Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 4, August 2026, pp. 3784~3796
ISSN: 2302-9285, DOI: 10.11591/eei.v15i4.10680

Driver drowsiness detection using YOLO based deep learning models

Helmi Wibowo[1], Muh Irhas Rafiqi[2]
[1]Department of Automotive Technology, Politeknik Keselamatan Transportasi Jalan, Tegal, Indonesia
[2]Department of Automotive Engineering Technology, Politeknik Keselamatan Transportasi Jalan, Tegal, Indonesia

Article Info

Article history:

Received May 14, 2025
Revised Apr 14, 2026
Accepted Jul 23, 2026

Keywords:

Computer vision
Deep learning
Driver drowsiness detection
Road safety
You only look once

ABSTRACT

The incidence of traffic accidents in Indonesia has been escalating, predominantly attributed to human factors such as fatigue and drowsiness. This study presents the implementation of a deep learning-based drowsiness detection system utilizing the you only look once (YOLO) architecture to enhance vehicular safety. Three YOLO model variants (YOLOv5, YOLOv8, and YOLOv10) were evaluated using a dataset comprising 1,000 annotated images across four classes: alert, low vigilance, drowsy, and microsleep. A quantitative experimental methodology was employed, with performance assessed through precision, recall, accuracy, and F1-score metrics. Experimental results demonstrate that YOLOv8 (medium and small variants) achieved superior overall performance, exhibiting a balanced optimization across all evaluation metrics. YOLOv5 yielded the highest recall, suggesting its suitability for comprehensive detection tasks, whereas YOLOv10 demonstrated enhanced computational efficiency without significant performance degradation. Based on these findings, YOLOv8 is recommended as the most effective model for real-world deployment, while YOLOv5 and YOLOv10 offer viable alternatives depending on specific operational requirements. This study contributes to the advancement of early warning systems for driver drowsiness detection, with the broader aim of mitigating traffic accident risks.

Corresponding Author:

Helmi Wibowo
Department of Automotive Technology, Politeknik Keselamatan Transportasi Jalan
Tegal, Central Java, Indonesia
Email: helmi.wibowo@pktj.ac.id

1. INTRODUCTION

Indonesia continues to face significant challenges in the transportation sector, particularly the persistently high rate of traffic accidents. According to the Ministry of Transportation, the number of accidents increased from 103,645 in 2021 to 137,851 in 2022, marking a sharp rise of 34,206 cases [1]. Although modern vehicles are increasingly equipped with active and passive safety technologies, these advancements alone have proven insufficient to guarantee road safety, which remains a primary concern for both drivers and other road users [2].

The causes of traffic accidents are generally attributed to human, vehicle, road, and environmental factors, with human-related errors being the most dominant [3], [4]. In particular, the National Transportation Safety Committee (KNKT) reported that nearly 80% of toll road accidents in Indonesia were caused by driver fatigue or drowsiness resulting from extended driving durations [5]. Driving in a drowsy state reduces concentration, slows reaction time, and increases the likelihood of losing vehicle control. Early signs, such as yawning, frequent blinking, and difficulty maintaining head posture, are often overlooked by drivers, even

though they represent early indicators of potential accidents [6]. Similar trends have also been reported globally, where drowsiness has been identified as a leading cause of severe road crashes, accounting for up to 20% of fatal accidents in several developed countries according to the World Health Organization (WHO) and the National Highway Traffic Safety Administration (NHTSA). This highlights that drowsiness detection is not only a local concern but also a global safety priority.

Several notable accidents worldwide illustrate the urgency of this issue. For example, on July 6, 2025, a 25-year-old driver experienced microsleep while traveling on the Bernina Pass road in Poschiavo, Switzerland. At approximately 2:50 p.m., the vehicle crossed into the oncoming lane and collided head-on with another car, followed by a secondary collision involving a third vehicle. Three individuals sustained injuries and required hospitalization, underscoring the critical risks associated with driver drowsiness [7]. Similarly, on December 2, 2025, a coroner determined that the fatal collision involving a night-shift worker and a school bus was primarily attributable to driver fatigue. The investigation emphasized that microsleep episodes lasting only a few seconds can severely compromise driving performance, reinforcing fatigue as a critical risk factor in road traffic accidents [8]. Such cases reinforce the urgent need for robust, real-time drowsiness detection systems to mitigate accident risks.

Existing studies have investigated drowsiness detection primarily through visual monitoring of the driver's face, particularly by analyzing blink frequency, eye closure duration, and head movement patterns [9]. Deep learning–based methods have achieved promising results. For example, Flores *et al.* [9] proposed a model using Mediapipe for feature extraction and several convolutional neural networks (CNN) architectures, with ResNet50V2 achieving the highest accuracy of 99.89%. Jahan *et al.* [10] trained VGG16, VGG19, and 4D CNN models on the MRL eye dataset, where the 4D CNN achieved superior accuracy, precision, and recall (97%) despite longer training times. Kanigoro and Asdyo [11] further demonstrated that applying data augmentation to YOLOv5s improved detection performance by 23% compared to non-augmented training. While these studies confirm the potential of deep learning, challenges remain in ensuring computational efficiency, scalability, and real-time performance under real driving conditions.

With the rapid advancement of artificial intelligence, CNNs, and the you only look once (YOLO) framework have emerged as powerful tools for object detection and pattern recognition [12]–[14]. Nevertheless, a critical research gap remains: most prior works focus on controlled datasets and offline evaluations, with limited emphasis on real-time implementation in naturalistic driving scenarios. Additionally, few studies have integrated CNN-based feature extraction with YOLO-based object detection for multimodal analysis of facial indicators such as eyes, mouth, and head movement simultaneously.

To address these gaps, this study proposes a hybrid deep learning approach that combines CNN and YOLO to detect early signs of driver drowsiness in real time. By focusing on critical facial indicators, the system is expected to generate timely alerts that enable drivers to take preventive action, thereby reducing accident risks. This research contributes not only to improving road safety in Indonesia but also to advancing the development of intelligent transportation safety systems that are efficient, adaptive, and globally relevant.

2. METHOD

This study aims to evaluate the effectiveness of implementing YOLOv5, YOLOv8, and YOLOv10 algorithms for driver drowsiness detection. According to Hussain [15], these YOLO variants demonstrate an optimal balance of speed, accuracy, and efficiency, making them suitable for resource-constrained environments. Based on this objective, the research adopts a quantitative experimental approach, as the data processed are numerical and directly support accuracy analysis. The study was conducted at PT. Bengawan Solo Trans from August 2024 to February 2025. Data collection was carried out using two main methods: observation and documentation. The observation method involved direct monitoring of research objects in real-world contexts, particularly the behavior of Batik Solo Trans bus drivers related to drowsiness [16]. The documentation method was employed to collect secondary data in the form of written records, images, and visual documents obtained from individuals or institutions. For this purpose, data were recorded using digital cameras and smartphones within the driver's cabin, focusing on four states: alert, low vigilance, drowsy, and microsleep. Supporting tools included writing instruments, a laptop, and a digital camera used to document activities and conditions during the study. Figure 1 presents the research flowchart, which systematically illustrates the stages and procedures undertaken throughout the research process.

2.1. Problem identification

In this stage, the study focuses on identifying potential issues and research gaps relevant to the topic under investigation. Problem identification serves as the initial step to introduce the challenges to be addressed and to define the research objectives based on these challenges [17]. In the context of transportation in Indonesia, a major concern is the high number of traffic accidents caused by drowsy driving. This issue was selected because limited research in Indonesia has explored the implementation of YOLO

algorithms for real-time drowsiness detection. Therefore, this study seeks to address this knowledge gap and contribute to the advancement of technology-based transportation safety.

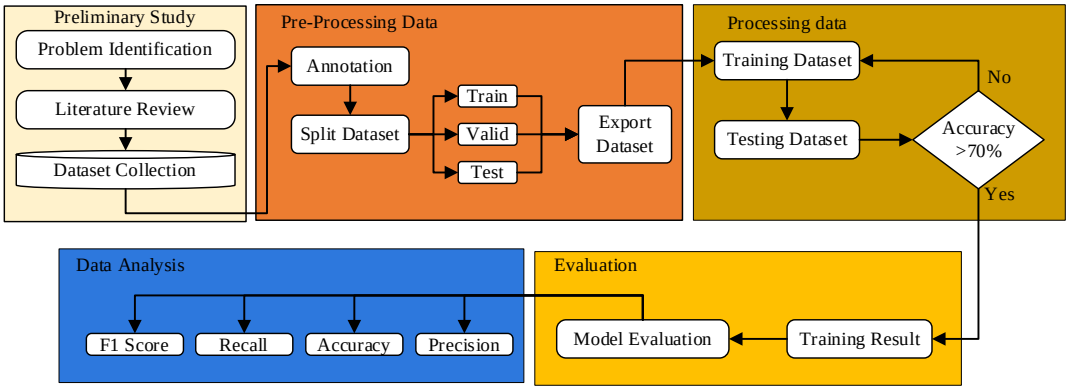


Figure 1. Research flowchart

2.2. Literature review

The literature review is the process of collecting data and information that serves as a reference for the research, relying on relevant sources such as journals, articles, proceedings, and books [18]. In the context of driver drowsiness detection, the reviewed literature includes studies on image-based drowsiness detection methods, eye and head movement recognition, as well as the implementation of YOLO in object detection tasks. This stage provides the foundation for designing the experimental procedure and justifying the methodological choices.

2.3. Dataset collection

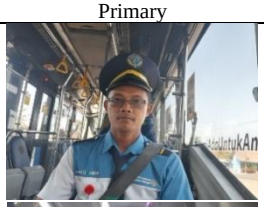
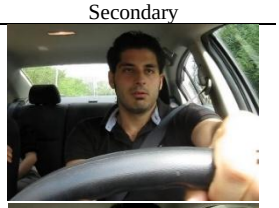
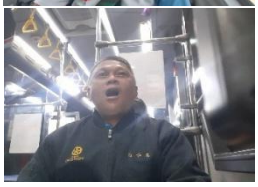
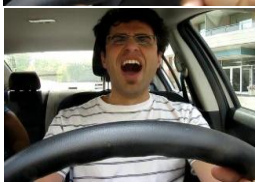
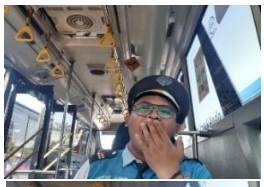
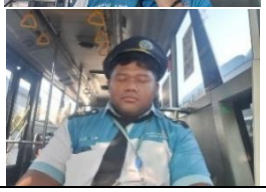
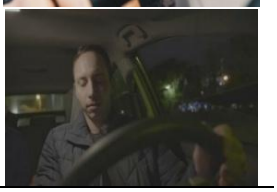
This study utilized primary data collected through direct image acquisition of 40 drivers' faces from Corridor 1 of an Indonesian transportation company, captured during driving using a Samsung Galaxy A30 smartphone camera. Complementarily, secondary data were obtained from open-source platforms such as Kaggle, consisting of four classes: alert, low vigilance, drowsy, and microsleep, which were initially unlabeled. The integration of primary and secondary datasets was designed not only to increase data diversity but also to address the limitations of relying on a single-source dataset, such as overfitting and limited generalization. By combining real-world observations with publicly available datasets, this study aims to develop a more robust and adaptable model for driver drowsiness detection under various conditions. To ensure a balanced dataset for model training and evaluation, an equal number of images was selected for each driver condition from both primary and secondary data sources. The distribution of the dataset according to driver condition and data source is summarized in Table 1.

Table 1. Distribution of the primary and secondary datasets

Driver condition	Primary data	Secondary data
Alert	125	125
Low vigilance	125	125
Drowsy	125	125
Microsleep	125	125
Total	500	500

Data collection was conducted in a real-world environment, specifically inside the cabin of Batik Solo Trans buses. The acquisition process was carried out across two times ranges to capture variations in natural lighting conditions: shift 1 (04:00–12:00) and shift 2 (12:00–20:00). Furthermore, each driver was documented from three different angles right oblique, left oblique, and frontal to provide a more representative perspective of facial conditions in realistic settings. Although the dataset incorporates variability in viewpoints and lighting conditions across two times ranges, demographic factors such as driver age and gender were not considered. This limitation highlights the need for future research to involve more diverse participants and additional environmental conditions, including nighttime lighting or varying road situations. Further details regarding the dataset employed for model training are provided in Table 2.

Table 2. Training data

No	Class	Dataset	
		Primary	Secondary
1.	Alert		
2.	Low vigilance		
3.	Drowsy		
4.	Microsleep		

2.4. Preprocessing

This stage involved a preprocessing procedure in which the dataset was labeled to classify facial images into four categories: alert, low vigilance, drowsy, and microsleep [19]. The labeling process was performed using the Roboflow platform, which transforms raw data into a test-ready format and enables export in various compatible structures. In this study, all datasets were standardized to a resolution of 640×640 pixels, following the YOLO input specification to ensure consistency and computational efficiency during the training process.

The annotation process refers to the stage of assigning class labels and spatial information to each image by defining the coordinates and confidence values of the bounding boxes. Roboflow provides several annotation tools, including the drag tool, bounding box tool, polygon tool, and AI-assisted annotation features to facilitate the labeling process. In this study, the bounding box tool was selected because it is well suited for facial object detection. Figure 2 illustrates the annotation interface in Roboflow, where each driver's face is enclosed using a bounding box and assigned the corresponding driver condition label. The resulting annotations provide the object location and class information required for training the proposed YOLO-based driver drowsiness detection model. As shown in Figure 2, this annotation process ensures consistent labeling before the dataset is divided into the training, validation, and testing sets.

Table 3 demonstrates that each version of the dataset was organized in a balanced manner to ensure a fair comparison of model performance. For example, in the first dataset consisting of 200 images, each class was represented by 50 images, with an equal proportion of primary data (25) and secondary data (25). In this study, no data augmentation techniques were applied in order to preserve the authenticity of the dataset distribution, ensuring that the model performance truly reflected its ability to recognize original data without artificial variations. Following the annotation process, the next stage involved dataset splitting for training and testing purposes. Dataset splitting is essential to separate training and testing data, thereby enabling the development and evaluation of the model. In this study, the dataset was divided into training (72%), validation (20%), and testing (8%) subsets. However, due to this splitting strategy, it is possible that the same driver identity (e.g., a frontal face included in the training set and a side face included in the testing set) appeared in multiple subsets. This may allow the model to more easily recognize driver identities, as certain facial features had already been observed during training. Consequently, the reported performance is likely to reflect the model's ability to recognize variations in facial perspectives (frontal, right, and left) of the same drivers, rather than its generalization capability to unseen drivers. This limitation is important to acknowledge, and it highlights the need for future research to consider identity-based dataset splitting strategies to improve model generalization.

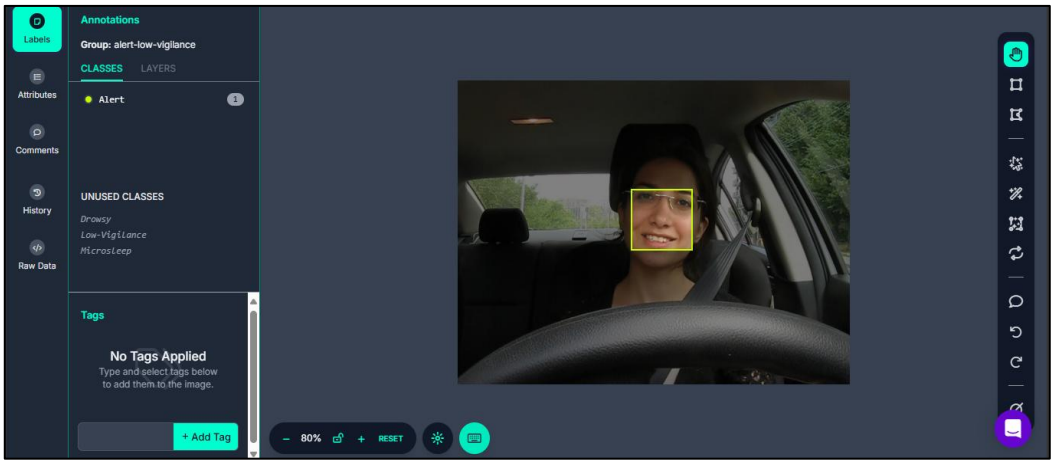


Figure 2. Example of image annotation using the Roboflow platform

Table 3. Number of annotations per class

Dataset size	Class types			
	Alert	Low vigilance	Drowsy	Microsleep
200 datasets	50	50	50	50
400 datasets	100	100	100	100
600 datasets	150	150	150	150
800 datasets	200	200	200	200
1000 datasets	250	250	250	250

2.5. Processing

In the processing stage, the dataset was divided into two subsets: training and testing. The training phase refers to the process of teaching the model to recognize driver drowsiness characteristics using the prepared dataset [20]. Several steps were involved in the training stage, including mounting Google Drive, checking GPU installation, configuring YOLO, creating a custom dataset, and initiating the custom training process. The training was conducted using Google Collaboratory with a T4 GPU and high RAM configuration to optimize performance and accelerate the training process [21]. After the annotation process was completed, the labeled dataset was exported from the Roboflow platform and imported into the Google Colaboratory environment for model training. Roboflow provides a download feature that automatically generates an API key and Python snippet, enabling seamless integration of the annotated dataset into the training notebook. Figure 3 illustrates the dataset export interface in Roboflow and its integration into the Google Colaboratory notebook used for training the YOLOv5, YOLOv8, and YOLOv10 models.

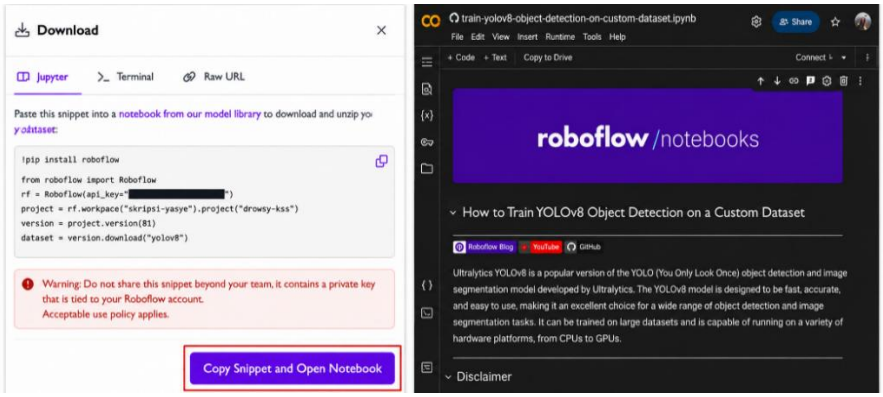


Figure 3. Dataset export from Roboflow and integration into the Google Colab notebook

The performance of deep learning models is strongly influenced by the selection of appropriate training hyperparameters. To ensure a fair and unbiased comparison among YOLOv5, YOLOv8, and

YOLOv10, the same training configuration was applied to all models. The selected hyperparameters include the number of training epochs, batch size, optimizer, and learning rate, as summarized in Table 4. These parameters were kept identical across all models to ensure that performance differences were attributable to the model architectures rather than variations in the training configuration.

Table 4. Training hyperparameters for YOLOv5, YOLOv8, and YOLOv10

Training parameter	YOLOv5	YOLOv8	YOLOv10	Training parameter	YOLOv5	YOLOv8	YOLOv10
Epoch	100	100	100	Optimizer	AdamW	AdamW	AdamW
Batch size	32	32	32	Learning rate	0.001	0.001	0.001

Using the hyperparameters summarized in Table 4, YOLOv5, YOLOv8, and YOLOv10 were trained under identical conditions to ensure a fair comparison of their performance. After the training process was completed, the resulting models were evaluated using the testing dataset. The testing dataset consisted of annotated images and videos that were used to assess the detection performance of each model. The experimental results are presented and discussed in Figure 4, which compares the performance of the evaluated YOLO models.

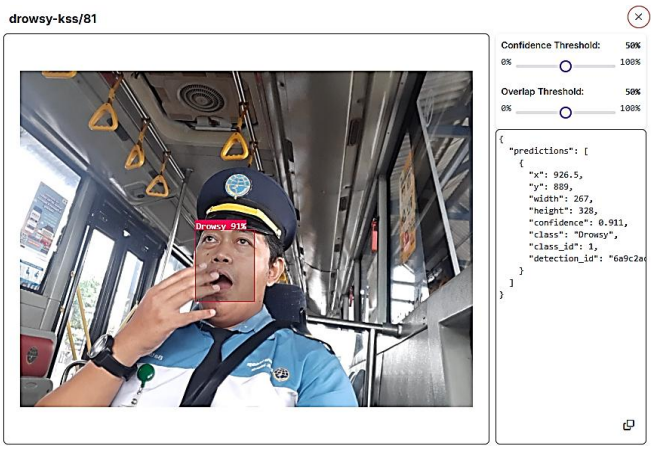


Figure 4. Dataset evaluation

2.6. Model evaluation

The model evaluation was conducted using the testing dataset to measure accuracy and confidence levels in classifying new data [22]. The evaluation metrics included the confusion matrix, accuracy, precision, recall, and F1-score, as illustrated in Figure 5. The output generated from the training process is a file named "best.pt", which represents the optimal result of the training phase. This file can be utilized with different datasets to evaluate whether the YOLO algorithm is capable of detecting driver drowsiness across various images and videos.

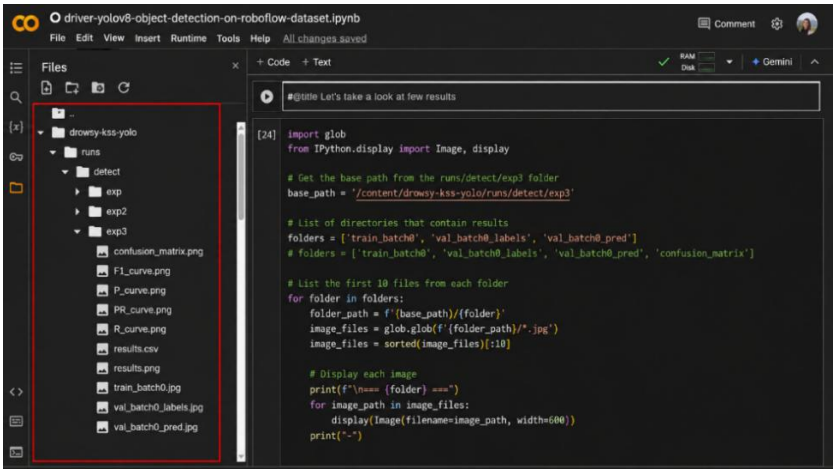


Figure 5. YOLO training results

2.7. Data analysis

Data analysis refers to the process of systematically organizing and structuring the collected information from interviews, field notes, and documentation. This process involves categorizing data, breaking it down into units, synthesizing, arranging it into patterns, identifying key elements for further examination, and drawing conclusions to facilitate interpretation. In this study, two analytical techniques were employed, namely comparative analysis and performance analysis.

2.7.1. Comparative analysis

To evaluate the influence of different YOLO architectures on driver drowsiness detection, this study compares several variants of YOLOv5, YOLOv8, and YOLOv10. Each architecture provides multiple model variants with different computational complexities and network sizes, allowing the selection of models that balance detection accuracy and computational efficiency. Table 5 summarizes the YOLO variants evaluated in this study, including the nano (n), small (s), medium (m), large (l), and extra-large (x) variants available for each architecture. As shown in Table 5, not all architectures provide the same set of model variants. Therefore, this study selected the available variants from each YOLO architecture to compare their performance under different dataset sizes. To investigate the effect of dataset size on model performance, five dataset scenarios with different numbers of images were prepared. The datasets consisted of 200, 400, 600, 800, and 1,000 images, allowing a systematic evaluation of how increasing the amount of training data influences the performance of the YOLO models. Table 6 summarizes the number of images included in each dataset scenario used throughout the experiments.

Table 5. Comparison of YOLO variants

Model variant	YOLOv5	YOLOv8	YOLOv10
Nano (n)	✓	✓	✓
Small (s)	✓	✓	✓
Medium (m)	✓	✓	✓
Large (l)	✓	✓	-
Extra large (x)	✓	-	-

Table 6. Comparison of dataset sizes used in the experiments

Dataset	Number of images
1	200
2	400
3	600
4	800
5	1000

2.7.2. Performance evaluation

The performance of an algorithm significantly influences the accuracy of detection outcomes. This study aims to evaluate the performance of the YOLO algorithm in detecting driver drowsiness. The evaluation metrics employed include the confusion matrix, accuracy, precision, recall, and F1-score, each of which serves a distinct purpose in assessing model performance [23].

a. Confusion matrix

The confusion matrix is a tabular representation used to evaluate the performance of a classification model by comparing the predicted class labels with the corresponding ground-truth labels [24]. It provides detailed information about the model's prediction outcomes by categorizing them into four possible results: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Figure 6 illustrates the structure of the confusion matrix used in this study. TP and TN represent correctly classified positive and negative instances, respectively, whereas FP and FN indicate incorrect predictions. This matrix serves as the basis for calculating the evaluation metrics, including accuracy, precision, recall, and F1-score.

b. Accuracy

Accuracy measures the proportion of correctly classified instances, encompassing both positive and negative predictions [25]. The value of accuracy can be calculated using (1):

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

c. Precision

Precision is defined as the ratio of true positive predictions to the total predicted positive outcomes, indicating the model's ability to minimize false positives [26]. The value of precision can be calculated using (2):

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

d. Recall

Recall is defined as the ratio of true positive predictions to the total actual positive instances, reflecting the model's sensitivity in detecting drowsiness cases [27]. The recall value can be calculated using (3):

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

e. F1-score

The F1-score represents the weighted harmonic mean of precision and recall, providing a balanced measure between the two metrics [28]. The F1-score value can be calculated using (4):

$$F1\ Score = \frac{2\ x\ precision\ x\ recall}{precision+recall} \quad (4)$$

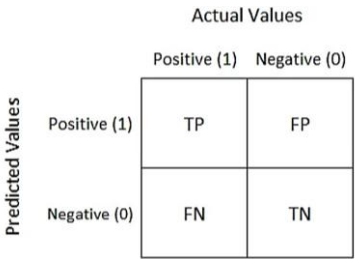


Figure 6. Confusion matrix

3. RESULTS AND DISCUSSION

This study evaluates the performance of several deep learning models, namely YOLOv5, YOLOv8, and YOLOv10, using evaluation metrics such as precision, accuracy, recall, and F1-score. The comparative results of model performance are presented in the graph below, providing a clearer illustration of each model's capability in object detection. A detailed explanation of the training outcomes is provided in the subsequent discussion.

3.1. Performance evaluation on Dataset 1

The first dataset consists of 200 images, which are divided into 144 training images, 40 validation images, and 16 testing images. Based on the evaluation results presented in Figure 7, the YOLOv8-small model achieved the highest precision value of 0.934, reflecting its ability to minimize false positive detections.

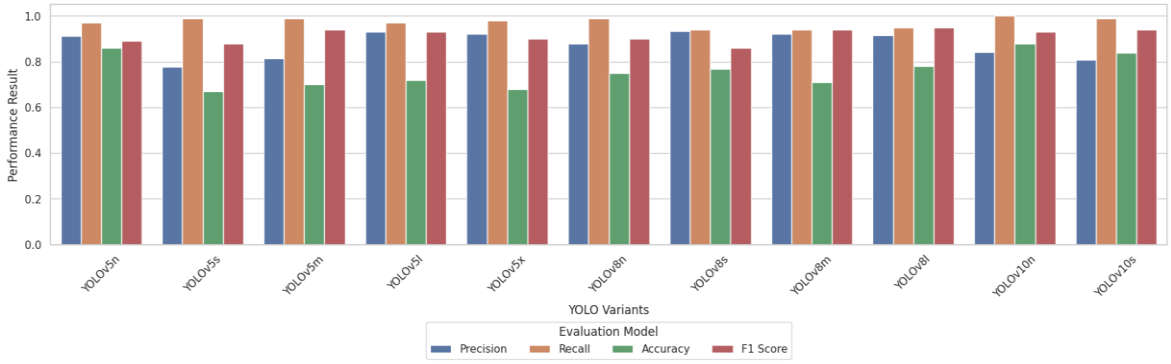


Figure 7. Performance of the YOLO algorithm on Dataset 1

The YOLOv5-medium and YOLOv5-large models demonstrated the most balanced performance, as indicated by their high F1-scores (0.94 and 0.93, respectively), along with relatively stable accuracy and recall values. In contrast, the YOLOv10-nano model achieved a perfect recall of 1.00, indicating its capability to detect all objects in the dataset, although its precision was lower (0.842). Meanwhile, the YOLOv10-small model exhibited consistently stable performance across all evaluation metrics, with an F1-score of 0.94, highlighting its efficiency and effectiveness as a lightweight model. Overall, YOLOv5 demonstrated superiority in balancing evaluation metrics, YOLOv8 excelled in precision, and YOLOv10 showed strength in recall as well as computational efficiency.

3.2. Performance evaluation on Dataset 2

The second dataset consisted of 400 images, comprising 288 images for training, 80 images for validation, and 32 images for testing. Figure 8 presents a comparative performance evaluation of the YOLOv5, YOLOv8, and YOLOv10 variants on Dataset 2 using four evaluation metrics: precision, recall, accuracy, and F1-score. The figure shows that YOLOv8-large achieved the highest precision (0.953), indicating its strong capability to minimize false-positive predictions. In contrast, YOLOv5x obtained the highest F1-score (0.98) with a perfect recall of 1.00, although its precision was relatively lower (0.839), reflecting robust detection performance with room for improvement in accuracy. All YOLOv5 variants exhibited superior recall (1.00), while YOLOv8 maintained consistent performance across different model sizes, with both the small and large variants achieving high F1-scores of 0.89 and 0.86, respectively. Meanwhile, YOLOv10 demonstrated stable performance, characterized by high recall (0.98–0.99) and reasonably good F1-scores (0.84–0.86), making it suitable for deployment in resource-constrained environments. Overall, Figure 8 indicates that YOLOv5 achieved the best recall and F1-score, YOLOv8 provided the highest precision, and YOLOv10 offered balanced performance across all evaluation metrics.

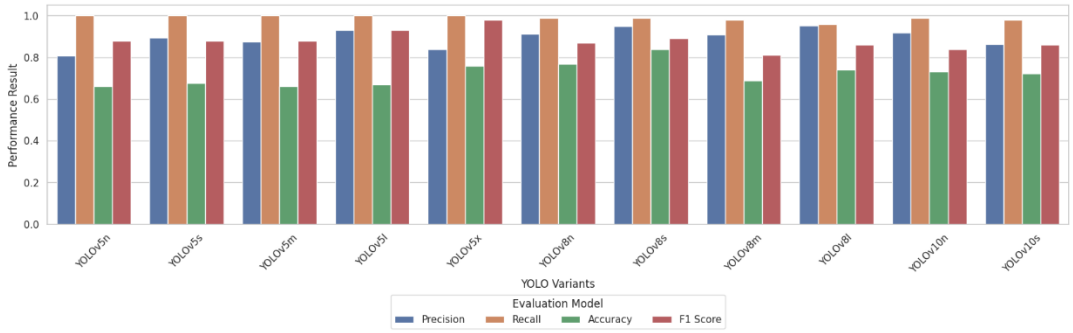


Figure 8. Comparison of precision, recall, accuracy, and F1-score of the evaluated YOLO variants on Dataset 2

3.3. Performance evaluation on Dataset 3

The third dataset consisted of 600 images, comprising 432 images for training, 120 images for validation, and 48 images for testing. Figure 9 presents a comparative performance evaluation of the YOLOv5, YOLOv8, and YOLOv10 variants on Dataset 3 using four evaluation metrics: precision, recall, accuracy, and F1-score.

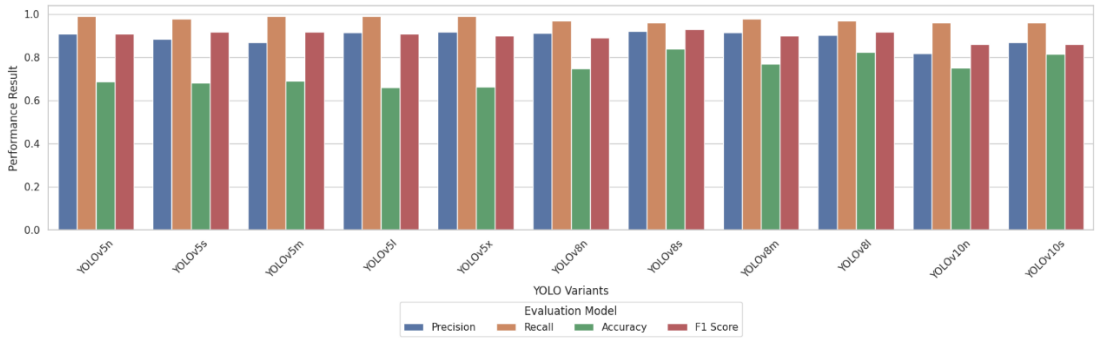


Figure 9. Comparison of precision, recall, accuracy, and F1-score of the evaluated YOLO variants on Dataset 3

Figure 9 shows that YOLOv8-small achieved the best overall performance, with a precision of 0.921, an accuracy of 0.84, and the highest F1-score of 0.93, indicating a strong balance between precision and recall. YOLOv5 also demonstrated competitive performance, particularly in terms of recall, which consistently reached 0.99, although its accuracy was slightly lower than that of YOLOv8. In contrast, YOLOv10 produced relatively lower precision and accuracy values, despite maintaining an acceptable recall of 0.96. Overall, Figure 9 indicates that YOLOv8-small provides the most balanced performance on Dataset 3, while YOLOv5 remains suitable for applications requiring high sensitivity and YOLOv10 offers stable but comparatively lower performance across the evaluated metrics.

3.4. Performance evaluation on Dataset 4

The fourth dataset consisted of 800 images, including 576 images for training, 160 images for validation, and 64 images for testing. Figure 10 presents a comparative performance evaluation of the YOLOv5, YOLOv8, and YOLOv10 variants on Dataset 4 using four evaluation metrics: precision, recall, accuracy, and F1-score. The figure shows that YOLOv8, particularly the medium variant, achieved the most balanced performance, with the highest F1-score of 0.95, a precision of 0.901, and a recall of 0.97, indicating a strong balance between detection accuracy and sensitivity. The small and large variants of YOLOv8 also demonstrated consistently strong performance across all evaluation metrics. Meanwhile, YOLOv5 exhibited the highest recall, with values ranging from 0.98 to 0.99 across all variants, highlighting its capability to detect nearly all positive instances. Although YOLOv10 produced slightly lower precision and F1-score values than YOLOv8, it maintained stable performance with high recall (0.98–0.99), making it a suitable alternative for computationally efficient applications. Overall, Figure 10 demonstrates that YOLOv8 provides the most balanced detection performance on Dataset 4, while YOLOv5 excels in recall and YOLOv10 offers competitive performance with a more efficient architecture.

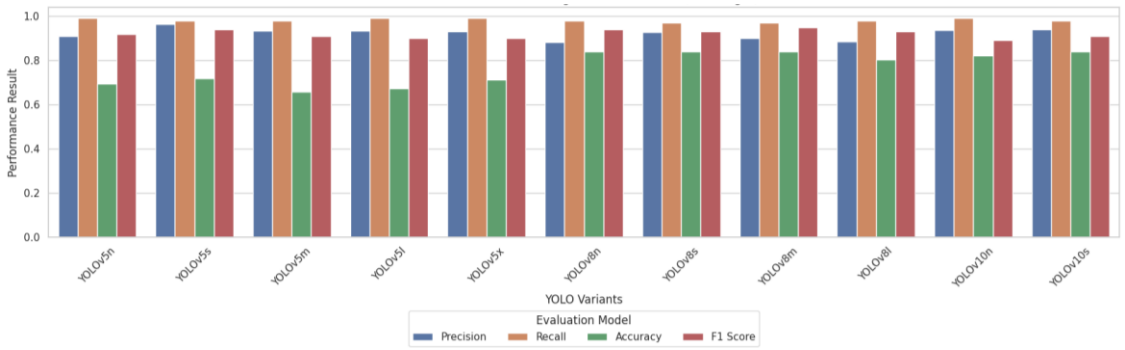


Figure 10. Comparison of precision, recall, accuracy, and F1-score of the evaluated YOLO variants on Dataset 4

3.5. Performance evaluation on Dataset 5

The fifth dataset consisted of 1,000 images, comprising 720 images for training, 200 images for validation, and 80 images for testing. Figure 11 presents a comparative performance evaluation of the YOLOv5, YOLOv8, and YOLOv10 variants on Dataset 5 using four evaluation metrics: precision, recall, accuracy, and F1-score. The figure shows that YOLOv5 achieved consistently high performance, with precision values reaching 0.959, recall of 0.99, and stable F1-scores ranging from 0.86 to 0.89, indicating reliable and comprehensive detection capability. Among all evaluated models, the YOLOv8 medium variant delivered the best overall performance, achieving a precision of 0.95, recall of 0.99, and the highest F1-score of 0.91, demonstrating an effective balance between detection accuracy and sensitivity. Meanwhile, YOLOv10, particularly the small variant, also produced competitive results, with a precision of 0.954, recall of 0.97, and an F1-score of 0.87, making it a suitable choice for applications with limited computational resources. Overall, Figure 11 demonstrates that YOLOv8 provides the most balanced performance on Dataset 5, while YOLOv5 offers excellent detection stability and YOLOv10 achieves competitive results with a more computationally efficient architecture.

3.6. Comprehensive analysis of you only look once performance

Overall, the evaluation results indicate that each YOLO variant possesses specific strengths and limitations. YOLOv5 stands out in terms of training efficiency and recall stability (≥0.97), demonstrating its ability to detect most instances with minimal misses, which makes it a promising candidate for real-time applications on resource-constrained devices. However, the relatively low accuracy on the Drowsy class (0.52 on Dataset 2) suggests that network depth and limited data augmentation remain challenges in capturing complex patterns, consistent with [29], which reported the limitations of YOLOv5 on minority classes. YOLOv8 exhibits superior precision and consistency, with the highest precision of 0.953 achieved by YOLOv8l on Dataset 2. This aligns with findings in [30], highlighting the contribution of the CSPDarknet and PANet architectures in strengthening feature representation. Among its variants, YOLOv8s demonstrates the most balanced performance (precision 0.928, accuracy 0.839, recall 0.97, F1-score 0.93 on Dataset 4), making it suitable for scenarios requiring a trade-off between accuracy and efficiency. Nevertheless, its longer training time and sensitivity to illumination changes may limit deployment in real-world environments with highly dynamic visual conditions.

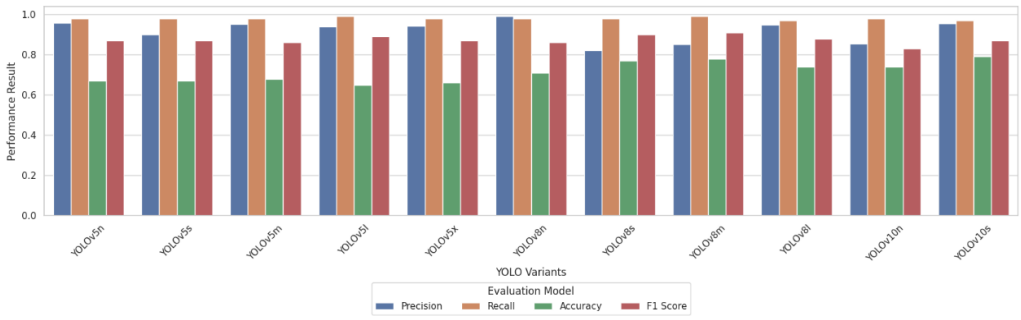


Figure 11. Performance of the YOLO algorithm on Dataset 5

YOLOv10, on the other hand, presents a promising compromise between efficiency and precision. With faster training times (e.g., 26 minutes for YOLOv10n on Dataset 5), the model achieves competitive detection performance, particularly YOLOv10s on Dataset 4, with a precision of 0.94, accuracy of 0.84, recall of 0.98, and an F1-score of 0.91. In addition to its detection accuracy, YOLOv10s demonstrates efficient inference performance, achieving an average inference time of 83.77 ms per image, corresponding to approximately 11.93 FPS on an NVIDIA Tesla T4 GPU. These findings support [31], which emphasized the efficiency and convergence advantages of YOLOv10 compared to previous versions. From an implementation perspective, the recall stability of YOLOv5 and the efficiency of YOLOv10 suggest potential for real-time driver monitoring systems, whereas the higher precision of YOLOv8 is more suitable for forensic analysis or applications requiring high accuracy [32]. Despite these findings, this study has several limitations. Specifically, no statistical significance testing was conducted to compare models, error analysis was not performed to identify failure patterns, and concrete measurements of inference time and memory consumption were not included. These limitations render efficiency claims, particularly for YOLOv10, relative rather than based on operational data. Future research is recommended to focus on: i) applying advanced data augmentation techniques to address class imbalance, particularly for minority classes such as Drowsy; ii) evaluating inference time and memory consumption to better reflect real-world implementation scenarios; iii) conducting systematic error analysis to identify model failure patterns; and iv) exploring lightweight architectures that are more adaptive to varying illumination conditions.

4. CONCLUSION

In summary, the evaluation across five datasets highlights that each YOLO variant has distinct strengths and limitations. YOLOv5 excels in training efficiency and recall stability, making it suitable for resource-constrained real-time applications, while YOLOv8 achieves the highest precision and overall consistency, particularly in scenarios requiring high accuracy. YOLOv10 offers a balanced trade-off between efficiency and precision, with faster convergence and strong performance in smaller variants. Nevertheless, challenges remain in detecting minority classes such as Drowsy, largely due to data imbalance and sensitivity to hyperparameter settings. Future work should address these limitations by incorporating advanced data augmentation, systematic error analysis, inference and resource evaluation, and the development of lightweight architectures adaptive to diverse environmental conditions.

ACKNOWLEDGMENTS

The authors would like to thank Politeknik Keselamatan Transportasi Jalan for providing research facilities and technical support. The authors also acknowledge PT. Bengawan Solo Trans, a public transportation company, for granting access to the research site. All acknowledged parties have provided their consent to be mentioned in this section.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Helmi Wibowo	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓
Muh Irhas Rafiqi		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

Written informed consent was obtained from all participants before data collection. All participants were informed about the purpose of the study, the use of facial image data, and their rights regarding privacy and data protection.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request. Due to the inclusion of facial image data, the dataset is not publicly available to protect participant privacy.

REFERENCES

[1] S. A. Hasibuan, N. F. Agus, and K. Hakim, "Effectivenessele electronic traffic law enforcement (ETLE) in the effort to regulate traffic directorate cross-regional police affairs within a region Bengkulu city," *International Journal of Policy and Public Administration*, vol. 5, no. 2, pp. 124–132, 2024.

[2] M. A. Puspasari *et al.*, "Effect of distraction and driving behaviour to traffic accidents in jakarta using partial least squares structural equation modeling (PLS-SEM)," *International Journal of Technology*, vol. 14, no. 7, pp. 1548–1559, Dec. 2023, doi: 10.14716/ijtech.v14i7.6676.

[3] N. Acharya and M. Henry, "Exploring the role of human behavior in road crash occurrence through analysis of crash variability in highly similar road environments," in *International Conference on Transportation and Development 2024: Transportation Safety and Emerging Technologies - Selected Papers from the International Conference on Transportation and Development 2024*, Tokyo: ASCE Library, 2024, pp. 249–260, doi: 10.1061/9780784485514.022.

[4] A. Bener, E. Yildirim, T. Özkan, and T. Lajunen, "Driver sleepiness, fatigue, careless behavior and risk of motor vehicle crash and injury: population based case and control study," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 4, no. 5, pp. 496–502, 2017, doi: 10.1016/j.jtte.2017.07.005.

[5] K. Khotimah and A. Sjafruddin, "Analysis of driver fatigue caused by highway hypnosis in monotonous geometrics of road: state of the arth review," in *International Conference on Civil, Structural and Transportation Engineering*, Jun. 2024, pp. 1–8, doi: 10.11159/iccste24.175.

[6] H. Pan *et al.*, "Exploring the effect of driver drowsiness on takeover performance during automated driving: an updated literature review," *Accident Analysis & Prevention*, vol. 216, p. 108023, Jun. 2025, doi: 10.1016/j.aap.2025.108023.

[7] Bluewin, "25 year old driver falls into oncoming lane due to microsleep - three injured," bluewin.ch, 2026. [Online]. Available: https://www.bluewin.ch/en/news/switzerland/25-year-old-driver-falls-into-oncoming-lane-due-to-microsleep-three-injured-2771276.html, (Accessed: Feb. 24).

[8] F. Blackweel, "Night shift worker died in school bus crash driving home, coroner warns of fatigue," nzherald.co.nz, 2025, [Online]. Available: https://www.nzherald.co.nz/nz/night-shift-worker-died-in-school-bus-crash-driving-home-coroner-warns-of-fatigue/XRJCTOHHI5F5XOMK7D45M5QTBM/. (Accessed: Feb. 24).

[9] R. Flores, F. Palomino-Quispe, R. J. Coaquira-Castillo, J. C. Herrera-Levano, T. Paixão, and A. B. Alvarez, "A CNN-based approach for driver drowsiness detection by real-time eye state identification," Applied Sciences, vol. 13, no. 13, 7849, 2023.

[10] I. Jahan *et al.*, "4D: a real-time driver drowsiness detector using deep learning," *Electronics (Switzerland)*, vol. 12, no. 1, p. 235, Jan. 2023, doi: 10.3390/electronics12010235.

[11] B. Kanigoro and B. Asdyo, "Facial landmark and YOLOv5 drowsiness detection system," *Procedia Computer Science*, vol. 245, pp. 548–554, 2024, doi: 10.1016/j.procs.2024.10.281.

[12] L. Shang *et al.*, "Research on fatigue detection of flight trainees based on face emf feature model combination with PSO-CNN algorithm," *Scientific Reports*, vol. 14, no. 1, pp. 1–11, 2024, doi: 10.1038/s41598-024-71192-x.

[13] S. A. El-Nabi, W. El-Shafai, E.-S. M. El-Rabaie, K. F. Ramadan, F. E. Abd El-Samie, and S. Mohsen, "Machine learning and deep learning techniques for driver fatigue and drowsiness detection: A review," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 9441–9477, 2024, doi: 10.1007/s11042-023-15054-0.

[14] A. Sinha, R. P. Aneesh, and S. K. Gopal, "Drowsiness detection system using deep learning," in *2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII)*, Chennai, India: IEEE, Mar. 2021, pp. 1–6, doi: 10.1109/ICBSII51839.2021.9445132.

[15] M. Hussain, "YOLOv5, YOLOv8 and YOLOv10: the go-to detectors for real-time vision," *aXiv Preprint*, vol. 4, pp. 1–12, 2024,

doi: 10.48550/arXiv.2407.02988.
[16] M. Darmanin, R. K. Matthew, M. Inglott, and H. Schembri, "The use of an observation proforma during a school-based physical activity programme: exploring the researchers' insights," *Journal of Education and Practice*, vol. 14, no. 32, pp. 1–9, Nov. 2023, doi: 10.7176/JEP/14-32-01.
[17] S. E. Atthar, "International conference on emerging markets (ICEM) 2026 ai-enabled digital twins for low-carbon logistics in emerging markets: a human-centric framework for cold-chain energy efficiency and cbam-ready supply chains in India "A human-centric framework for," *International Journal of Research and Innovation in Social Science (IJRISS)*, vol. 10, no. 19, pp. 41–57, 2026, doi: 10.47772/IJRISS.
[18] I. K. Bakti, Zulkarnain, A. Yarun, Rusdi, M. Syaifudin, and H. Syafaq, "The role of artificial intelligence in education: a systematic literature review," *Jurnal Iqra': Kajian Ilmu Pendidikan*, vol. 8, no. 2, pp. 182–197, Nov. 2023, doi: 10.25217/ji.v8i2.3194.
[19] E. Essel, F. Lacy, F. Albalooshi, W. Elmedany, and Y. Ismail, "Drowsiness detection in drivers using facial feature analysis," *Applied Sciences*, vol. 15, no. 1, p. 20, Dec. 2024, doi: 10.3390/app15010020.
[20] G. S. Krishna, K. Supriya, J. Vardhan, and M. R. K, "Vision transformers and yolov5 based driver drowsiness detection framework," *aXiv Preprint*, 2022, doi: 10.48550/arXiv.2209.01401.
[21] C. G. Ajudiya and S. S. Panchal, "A review on ai-driven drowsiness detection systems using deep-learning," in *EPJ Web of Conferences*, Jan. 2026, p. 04001, doi: 10.1051/epjconf/202634804001.
[22] S. Cao, P. Feng, W. Kang, Z. Chen, and B. Wang, "Optimized driver fatigue detection method using multimodal neural networks," *Scientific Reports*, vol. 15, no. 1, p. 12240, Apr. 2025, doi: 10.1038/s41598-025-86709-1.
[23] M. Trigka and E. Dritsas, "A comprehensive survey of machine learning techniques and models for object detection," *Sensors*, vol. 25, no. 1, p. 214, Jan. 2025, doi: 10.3390/s25010214.
[24] D. Widyawati and A. Faradibah, "Comparison analysis of classification model performance in lung cancer prediction using decision tree, naive bayes, and support vector machine," *Indonesian Journal of Data and Science*, vol. 4, no. 2, pp. 80–89, 2023, doi: 10.56705/ijodas.v4i2.76.
[25] A. Bhamidipati, "Mathematical ranking of ai classifiers using confusion matrix and matthews correlation coefficient," *Journal of Student Research*, vol. 11, no. 3, pp. 1–13, Aug. 2022, doi: 10.47611/jsrhs.v11i3.3113.
[26] E. J. Michaud, Z. Liu, and M. Tegmark, "Precision machine learning," *Entropy*, vol. 25, no. 1, pp. 1–17, 2023, doi: 10.3390/e25010175.
[27] M. F. Amin, "Confusion matrix in three-class classification problems: a step-by-step tutorial," *Journal of Engineering Research*, vol. 7, no. 1, p. 26, 2023, doi: 10.21608/erjeng.2023.296718.
[28] S. Sharma and K. Guleria, "A deep learning based model for the detection of pneumonia from chest x-ray images using vgg-16 and neural networks," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 357–366, doi: 10.1016/j.procs.2023.01.018.
[29] D. Herath, C. Abeyrathne, and P. Jayaweera, "Vision-based driver drowsiness monitoring: comparative analysis of yolov5-v11 models," *aXiv Preprint*, pp. 1–13, 2025, doi: 10.48550/arXiv.2509.17498.
[30] M. A. Megaarta, "Comparative evaluation of yolov5 and yolov8 models in detecting smoking behavior," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 4, no. 3, pp. 2048–2056, Jun. 2025, doi: 10.59934/jaiea.v4i3.1089.
[31] V. Hendriko and D. Hermanto, "Performance comparison of yolov10, yolov11, and yolov12 models on human detection datasets," *Brilliance: Research of Artificial Intelligence*, vol. 5, no. 1, pp. 440–450, Jul. 2025, doi: 10.47709/brilliance.v5i1.6447.
[32] A. S. Geetha and M. Hussain, "A comparative analysis of yolov5, yolov8, and yolov10 in kitchen safety," *aXiv Preprint*, pp. 1–17, 2024, doi: 10.48550/arXiv.2407.20872.

BIOGRAPHIES OF AUTHORS

Helmi Wibowo received a Bachelor's degree in Electrical Engineering Education (S.Pd.) from Universitas Pendidikan Indonesia in 2013 and a Master's degree in Engineering (M.T.) from Institut Teknologi Bandung in 2016. His research focused on computer vision and control systems related to transportation safety. He can be contacted at email: helmi.wibowo@pktj.ac.id.

Muh Irhas Rafiqi was born in November 2002 in Indonesia. He obtained a Bachelor of Applied Engineering (S.Tr.T.) degree in Automotive Engineering Technology from the Polytechnic of Road Transportation Safety in 2025. His academic interests focus on intelligent transportation systems, computer vision, and the application of deep learning in transportation safety. He can be contacted at email: irhas.rafiqi@gmail.com.