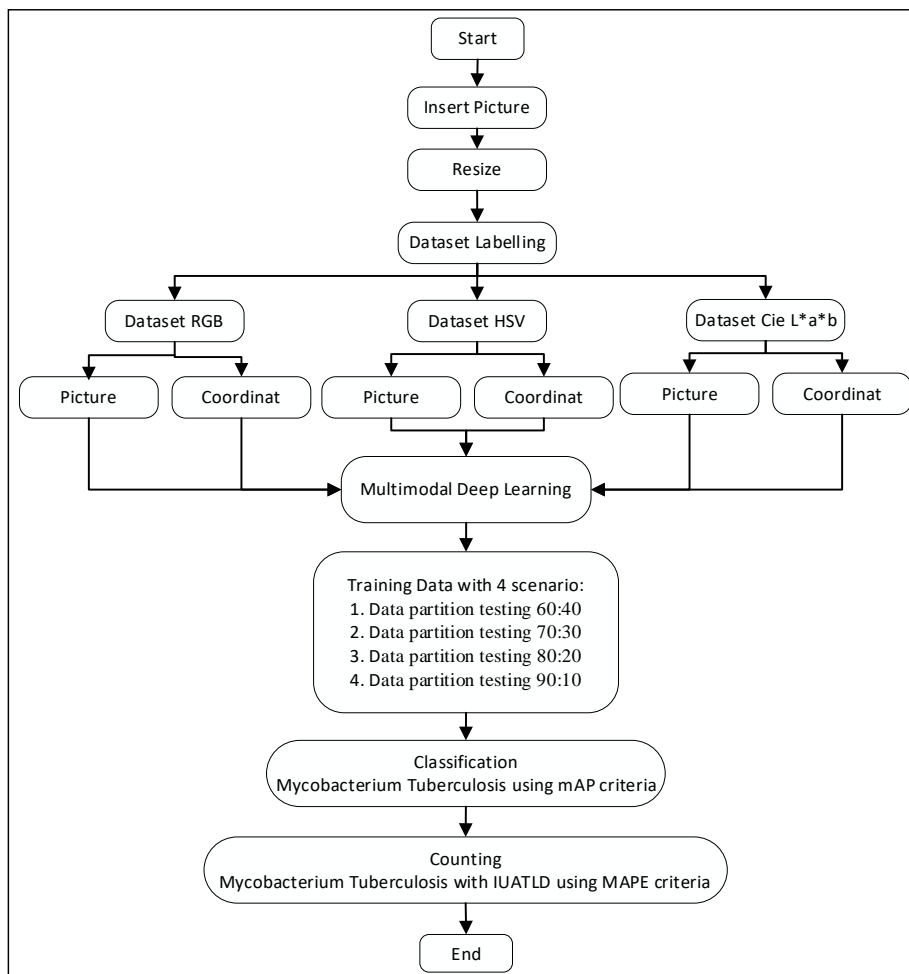


Figure 1. Flow chart research method detection MTB

2.2. Image segmentation

Deep learning has now become a widely accepted and powerful approach for image segmentation, frequently utilized to separate uniform areas as a fundamental step in diagnostic and treatment workflow. Segmentation involves partitioning an image into distinct regions to identify objects or areas of interest, aiming to simplify the representation of the image into clear, meaningful sections [36]. The HSV (Hue, Saturation, Value) color space aligns closely with the RGB color space, reflecting how humans perceive and describe color sensations, which often makes it preferable for color image segmentation [37]. CIE Lab* (CIE L*a*b*) is the most comprehensive color spaces defined by the International Commission on Illumination, capturing all colors visible to the human eye [38]. Designed as a device-independent reference, CIELAB uses three coordinates: L* for lightness (from 0 for black to 100 for diffuse white), a* for the red/magenta to green spectrum (with positive values indicating magenta and negative values indicating green), and b* for the yellow to blue range (positive for yellow and negative for blue). The results of image segmentation in RGB, HSV, and CIELAB are shown in Figure 2.

2.3. Classification using multimodal deep learning

To build a model using the YOLO technique, a dataset is required for training. In this research, the dataset comprises 1,265 microscopic images of TB, obtained through Acid Fast Bacteria staining using the ZN method to assist in the initial microscopic diagnosis of TB. The fixation process in this staining technique

opens the bacterial cell walls, allowing them to absorb the coloring agents used. In total, 3,795 images were utilized for this study, divided into 1,265 images each in RGB, HSV, and CIE LAB color spaces. Combining multiple modalities enables the deep learning model to better interpret its environment, as certain features are present only within specific modalities, as shown in Figure 3.

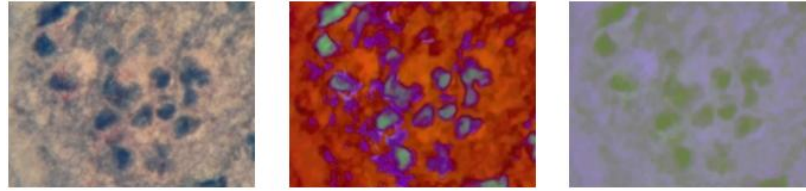


Figure 2. MTB in RGB, HSV and CIE LAB

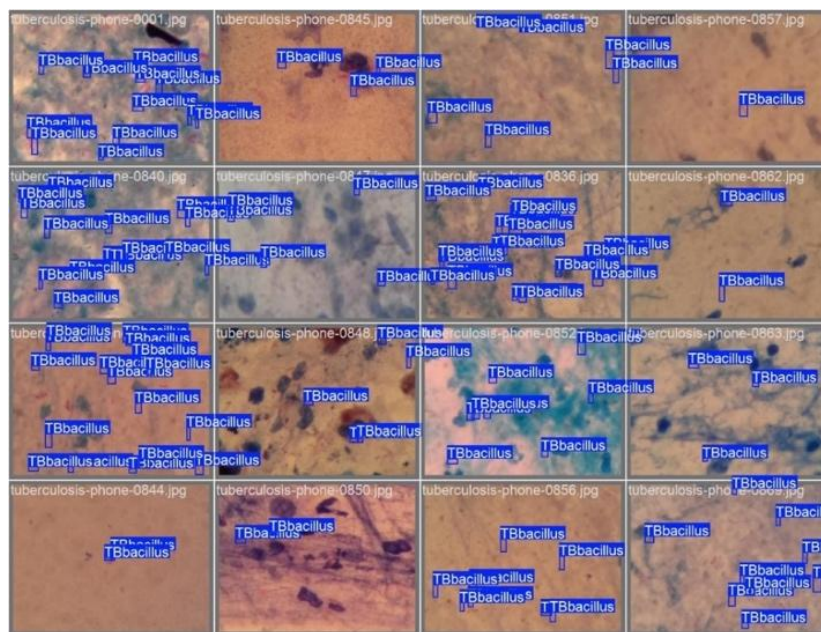


Figure 3. Result of classification of MTB with multimodal deep learning

2.4. Counting *Mycobacterium tuberculosis* with International Union Against Tuberculosis and Lung Disease

MTB counts are determined using the classification results, with four models created based on four different data partitioning scenarios. Grading of TB in sputum samples follows the IUATLD scale, defined as follows: negative (no AFB observed in 100 fields), 1+ (10-99 AFB in 100 fields), 2+ (1-9 AFB per field in at least 50 fields), and 3+ (more than 10 AFB per field in at least 20 fields).

3. RESULTS AND DISCUSSION

3.1. Results

A confusion matrix is used to assess the performance of a machine learning model by comparing actual classifications with predicted classifications [39]. It consists of four components: true negative (TN), true positive (TP), false negative (FN), and false positive (FP), each representing a combination of actual and predicted outcomes. TP indicates the count of correctly classified positive samples, TN is the count of accurately classified negative samples, FP refers to negative samples incorrectly labeled as positive, and FN represents positive samples incorrectly labeled as negative [40]. Figure 4 shows the confusion matrix for MTB classification using multimodal deep learning. The dataset partitioning for testing follows the holdout method, splitting data into training and test sets. The dataset includes 3,795 samples across two classes, stored in training and testing folders according to predetermined portions. This training process enables the

model to learn and recognize predefined objects. The dataset will undergo four training scenarios, outlined as follows:

- Data train 60% (2277 data) and data test 40% (1518 data).
- Data train 70% (2656 data) and data test 30% (1139 data).
- Data train 80% (3036 data) and data test 20% (759 data).
- Data train 90% (3415 data) and data test 10% (380 data).

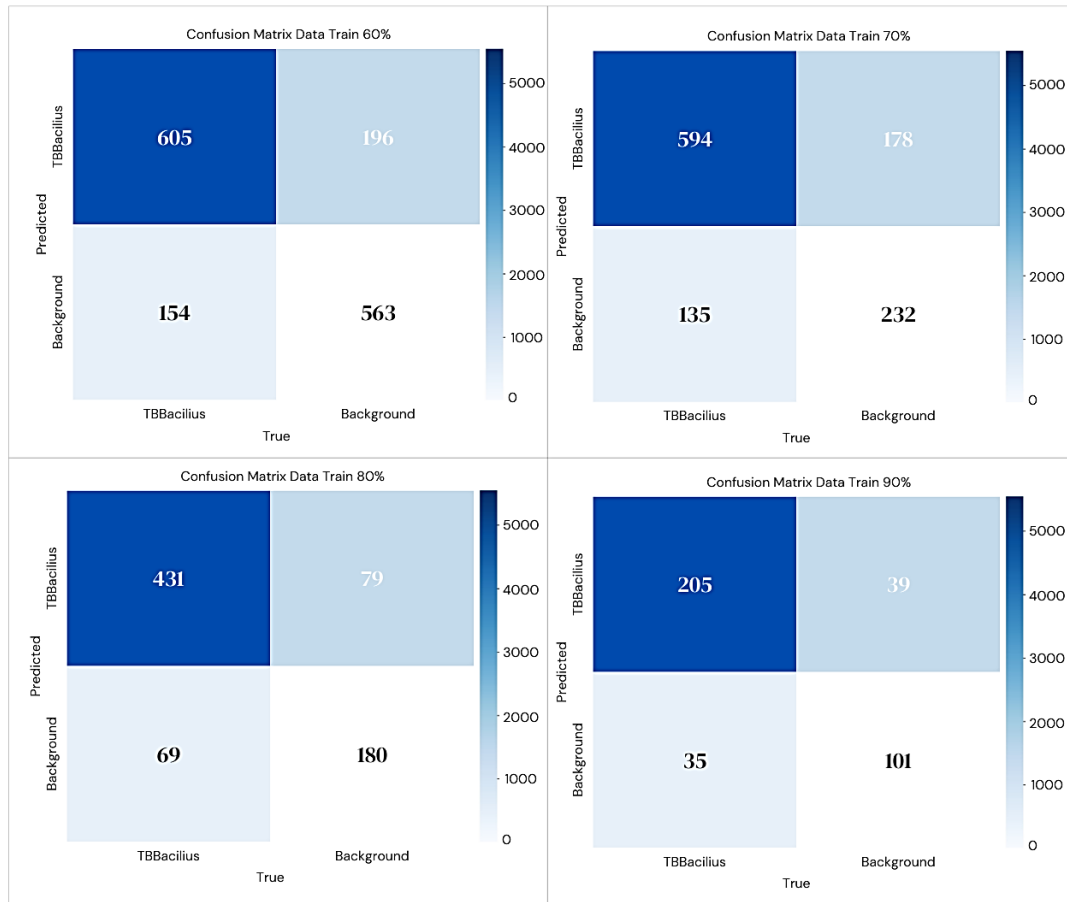


Figure 4. Confusion matrix from classification MTB with multimodal deep learning

Figure 4 presents the confusion matrix for MTB classification across four data partition scenarios. Using (1) to (4), we calculate metrics such as accuracy, precision, recall, F1-score, and mAP. In the first data partition (60:40), we obtained 605 TP, 196 FP, 154 FN, and 563 TN. For the second partition (70:30), the values were 594 TP, 178 FP, 135 FN, and 232 TN. In the third partition (80:20), there were 431 TP, 79 FP, 69 FN, and 180 TN. Finally, for the fourth partition (90:10), the results showed 205 TP, 39 FP, 35 FN, and 101 TN. These values enable the calculation of precision, accuracy, recall, F1-score, and mAP.

Precision is the ratio of TP to the total number of predicted positive data. In the denominator, there is the variable FP as the divisor. This can be written using (1) [41]:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

Accuracy is the percentage of correctly identified cases. Thus, can be written using (2) [41]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

On the other hand, recall is defined as the ratio of TP to the total number of actually positive instances. The denominator includes FN as the divisor, and it can be written using (3) [42]:

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

When recall is very high, precision will be very low, and vice versa. There is a trade-off relationship between precision and recall. This trade-off relationship implies that the sum of these two variables equals 1. The harmonization of the average between precision and recall is called the F1-score. Based on (4) [42], the best value for the F1-score is 1.0, while the worst value is 0.0.

$$F1Score = \frac{2xPrecisionxRecall}{Precision+Recall} \tag{4}$$

mAP averages the precision and recall scores for each object class to determine the overall accuracy of the object detector. This metric represents the average of the average precision (AP) calculated for all the classes being detected, and it can be written using (5) [42]:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \tag{5}$$

Interpretation criteria of mAP, can be seen in Table 2 where mAP>70% which means that model very accurate. 50%≤mAP≤70% which means that the model has adequate performance and mAP<50% which means that the model not good.

Table 2. Interpretation criteria of mAP [42]

Interpretation (%)	Value
mAP>70	Very accurate. The model is very accurate in detecting and classifying objects.
50≤mAP≤70	Good. The model has adequate performance, but there may still be room for improvement.
mAP<50	Not good. The model has significant problems detecting or classifying objects correctly.

Figure 5 illustrates the accuracy scores across four test scenarios: data partition 1 achieved 76.94%, data partition 2 scored 72.52%, and both data partition 3 and data partition 4 reached 80.5%. Precision scores were as follows: 75.53% for data partition 1, 76.94% for data partition 2, 84.51% for data partition 3, and 84.01% for data partition 4. Recall scores showed 81.48% for both data partitions 1 and 2, 86.2% for data partition 3, and 85.42% for data partition 4. F1 scores were 77.56% for data partition 1, 79.14% for data partition 2, 85.34% for data partition 3, and 84.71% for data partition 4. The mAP scores were 84.14% in data partition 1, 83.34% in data partition 2, 92.24% in data partition 3, and 87.39% in data partition 4.

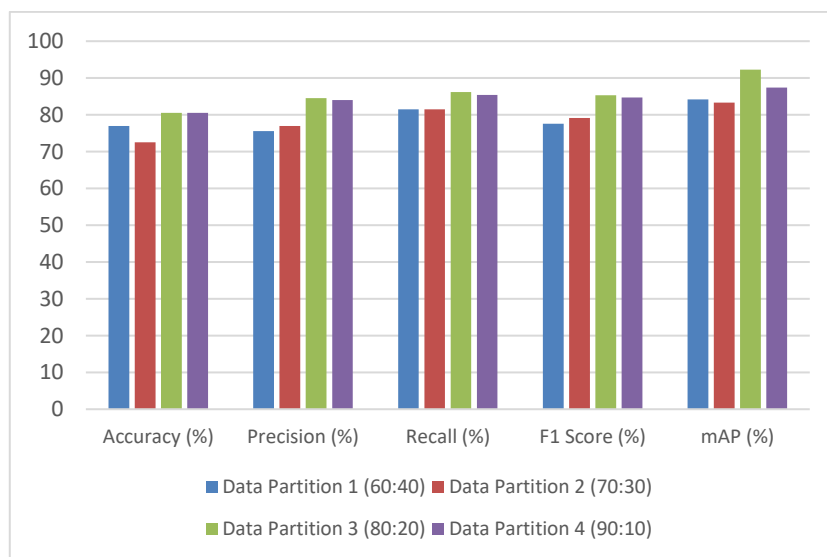


Figure 5. Results of data partition testing

The IUATLD calculation process on 3795 images was obtained from 13 patients, where every 100 images were the result of taking MTB sputum from one patient. Dataset in this study consists of 3795 images, where 1265 images in RGB, 1265 images in HSV and 1265 images in CIELAB, then there are 13 patients who can be classified using IUATLD calculations. The classification results used for IUATLD calculations in Table 2 use a classification model from 80:20 data partition, which has the highest score from Table 3.

Table 3. Results of IUATLD from the third data partition testing

No	Number of visual fields with condition 1+	Number of visual fields with condition 2+	Number of visual fields with condition 3+	Score IUATLD
1	97	70	27	2+
2	98	76	22	2+
3	100	75	25	2+
4	100	80	20	3+
5	96	71	25	3+
6	95	68	27	3+
7	98	67	31	3+
8	97	70	27	3+
9	99	82	17	2+
10	98	74	24	3+
11	98	75	23	3+
12	99	75	24	3+
13	37	27	10	1+

Results of the IUATLD calculation, as can be seen in Table 3, from the 13 patients there was one patient who was categorized as 1+, then one patient who were categorized as 2+ and eleven patients who were categorized as 3+. Based on [43] a MAPE value of 3.9% has a very accurate interpretation of forecasting results. The MAPE value in this study was obtained by comparing the number of MTB classifications produced by the proposed model with the number of MTB calculated by Department of Clinical Pathology, Faculty of Medicine, Universitas Airlangga in 1265 images. This research already noted in Ethical Exemption Number 53/EC/KEPK/FKUA/2023. The difference between the observed value and the forecast value is often referred to as the residual. Various techniques in measuring accuracy both in numbers and in percent such as MAPE. Calculations with MAPE are carried out with absolutes in the form of a percent divided by a lot of data to be measured by period, is shown in (6).

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - F_i}{A_i} \right| \times 100\% \quad (6)$$

where n is sample size, A_i is actual data value, and F_i is forecasting data value

MAPE is scale-independent and easy to interpret, which makes it popular with industry practitioners. It is recommended in most textbooks. The interpretation of the MAPE value is <10%, including very accurate forecasting; 10-20%, including good forecasting; 20-50%, including reasonable forecasting; and >50%, including inaccurate forecasting, as seen in Table 4 [43].

From the Table 5, we can see the results of MAPE in data partition 60:40 has MAPE score in 0.76%. In data partition 70:30 has MAPE score in 0.69%. In data partition 80:20 has MAPE score in 0.23% and in data partition 90:10 has MAPE score in 0.53%. The interpretation of the MAPE value is <10%, including very accurate forecasting; 10-20%, including good forecasting; 20-50%, including reasonable forecasting; and >50%, including inaccurate forecasting [43]. The best MAPE score in this research is 0.23% which means very accurate forecasting.

Table 4. Interpretation criteria of MAPE [43]

Interpretation	Value
MAPE<10%	Very accurate forecasting
10-20%	Good forecasting
20-50%	Reasonable forecasting
MAPE>50%	Inaccurate forecasting

Table 5. Results of MAPE

No	YOLO	Data partition (%)	MAPE (%)
1	YOLOv8	60:40	0.76
2		70:30	0.69
3		80:20	0.23
4		90:10	0.53

3.2. Discussion

In epidemiology, health issues can be analyzed from various angles, such as person, time, and place [44]. TB remains a leading cause of mortality globally, with young children at the highest risk of infection,

highlighting the urgent need for improved diagnostic and treatment methods [45]. This research is unique in its integrative approach, combining multiple color spaces to boost diagnostic accuracy. Unlike traditional methods that rely on a single color space, this model incorporates RGB, HSV, and CIELAB simultaneously. Recent advancements also include hyperspectral imaging or image clustering using the HSV color space, which aids in detecting biological colonization and assessing color variations on surfaces [46]. The innovative use of bounding boxes for detecting and segmenting sputum images offers several advantages:

- Enhanced precision and speed: YOLOv8's architecture allows for rapid and precise detection of objects within images. This is crucial in medical diagnostics where timely and accurate analysis can significantly impact patient outcomes [44].
- Real-time processing: the ability of YOLOv8 to process images in real-time ensures that the segmentation of sputum images can be performed quickly, facilitating faster diagnosis and treatment planning [47].
- Detailed localization: by producing both images and coordinates for each canal, the model provides detailed localization of the segmented areas. This level of detail is essential for thorough analysis and can aid in identifying specific regions of interest within the sputum samples [46].
- Scalability and adaptability: the model's reliance on YOLOv8 means it can be easily adapted to other types of medical image segmentation tasks. This scalability makes the approach versatile and applicable to a wide range of diagnostic applications [44].

The proposed model introduces an innovative method for sputum image segmentation by combining models from three different channels to extract distinguishing features through multimodal deep learning techniques. By incorporating segmentation models across these three channels, the model achieves a thorough analysis of sputum images, which is essential for differentiating among various types of sputum samples and enhancing diagnostic and disease classification accuracy. The use of multimodal deep learning enables effective processing and integration of information from diverse sources, enhancing the model's capacity to learn complex patterns and relationships in the data, resulting in improved segmentation outcomes [48]. This combination of multiple channel models with multimodal deep learning signifies a notable advancement in medical image analysis, providing a powerful tool for extracting critical features from sputum images and contributing to more precise and efficient diagnostics.

This study introduces an innovative multimodal deep learning model designed for the detection and classification of MTB in sputum images. By combining image segmentation from RGB, HSV, and CIELAB color channels and employing the YOLOv8 model, the researchers attained an impressive accuracy of 92.24% and a MAPE of 0.23% for precise predictions. The model's effectiveness was assessed through four different data partition scenarios, with the 80:20 partition providing the optimal outcomes. The IUATLD grading scale was utilized to evaluate TB severity, resulting in accurate classification of patients into negative, 1+, 2+, and 3+ categories.

The findings of this research have important implications for TB diagnostics. By integrating multimodal image segmentation utilizing RGB, HSV, and CIELAB color channels, we have established a novel approach that significantly improves both diagnostic accuracy and speed. Our critical comparison with traditional single-channel methods demonstrates the advantages of our multimodal model in effectively capturing the complex features of sputum samples. This advancement not only overcomes the limitations of current diagnostic techniques but also opens avenues for future studies to investigate similar multimodal strategies in other medical imaging fields. The application of the YOLOv8 model for real-time detection and segmentation further emphasizes the potential for implementing this technology in clinical environments, ultimately enhancing patient outcomes through timely and accurate diagnoses. The model also uses the IUATLD grading scale to assess the severity of TB infection, offering a standardized and precise method for counting and classifying MTB, thus improving diagnostic reliability.

This article has several limitations that should be acknowledged. First, the research relied on a limited dataset of 3,795 images collected from just 13 patients, which may not adequately represent a larger population. Second, the focus on specific color channels (RGB, HSV, and CIELAB) for image segmentation might not encompass all pertinent features, indicating a need for additional modalities or improvements to the existing method. Finally, external validation is lacking; the findings have not been thoroughly tested across diverse clinical and geographical contexts, which raises questions about their generalizability and applicability.

4. CONCLUSION

This research introduces an innovative method for classifying MTB through a multimodal deep learning model that incorporates sputum image segmentation across RGB, HSV, and CIELAB color channels. Utilizing the YOLOv8 model for real-time bounding box detection and segmentation, along with the IUATLD grading scale for classification, our approach achieves remarkable accuracy of 92.24% and a

MAPE of 0.23%. The findings highlight that the multimodal deep learning approach significantly improves both diagnostic accuracy and speed, serving as a dependable tool for early TB detection. This method harnesses advanced image processing and machine learning techniques, making it applicable in various healthcare environments, even those with limited resources. This study showcases the potential of combining multimodal image processing and deep learning to enhance diagnostic results, paving the way for further exploration of artificial intelligence applications in medical diagnostics, particularly for infectious diseases. For communities in high TB burden regions, our approach presents a promising solution to the difficulties associated with TB diagnosis. By offering a more precise and accessible diagnostic tool, we can enhance early detection and treatment outcomes, ultimately curbing the spread of TB and bolstering public health.

The proposed model has the potential to be adapted for other infectious diseases that necessitate image-based diagnostics. Future studies might investigate the incorporation of additional data modalities, such as genomic information, to further improve diagnostic precision. Furthermore, creating portable diagnostic devices that utilize this technology could transform point-of-care testing in remote and resource-constrained environments.

ACKNOWLEDGEMENTS




This research received funding from PUSLAPDIKTI through the *Beasiswa Pendidikan Indonesia* (BPI) program, as outlined in individual decision letter Number 00170/BPPT/BPI.06/9/2023 for the year 2023.

REFERENCES




- [1] World Health Organization, "Monitoring Health For The SDGs," 2021.
- [2] R. O. Panicker, B. Soman, G. Saini, and J. Rajan, "A Review of Automatic Methods Based on Image Processing Techniques for Tuberculosis Detection from Microscopic Sputum Smear Images," *Journal of Medical Systems*, vol. 40, no. 1, p. 17, Jan. 2016, doi: 10.1007/s10916-015-0388-y.
- [3] R. Kurniawan, I. Muhimmah, A. Kurniawardhani, and S. Kusumadewi, "Segmentation of Tuberculosis Bacilli Using Watershed Transformation and Fuzzy C-Means," *CommIT (Communication and Information Technology) Journal*, vol. 13, no. 1, p. 9, May 2019, doi: 10.21512/commit.v13i1.5119.
- [4] M. I. Shah *et al.*, "Ziehl-Neelsen sputum smear microscopy image database: a resource to facilitate automated bacilli detection for tuberculosis diagnosis," *Journal of Medical Imaging*, vol. 4, no. 2, p. 027503, Jun. 2017, doi: 10.1117/1.JMI.4.2.027503.
- [5] J. Du, "Understanding of Object Detection Based on CNN Family and YOLO," *Journal of Physics: Conference Series*, Apr. 2018, vol. 1004, p. 012029, doi: 10.1088/1742-6596/1004/1/012029.
- [6] C. Hofmann, F. Particke, M. Hiller, and J. Thielecke, "Object Detection, Classification and Localization by Infrastructural Stereo Cameras," in *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, SCITEPRESS - Science and Technology Publications, 2019, pp. 808–815, doi: 10.5220/0007370408080815.
- [7] J. Kaur and W. Singh, "A systematic review of object detection from images using deep learning," *Multimedia Tools and Applications*, vol. 83, no. 4, pp. 12253–12338, Jan. 2024, doi: 10.1007/s11042-023-15981-y.
- [8] X. W. Gao and Y. Qian, "Prediction of Multidrug-Resistant TB from CT Pulmonary Images Based on Deep Learning Techniques," *Molecular Pharmaceutics*, vol. 15, no. 10, pp. 4326–4335, Oct. 2018, doi: 10.1021/acs.molpharmaceut.7b00875.
- [9] L. Ma *et al.*, "Developing and verifying automatic detection of active pulmonary tuberculosis from multi-slice spiral CT images based on deep learning," *Journal of X-Ray Science and Technology*, vol. 28, no. 5, pp. 939–951, Sep. 2020, doi: 10.3233/XST-200662.
- [10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
- [11] Z. Guo, J. Wang, J. Wang, and J. Yuan, "Lightweight YOLOv4 with Multiple Receptive Fields for Detection of Pulmonary Tuberculosis," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–13, Mar. 2022, doi: 10.1155/2022/9465646.
- [12] Y. Huang, A. Kruyer, S. Syed, C. B. Kayasandik, M. Papadakis, and D. Labate, "Automated detection of GFAP-labeled astrocytes in micrographs using YOLOv5," *Scientific Reports*, vol. 12, no. 1, p. 22263, Dec. 2022, doi: 10.1038/s41598-022-26698-7.
- [13] J. Liu *et al.*, "CT Image Detection of Pulmonary Tuberculosis Based on the Improved Strategy YOLOv5," *International Journal of Swarm Intelligence Research*, vol. 14, no. 1, pp. 1–12, Aug. 2023, doi: 10.4018/IJSIR.329217.
- [14] C. Guichaoua, P. Lascabettes, and E. Chew, "End-to-End Bayesian Segmentation and Similarity Assessment of Performed Music Tempo and Dynamics without Score Information," *Music & Science*, vol. 7, Jan. 2024, doi: 10.1177/20592043241233411.
- [15] Z. Mahmood, "Digital Image Processing: Advanced Technologies and Applications," *Applied Sciences*, vol. 14, no. 14, p. 6051, Jul. 2024, doi: 10.3390/app14146051.
- [16] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges," *Journal of Digital Imaging*, vol. 32, no. 4, pp. 582–596, Aug. 2019, doi: 10.1007/s10278-019-00227-x.
- [17] D. Ramachandram and G. W. Taylor, "Deep Multimodal Learning: A Survey on Recent Advances and Trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, Nov. 2017, doi: 10.1109/MSP.2017.2738401.
- [18] Y. Zhang, T. Zuo, L. Fang, J. Li, and Z. Xing, "An Improved MAHAKIL Oversampling Method for Imbalanced Dataset Classification," *IEEE Access*, vol. 9, pp. 16030–16040, 2021, doi: 10.1109/ACCESS.2020.3047741.
- [19] Y. Zhen and D.-Y. Yeung, "A probabilistic model for multimodal hash function learning," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2012, pp. 940–948, doi: 10.1145/2339530.2339678.

- [20] S. Malik, K. Muhammad, and Y. Waheed, "Artificial intelligence and industrial applications-A revolution in modern industries," *Ain Shams Engineering Journal*, vol. 15, no. 9, p. 102886, Sep. 2024, doi: 10.1016/j.asej.2024.102886.
- [21] W. C. Sleeman, R. Kapoor, and P. Ghosh, "Multimodal Classification: Current Landscape, Taxonomy and Future Directions," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1-31, doi: 10.1145/3543848.
- [22] N. Chamidah, Y. S. Yonani, E. Ana, and B. Lestari, "Identification the number of Mycobacterium tuberculosis based on sputum image using local linear estimator," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 5, pp. 2109-2116, Oct. 2020, doi: 10.11591/eei.v9i5.2021.
- [23] D. Hema and Dr. S. Kannan, "Interactive Color Image Segmentation using HSV Color Space," *Science & Technology Journal*, vol. 7, no. 1, pp. 37-41, Jan. 2019, doi: 10.22232/stj.2019.07.01.05.
- [24] M. Torres-González, J. Valença, B. O. Santos, A. Silva, and M. P. Mendes, "StainView: A Fast and Reliable Method for Mapping Stains in Facades Using Image Classification in HSV and CIELab Colour Space," *Remote Sensing*, vol. 15, no. 11, p. 2895, Jun. 2023, doi: 10.3390/rs15112895.
- [25] K. K. Rao, R. K. B. S. Rao, and L. K., "Development of ExG, ExR, ExGR, HSV, CIELAB Images from RGB Images Using Image Segmentation Algorithm in Computer Vision Based Herbicide Spraying Applications," *Journal of Scientific Research and Reports*, vol. 30, no. 10, pp. 501-508, Oct. 2024, doi: 10.9734/jjsrr/2024/v30i102477.
- [26] V. Radu *et al.*, "Multimodal Deep Learning for Activity and Context Recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 1-27, Jan. 2018, doi: 10.1145/3161174.
- [27] A. Shaban and S. Yousefi, "Multimodal Deep Learning," *Multimodal and Tensor Data Analytics for Industrial Systems Improvement*, 2024, pp. 209-219, doi: 10.1007/978-3-031-53092-0_10.
- [28] S. Kumari and M. P. Singh, "A Deep Learning Multimodal Framework for Fake News Detection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16527-16533, Oct. 2024, doi: 10.48084/etasr.8170.
- [29] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2887-2905, Jan. 2021, doi: 10.1007/s11042-020-08836-3.
- [30] S. Y. Choi, A. Choi, S.-E. Baek, J. Y. Ahn, Y. H. Roh, and J. H. Kim, "Effect of multimodal diagnostic approach using deep learning-based automated detection algorithm for active pulmonary tuberculosis," *Scientific Reports*, vol. 13, no. 1, p. 19794, Nov. 2023, doi: 10.1038/s41598-023-47146-0.
- [31] Y. Minyilu, M. Abebe, and M. Meshesha, "Applying Multimodal Data Fusion based on Deep Learning Methods for the Diagnoses of Neglected Tropical Diseases: A Systematic Review," *medRxiv*, pp. 1-24, Jan. 2024, doi: 10.1101/2024.01.07.24300957.
- [32] M. Ramirez-Hidalgo *et al.*, "Time to sputum conversion in patients with pulmonary tuberculosis: A score to estimate the infectious period," *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases*, vol. 31, p. 100361, May 2023, doi: 10.1016/j.jctube.2023.100361.
- [33] O. N. Putra, A. Damayanti, N. W. D. Nurrahman, T. Devi, and W. Aluf, "Evaluation of Category I of Anti-tuberculosis Therapy in Intensive Phase Pulmonary TB by Conversion of Acid-Fast Bacilli Sputum," *Pharmaceutical Sciences and Research*, vol. 6, no. 3, Dec. 2019, doi: 10.7454/psr.v6i3.4483.
- [34] Misnarliah, Mudrika, and A. A. Basir, "Effect of Preparate Coloring Delay Achid Resistant Bacteria With Ziehl Neelsen Method On The Result of Microscopic Examination," *International Journal of Science, Technology & Management*, vol. 2, no. 2, pp. 536-541, Feb. 2021, doi: 10.46729/ijstm.v2i2.181.
- [35] M. M. Mahasin, A. Naba, C. S. Widodo, and Y. Yueniwati, "Development of a Modified UNet-Based Image Segmentation Architecture for Brain Tumor MRI Segmentation," *Proceedings of the International Conference of Medical and Life Science (ICoMELISA 2021)*, 2023, pp. 37-43, doi: 10.2991/978-94-6463-208-8_7.
- [36] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-1, 2021, doi: 10.1109/TPAMI.2021.3059968.
- [37] M. Singh *et al.*, "Evolution of Machine Learning in Tuberculosis Diagnosis: A Review of Deep Learning-Based Medical Applications," *Electronics*, vol. 11, no. 17, p. 2634, Aug. 2022, doi: 10.3390/electronics11172634.
- [38] A. Ray and M. H. Kolekar, "Image Segmentation and Classification Using Deep Learning," in *Machine Learning Algorithms for Signal and Image Processing*, Wiley, 2022, pp. 19-36, doi: 10.1002/9781119861850.ch2.
- [39] I. P. Sary, S. Andromeda, and E. U. Armin, "Performance Comparison of YOLOv5 and YOLOv8 Architectures in Human Detection using Aerial Images," *Ultima Computing : Jurnal Sistem Komputer*, pp. 8-13, Jun. 2023, doi: 10.31937/sk.v15i1.3204.
- [40] E. Casas, L. Ramos, C. Romero, and F. Rivas-Echeverría, "A comparative study of YOLOv5 and YOLOv8 for corrosion segmentation tasks in metal surfaces," *Array*, vol. 22, p. 100351, Jul. 2024, doi: 10.1016/j.array.2024.100351.
- [41] F. M. Juan, C. P. Carolina, C. O. Patricia, and P. V. Carlos, "Collaborative desing in web application development to improve tuberculosis diagnostic," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 3, p. 1821, Jun. 2023, doi: 10.11591/ijeecs.v30.i3.pp1821-1828.
- [42] B. O. Santos, J. Valença, and E. Júlio, "Classification of cracks of biological colonization on concrete surface using false colour HSV images, including near-infrared information," in *Proceedings, Optical Sensing and Detection V*, May 2018, vol. 10680, p. 2, doi: 10.1117/12.2307728.
- [43] A. R. Lubis, S. Prayudani, Y. Fatmi, M. Lubis, and Al-Khowarizmi, "MAPE accuracy of CPO Forecasting by Applying Fuzzy Time Series," in *2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, IEEE, Oct. 2021, pp. 370-373, doi: 10.23919/EECSI53397.2021.9624303.
- [44] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1680-1716, Nov. 2023, doi: 10.3390/make5040083.
- [45] V. A. Kich *et al.*, "Precision and Adaptability of YOLOv5 and YOLOv8 in Dynamic Robotic Environments," in *2024 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE International Conference on Robotics, Automation and Mechatronics (RAM)*, IEEE, Aug. 2024, pp. 514-519, doi: 10.1109/CIS-RAM61939.2024.10673292.
- [46] M. Talib, A. H. Y. Al-Noori, and J. Suad, "YOLOv8-CAB: Improved YOLOv8 for Real-time object detection," *Karbala International Journal of Modern Science*, vol. 10, no. 1, Jan. 2024, doi: 10.33640/2405-609X.3339.
- [47] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Deep Learning-Based Image Segmentation on Multimodal Medical Imaging," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 162-169, Mar. 2019, doi: 10.1109/TRPMS.2018.2890359.
- [48] E. Warner *et al.*, "Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects," *International Journal of Computer Vision*, vol. 132, no. 9, pp. 3753-3769, Sep. 2024, doi: 10.1007/s11263-024-02032-8.




BIOGRAPHIES OF AUTHORS

Nia Saurina    received the Bachelor degree in Electronic Engineering Polytechnic Institut Surabaya (EEPIS) in 2006, and the Master Degree in Informatics Engineering from Institut Teknologi Sepuluh Nopember Surabaya (ITS) in 2009. She starts works in Universitas Wijaya Kusuma Surabaya as a lecturer from 2010 in Informatics Departments. Her research interest are software engineering and deep learning modelling. She can be contacted at email: niasaurina@gmail.com.






Nur Chamidah    is a Professor at the Mathematics Department, Faculty of Science and Technology, Airlangga University, Indonesia. She received Bachelor (S.Si.) degree in Mathematics from Airlangga University, Surabaya, Indonesia in 1997, Masters (M.Si.) degree in Statistics from Sepuluh Nopember Institute of Technology, Surabaya, Indonesia in 2002, and Doctorate (Dr.) degree in Statistics from Sepuluh Nopember Institute of Technology, Surabaya, Indonesia in 2014. She is a head of Statistical Modeling in Life Science (SMiLeS) Research Group. Her research interests are statistical modelling in growth chart and nutritional status of toddlers, non-communicable and infectious diseases, detection and classification diseases based on images especially using nonparametric and semiparametric regressions models. She can be contacted at email: nur-c@fst.unair.ac.id.



Riries Rulaningtyas    is a senior lecturer in Biomedical Engineering Study Program, Department of Physics, Faculty of Science and Technology, Universitas Airlangga, Indonesia. She got her Bachelor and Master from Institut Teknologi Sepuluh Nopember, Surabaya. She received her Ph.D. degree from School of Electrical Engineering and Informatics majoring Biomedical Engineering, Institut Teknologi Bandung. Her research interests are medical signal processing, medical image processing, and artificial intelligence. She can be contacted at email: riries-r@fst.unair.ac.id.



Aryati Aryati    is a Professor at the Faculty of Medicine, Universitas Airlangga, Surabaya. She was born in Surabaya in August 1963. She completed his Bachelor Degree of Medicine and Medical Professional Education at Universitas Airlangga in 1988, completed her Masters Degree in Immunology at Universitas Airlangga in 1992, completed her specialist in Clinical Pathology at Universitas Airlangga in 2000, completed her Doctoral Degree in Medical Sciences at Universitas Airlangga in 2006, and became an infectious disease consultant in 2007, also became Immunology consultant in 2022. Currently she works as a lecturer and doctor in the Department of Clinical Pathology, Faculty of Medicine, Universitas Airlangga/Dr. Soetomo Surabaya's Hospital. Her current position is the head of the Clinical Pathology Subspecialist Study Program, Faculty of Medicine, Universitas Airlangga and Chairman of Indonesian Association of Clinical Pathologist and Laboratory Medicine (IACPALM). She is also active as a member of the medical expert team for the COVID-19 task force and a member of the steering team for the advocacy team for the implementation of the PB IDI vaccine. Recently, she contributed as an author to the book "guidelines for the prevention and control of corona virus disease (COVID-19) Ministry of Health Revision 4 and Revision 5". Apart from that, she was also a contributor to the book "Standard Guidelines for Doctor Protection in the COVID-19 Pandemic Era". Her research interests are in the field of laboratory infectious diseases and immunology and molecular. She can be contacted at email: dr_aryati@yahoo.com.