DOI: 10.22146/jnteti.v15i2.24611

Student Behavior Detection Using YOLOv10 for Classroom Engagement Analysis

Resa Pramudita[1], Mochamad Rizal Fauzan[2], Ilyasa Nafan Faza[1], Jaja Kustija[3], Ibnu Hartopo[1], Muhammad Adli Rizqulloh[4]

[1] Industrial Automation and Robotics Engineering Education Study Program, Faculty of Engineering Education and Industry, Universitas Pendidikan Indonesia, Bandung, Jawa Barat 40154, Indonesia
[2] Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Taipei University of Technology, Taipei 10608, Taiwan (R.O.C.)
[3] Electrical Engineering Education Study Program, Faculty of Engineering Education and Industry, Universitas Pendidikan Indonesia, Bandung, Jawa Barat 40154, Indonesia
[4] Department of Computer Engineering, College of Computing and Mathematics, King Fahd University of Petroleum & Minerals Dhahran, Dhahran 31261, Saudi Arabia

[Received: 2 October 2025, Revised: 13 December 2025, Accepted: 6 March 2026]
Corresponding Author: Mochamad Rizal Fauzan (email: rizalfauzan2002@gmail.com)

ABSTRACT — Student engagement is a critical determinant of learning effectiveness, yet manual observation in classroom environments remains labor-intensive, subjective, and difficult to scale. This study examined a student behavior detection framework built on You Only Look Once (YOLO) version 10 or YOLOv10, the latest generation of real-time object detection models. A dataset of 2,600 annotated classroom images covering eight behavioral categories was collected under diverse conditions, including variations in lighting, camera perspectives, and occlusion. Five YOLOv10 variants (n, s, m, l, x) were trained and evaluated using precision, recall, F1 score, and mean average precision (mAP). The best-performing configuration achieved an overall mAP@0.5 of 0.821 and mAP@0.5:0.95 of 0.640, with strong performance on upright (AP = 0.967), bow head (AP = 0.958), and sleep (AP = 0.943), while more subtle behaviors such as writing (AP = 0.519) and hand-raising (AP = 0.650) proved challenging. Importantly, the system maintained real-time inference speeds ranging from 40 to 88 FPS depending on the YOLOv10 variant, when evaluated on an RTX 2060 GPU, thereby demonstrating its robustness for deployment in classroom settings. To ensure usability, the optimized YOLOv10 model was integrated into a Streamlit-based interactive dashboard, enabling educators to monitor engagement levels and respond with timely interventions. By combining state-of-the-art YOLOv10 architecture with real-time behavioral analytics, this work establishes a scalable foundation for intelligent classroom monitoring and contributes to advancing technology-enhanced education.

KEYWORDS — Student Engagement, YOLOv10, Classroom Behavior Detection, Deep Learning, Real-Time Analytics, Educational Technology.

I. INTRODUCTION

Student engagement is widely recognized as a fundamental determinant of learning effectiveness and academic success. Active participation such as raising hands, reading, or maintaining attention has been proven to enhance knowledge retention and improve classroom performance. In contrast, disengaged behaviors such as smartphone use, head bowing, or sleeping can reduce comprehension levels by up to 40% [1], leading to significant learning gaps and reduced instructional efficiency in large or dynamic classroom environments. For educators, the ability to monitor engagement in real time is therefore critical to providing timely feedback and adaptive interventions [2], [3].

The integration of computer vision and deep learning has created new opportunities to address this challenge. Conventional convolutional neural networks and transformer-based models have been used in behavioral recognition with accuracy levels often exceeding 85% [4], [5]. However, these approaches are limited by high computational costs and relatively slow inference, making them unsuitable for real-time classroom monitoring [6]. To overcome such challenges, the YOLO family of object detection models has emerged as a dominant solution due to its ability to achieve both high accuracy and high inference speed [7].

Recent studies have highlighted the effectiveness of You Only Look Once (YOLO) in real-world applications. For instance, YOLOv8-based systems for human activity and driver monitoring achieved 97.5% precision, 92.8% recall, 95.1% F1-score, and 96.7% mean average precision (mAP), demonstrating robust performance across varied environments [8], [9]. More recently, YOLOv9 advanced these capabilities, with results showing 99.4% precision, 99.6% recall, 99.5% F1 score, and 85.5% mAP@50–95, while maintaining real-time inference speeds of 52.08 FPS in drowsiness detection tasks [10]. These findings confirm the suitability of YOLO architectures for time-sensitive, safety-critical, and dynamic environments [11], [12].

Despite these advances, applications in educational settings remain limited. Unlike driver monitoring scenarios where facial landmarks are more constrained, classroom environments introduce additional complexities, including occlusion among students, overlapping gestures, variable camera angles, and similar visual postures. Such challenges hinder the accuracy of existing models and reduce their ability to generalize across diverse learning contexts. Furthermore, most prior approaches do not integrate detection models with interactive interfaces that allow teachers to interpret and act upon engagement data in real time [13]. However, existing studies predominantly focus on earlier YOLO versions and lack a systematic evaluation of YOLOv10 for real-time student engagement analysis in complex classroom environments.

To address these limitations, this study developed a student behavior detection framework based on YOLOv10. A dataset of 2,600 classroom images was collected and manually annotated, covering both active behaviors, such as hand raising, reading, and upright attention, and non-active behaviors,

including phone use, bowing the head, and sleeping. Multiple YOLOv10 variants (n, s, m, l, and x) were trained and systematically evaluated using standard object detection metrics, including precision, recall, F1 score, and mAP. Through this framework, the proposed approach aims to achieve accurate and real-time analysis of student engagement under diverse classroom conditions.

Finally, to ensure practical applicability, the selected YOLOv10 model was integrated into a Streamlit-based interactive dashboard for real-time classroom monitoring. Unlike previous studies that primarily relied on YOLOv8 or YOLOv9, this work introduces YOLOv10 into the educational context for the first time, offering enhanced accuracy and efficiency. The contributions of this study are: (a) the development of a YOLOv10-based framework for classroom engagement detection, (b) the comprehensive evaluation of multiple YOLOv10 variants under diverse classroom conditions, and (c) the deployment of an interactive dashboard that bridges advanced object detection models with practical educational needs. By combining state-of-the-art YOLOv10 architecture with a real-time implementation framework, this work aims to establish a new benchmark in intelligent classroom behavior analysis.

II. METHODOLOGY

This study adopted a systematic methodology to develop a real-time student behavior detection system based on YOLOv10. The overall research workflow was designed to ensure clarity, reproducibility, and practical applicability in real classroom environments. As illustrated in Figure 1, the proposed methodology consists of six main stages: data collection, annotation and labeling, preprocessing and data augmentation, YOLOv10 model training, model evaluation, and real-time deployment through an interactive dashboard.

In the first stage, classroom images were collected under diverse conditions to capture realistic learning environments, including variations in lighting, camera angles, and student seating arrangements. The collected images were then manually annotated by assigning bounding boxes and behavior labels to each student instance. Subsequently, preprocessing and data augmentation techniques were applied to standardize input dimensions and enhance model robustness against environmental variations.

Next, multiple YOLOv10 model variants are trained using the prepared dataset to analyze the trade-off between detection accuracy and computational efficiency. The trained models are evaluated using standard object detection metrics, including precision, recall, F1 score, and mAP. Finally, the selected model was deployed in a real-time interactive dashboard, enabling practical classroom monitoring and visualization of student engagement. The details of each stage in the proposed methodology are described in the following subsections.

A. DATASET COLLECTION

The dataset used in this study comprised 2,600 classroom images collected under diverse real-world conditions, including variations in lighting (daylight and dim indoor settings) and camera perspectives (frontal, lateral, and diagonal angles). These variations were intentionally incorporated to capture realistic classroom dynamics, such as partial occlusion, overlapping student postures, and diverse behavioral expressions commonly encountered in actual learning environments.

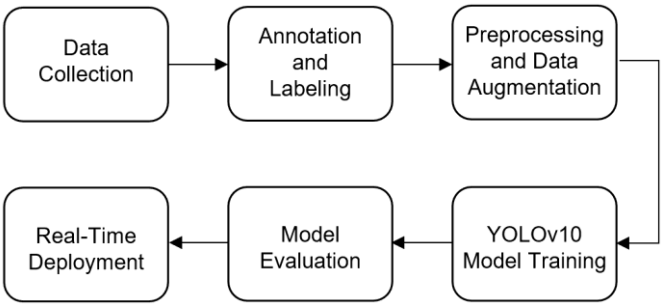

Figure 1. Proposed YOLOv10-based student behavior detection system.

The dataset was categorized into eight distinct student behavior classes representing active and nonactive engagement. The active engagement category included behaviors such as hand raising, reading, and raising the head in an attentive posture, while the nonactive engagement category consisted of using a mobile phone, bending forward, bowing the head, sleeping, and turning the head away. This classification scheme was defined based on prior studies on student engagement analysis and commonly observed classroom participation and distraction patterns reported in educational research, ensuring that the selected behaviors were pedagogically relevant and grounded in the literature.

Table I presents representative sample images for each of the eight behavioral categories, illustrating the visual diversity and complexity of student behaviors in classroom settings. These examples highlight the differences between engaged and disengaged behaviors that form the foundation for training the YOLOv10-based detection model.

To avoid bias during training, the dataset was deliberately constructed to be approximately balanced, with each behavior category containing between 300 and 350 annotated samples. This controlled distribution eliminated the need for additional resampling techniques and supported fair and reliable model learning across all behavior classes. The balanced design of the dataset contributed to the robustness of the experimental results and enhanced the generalizability of the proposed framework for real-world classroom deployment.

B. DATASET PREPARATION AND LABELING

The dataset preparation and labeling process was conducted in a structured and systematic manner to ensure high data quality and consistency for model training. Initially, each classroom image was manually annotated using a graphical user interface annotation tool. During this stage, bounding boxes were drawn around individual students, and each bounding box was assigned to one of the eight predefined student behavior categories. This manual annotation process provided accurate ground truth labels that served as the basis for supervised learning.

Following the annotation stage, a preprocessing pipeline was applied to prepare the data for model training. Using the annotated bounding boxes, regions of interest (ROIs) corresponding to individual students were extracted to remove irrelevant background information. These ROIs were then resized to a standardized resolution of 320 × 320 pixels and normalized to values between 0 and 1, ensuring uniform input dimensions and stable convergence during training.

To enhance robustness and mitigate overfitting, data augmentation techniques were applied to the cropped ROIs.

TABLE I
EXAMPLE IMAGES FROM THE DATASET USED

No.	Student Condition	Images
1.	Using phone	
2.	Bend	
3.	Bow head	
4.	Hand raising	
5.	Sleep	
6.	Reading	
7.	Turn head	
8.	Raise Head	

The augmentation process included horizontal flipping, random rotations within ±15°, and brightness adjustments of up to ±20%. These transformations were designed to simulate natural variations commonly observed in classroom environments, such as changes in viewing angle, posture orientation, and illumination conditions.

Figure 2 illustrates the overall preprocessing workflow, starting from the original classroom image, followed by manual annotation, ROI extraction based on annotated bounding boxes, and final behavior classification with an associated confidence score. As shown in the figure, the student's behavior was correctly identified as writing, demonstrating the effectiveness of the labeling and preprocessing pipeline. This structured preparation and labeling workflow ensured high annotation quality, balanced representation across behavior categories, and realistic data variation, all of which were essential for reliable training and evaluation of the proposed YOLOv10-based student behavior detection model.

C. YOLOv10 MODEL ARCHITECTURE

This research employed YOLOv10, the most recent generation of the YOLO real-time object detection models developed as the latest YOLO generation by the Ultralytics team. Unlike its predecessors, YOLOv10 introduces a holistic redesign that eliminates the need for non-maximum suppression (NMS) during inference by leveraging consistent dual assignments in training [14]. This NMS-free approach reduces inference latency while maintaining prediction reliability, making YOLOv10 highly efficient for real-time applications [15].

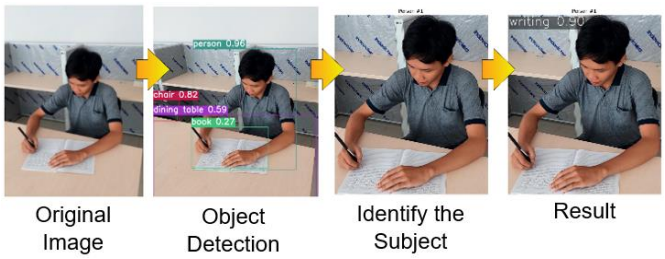

Figure 2. Preprocessing workflow of the dataset, starting from the original image, object detection with bounding boxes, ROI extraction, and final classification of the student's behavior.

The model architecture consisted of three primary components. The backbone, based on an enhanced cross stage partial network (CSPNet), was responsible for feature extraction with improved gradient flow and reduced redundancy. The neck, implemented using a path aggregation network (PAN), fused multiscale features to strengthen object detection across varying sizes [16]. Finally, the dual detection heads were divided into a one-to-many head used during training, generated rich supervisory signals, and a one-to-one head used during inference, thereby ensuring a single best prediction per object and enables NMS-free detection [17].

YOLOv10 also integrates several efficiency and accuracy enhancements. From the efficiency perspective, it adopts a lightweight classification head, spatial-channel decoupled down-sampling, and rank-guided block design to minimize redundancy and computational cost [18]. From the accuracy perspective, it incorporates large-kernel convolutions to enlarge the receptive field and partial self-attention (PSA) modules to improve global feature representation without significant overhead.

Multiple YOLOv10 variants were tested in this study, including YOLOv10n, YOLOv10s, YOLOv10m, YOLOv10l, and YOLOv10x, each designed to balance accuracy and computational requirements for different deployment scenarios. According to Ultralytics benchmarks on the COCO dataset, YOLOv10s achieved 46.8% mAP with 21.6 GFLOPs and 2.49 ms latency, while YOLOv10x reached 54.4% mAP with 160.4 GFLOPs and 10.7 ms latency. These results highlight the superior accuracy–latency trade-offs of YOLOv10 compared to YOLOv9 and other state-of-the-art detectors.

The simplified architecture of YOLOv10 used in this study is illustrated in Figure 3, highlighting the flow from the input image through the CSPNet backbone and PAN neck to the dual detection heads, consisting of a one-to-many head for training and a one-to-one head for inference. By adopting this architecture, the present study leveraged YOLOv10's unique balance of speed accuracy to address the challenges of classroom environments, including occlusion, overlapping gestures, and visually similar postures.

D. TRAINING CONFIGURATION AND OPTIMIZATION

The training of the YOLOv10 models was done using the Ultralytics framework on an NVIDIA GPU with CUDA acceleration to ensure efficient computation. The dataset was

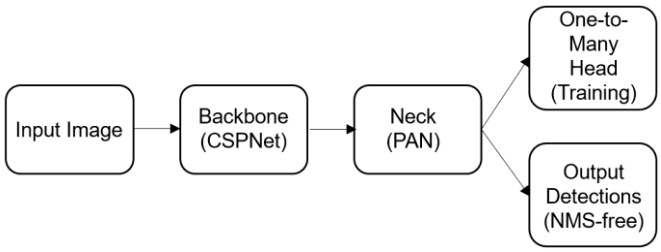

Figure 3. Simplified architecture of YOLOv10, showing the CSPNet backbone, PAN neck for multi-scale feature fusion, and dual detection heads for training and NMS-free inference.

split into 80% training, 10% validation, and 10% testing for balanced representation in all eight behavioral categories.

The training process was configured with 100 epochs, a batch size of 32, and an initial learning rate of 0.001 following a cosine decay schedule for stable convergence. The AdamW optimizer was employed, thereby balancing adaptive learning with weight decay regularization to mitigate overfitting. The loss function integrated classification, objectness, and bounding box regression losses through the task-aligned assignment (TAL) mechanism, which aligns one-to-many and one-to-one strategies to enhance training quality [19].

During training, all images were resized to 640 × 640 pixels, while augmentation techniques such as flipping, random rotations of up to ±15°, and brightness adjustments of ±20% were applied to improve generalization. Early stopping with a patience of 20 epochs was enabled to prevent unnecessary computation once performance plateaued.

Performance monitoring relied on mAP (mAP@0.5 and mAP@0.5:0.95), precision, recall, and F1 score metrics evaluated on the validation set. The best-performing model weights, based on validation mAP, were saved for final testing on the unseen dataset. The detailed hyperparameters used for model training are summarized in Table II.

TABLE II
TRAINING HYPERPARAMETERS

Hyperparameter	Value/Setting	Description
Input Size	640 × 640 pixels	Resized resolution for YOLOv10 training
Batch Size	32	Number of images per iteration
Epochs	100	Maximum training iterations
Learning Rate	0.001 (cosine decay scheduler)	Initial learning rate with adaptive scheduling
Optimizer	AdamW	Adaptive learning with weight decay to prevent overfitting
Loss Function	Classification + Objectness + Box (TAL)	Task-Aligned Assignment combining one-to-many and one-to-one strategies
Early Stopping	Patience = 20 epochs	Stop training if no improvement on validation set
Data Split	80% train / 10% validation / 10% test	Balanced distribution across all behavior categories
Augmentation	Flip, rotation (±15°), brightness (±20%)	Improve generalization and robustness
Hardware	NVIDIA GPU RTX 2060 with CUDA acceleration	High-performance training environment

E. EVALUATION MODEL

The performance of each trained model was evaluated using a series of quantitative metrics, including precision, recall, F1 score, mAP, and processing speed (inference latency). This evaluation was conducted under multiple scenarios to test the robustness of the models, such as varying lighting conditions, different camera positions, and diverse viewing perspectives.

1) PRECISION

Precision measures the accuracy of positive predictions by comparing the number of correct positive predictions (true positives, TP) with the total number of predicted positives, namely TP + false positive (FP). High precision indicates that the model makes fewer false-positive errors.

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

2) RECALL

Recall evaluates the completeness of detection, i.e., the ability of the model to identify all relevant objects. It is defined as the ratio between correctly predicted positives and all actual positives, namely TP + false negative (FN).

$$P = \frac{TP}{TP+FN}. \quad (2)$$

3) F1 SCORE

F1 score represents the harmonic mean of precision and recall, balancing both aspects into a single performance indicator.

$$F1 = \frac{2 \times P \times R}{P+R}. \quad (3)$$

4) MEAN AVERAGE PRECISION (mAP)

MAP provides a comprehensive assessment across all classes by averaging the average precision (AP) values for each class q in the dataset. In this study, two variations were used: mAP@0.5 and mAP@0.5:0.95. The general formula is presented in (4).

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q} AP(q). \quad (4)$$

where Q is the total number of classes.

In addition to accuracy-related metrics, inference speed (FPS) was recorded to evaluate the feasibility of deploying the models in real-time classroom environments. Together, these metrics provide a holistic view of model performance, balancing accuracy, robustness, and efficiency.

III. RESULTS AND DISCUSSION

A. QUANTITATIVE RESULTS

Table III provides a comparative analysis of YOLOv8s, YOLOv9s, and the YOLOv10 family on the classroom behavior dataset. The results demonstrated that YOLOv10 consistently surpassed earlier generations, establishing new performance baselines.

Specifically, YOLOv10s achieved 0.82 mAP@0.5 and 0.59 mAP@0.5:0.95, representing a relative improvement of 9.3% over YOLOv8s and 5.4% over YOLOv9s in terms of mAP@0.5:0.95. Similarly, YOLOv10m and YOLOv10l further improved performance, reaching 0.61 and 0.63

TABLE III
PERFORMANCE COMPARISON OF YOLO VARIANTS ON THE CLASSROOM BEHAVIOR DATASET

Model	mAP@0.5	mAP@0.5:0.95	Precision	Recall	F1 Score	FPS
YOLOv8s	0.75	0.52	0.74	0.76	0.75	85
YOLOv9s	0.78	0.55	0.76	0.77	0.76	80
YOLOv10n	0.77	0.54	0.75	0.77	0.76	95
YOLOv10s	0.82	0.59	0.81	0.80	0.80	88
YOLOv10m	0.84	0.61	0.82	0.82	0.82	70
YOLOv10l	0.85	0.63	0.83	0.82	0.83	55
YOLOv10x	0.86	0.64	0.84	0.83	0.84	40

mAP@0.5:0.95, while maintaining inference speeds above 50 FPS on an RTX 2060 GPU, thereby sustaining real-time operability. These improvements can be attributed to several architectural innovations. The NMS-free dual-head design enables more efficient training–inference alignment, reducing prediction redundancies while preserving accuracy. The CSPNet-based backbone strengthens gradient flow, while the PAN neck facilitates effective multi-scale feature fusion, which is particularly beneficial in classroom settings where student gestures and postures vary widely in scale and occlusion. Furthermore, the integration of large-kernel convolutions and partial self-attention modules enhances the receptive field and global representation, contributing to the superior recall and F1 scores compared to YOLOv8 and YOLOv9.

Figure 4 presents the training and validation curves of YOLOv10, highlighting the learning dynamics over 100 epochs. The steady decline in box, classification, and distribution focal losses indicates smooth convergence, while the simultaneous improvements in precision, recall, and mAP metrics illustrate that the model effectively generalizes without severe overfitting. Notably, precision stabilized above 0.82 and recall maintained around 0.80 in the later epochs, corroborating the numerical results in Table III.

The combination of Table III and Figure 4 provides complementary evidence of YOLOv10's robustness. While YOLOv10x offered the highest detection accuracy (0.64 mAP@0.5:0.95), the trade-off in inference speed (40 FPS) suggests it was more suitable for offline analysis rather than real-time classroom deployment. In contrast, YOLOv10s and YOLOv10m provided the best balance, achieving real-time speeds (70–88 FPS) with competitive accuracy, making them ideal candidates for intelligent classroom monitoring systems. These findings are consistent with recent benchmarks in real-time detection models, reinforcing YOLOv10's position as a state-of-the-art detector for education-related computer vision tasks.

B. CLASS-WISE PERFORMANCE

The class-wise evaluation revealed noticeable differences in detection performance across various student behaviors. Distinct and visually salient actions, such as upright attention (AP = 0.967), bowing the head (AP = 0.958), and sleeping (AP = 0.943), achieved high detection accuracy, indicating that these behaviors possess strong and consistent visual characteristics that are easily distinguishable by the model. In contrast, more subtle behaviors, including writing (AP = 0.519) and hand raising (AP = 0.650), exhibited lower detection performance due to their visual similarity with other actions and higher intra-class variability. Overall, the model achieved an average mAP@0.5 of 0.821, confirming robust recognition performance across the majority of behavior classes.

As illustrated in Figure 5, the precision–recall (PR) curves provide deeper insight into the detection behavior of each class. Classes associated with distinct postures demonstrated higher and more stable PR curves, maintaining strong recall even at high precision levels. Conversely, behaviors with subtle visual differences showed steeper precision–recall trade-offs, where recall decreased more rapidly as precision increases. This trend indicates that the model becomes more conservative when distinguishing visually ambiguous behaviors, prioritizing precision at the expense of recall.

Further analysis using the normalized confusion matrix, shown in Figure 6, highlights the specific sources of misclassification. The strong diagonal dominance confirmed high classification accuracy for most behavior categories. However, certain confusion patterns were observed, such as writing being misclassified as reading or hand raising being confused with upright attention. These errors primarily arose from fine-grained posture similarities and partial occlusions in crowded classroom environments. The findings suggest that while the YOLOv10-based model performs effectively overall, incorporating temporal information or multimodal cues in future work may further improve the discrimination of visually similar behaviors.

C. ROBUSTNESS UNDER ENVIRONMENTAL VARIATIONS

To assess robustness, the YOLOv10 model was evaluated across diverse environmental variations that often occur in real classrooms. These included differences in illumination (bright daylight vs. dim artificial lighting), variations in seating arrangements where students were positioned at different distances from the camera, and multiple viewing angles resulting from shifted camera perspectives. Such factors are known to influence object detection performance, as shadows, occlusions, and perspective distortions can reduce classification accuracy.

Figure 7 presents representative detection outcomes obtained under these challenging conditions. Despite the increased complexity, the model consistently produced reliable detections with confidence scores typically above 0.70. Frequent behaviors such as upright sitting, bending, and bowing head were recognized with high stability, while more subtle postures such as turning head or raising head were still detected with reasonable accuracy. Importantly, even when multiple students appeared in crowded settings, YOLOv10 maintained clear differentiation of individual behaviors without introducing significant false positives [20]. These results demonstrate that the model generalizes well beyond controlled laboratory setups, confirming its robustness and suitability for real-world classroom monitoring applications.

D. COMPARATIVE ANALYSIS WITH STATE-OF-THE-ART

The comparative evaluation highlights the performance of YOLOv10 against prior state-of-the-art detectors, namely

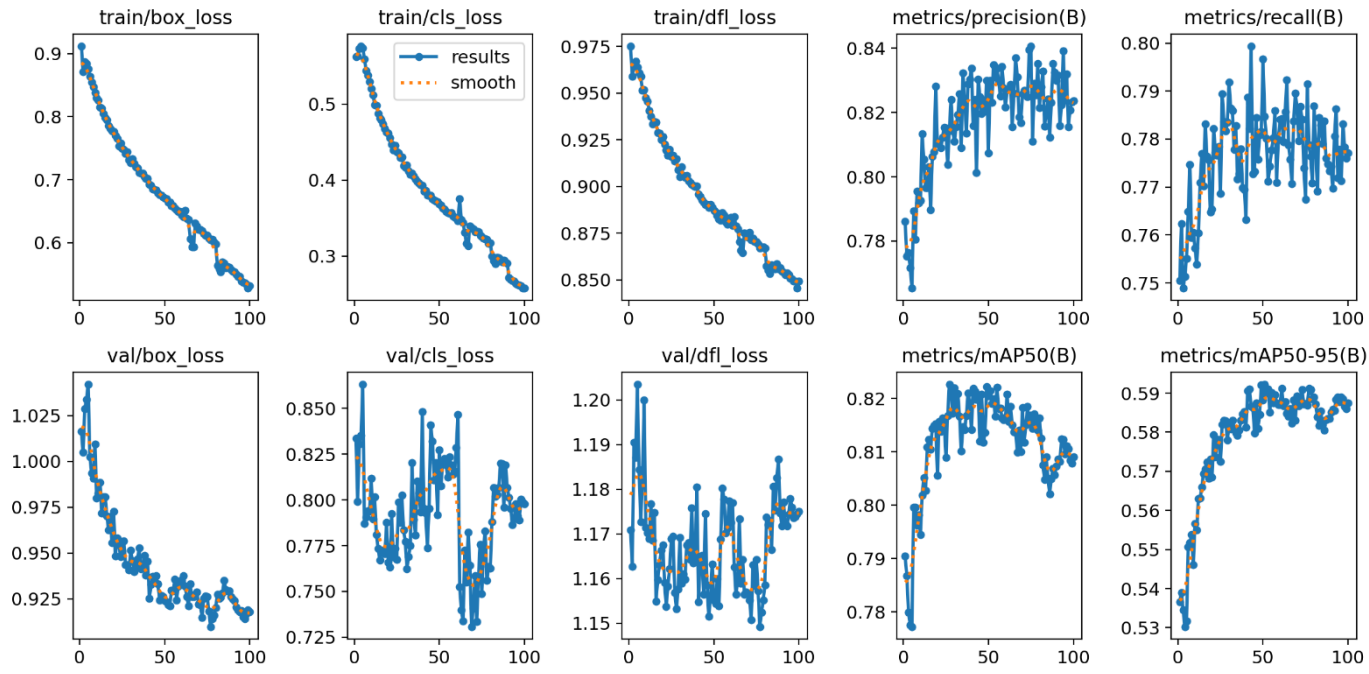

Figure 4. Training and validation curves of YOLOv10 on the classroom behavior dataset. The plots demonstrate consistent loss reduction and metric improvements, reflecting stable convergence and generalization.

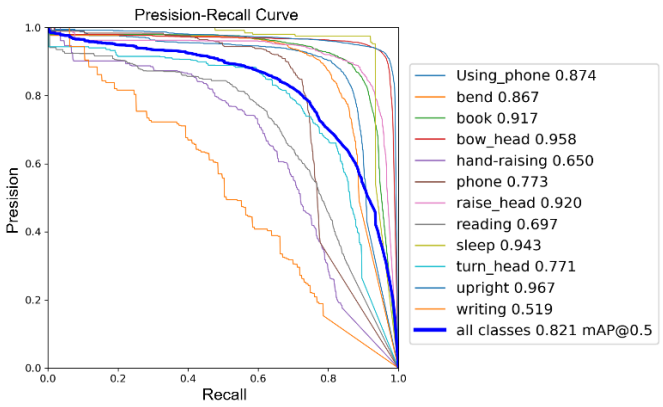

Figure 5. Precision–recall (PR) curves for each behavior class, showing variability in per-class detection performance.

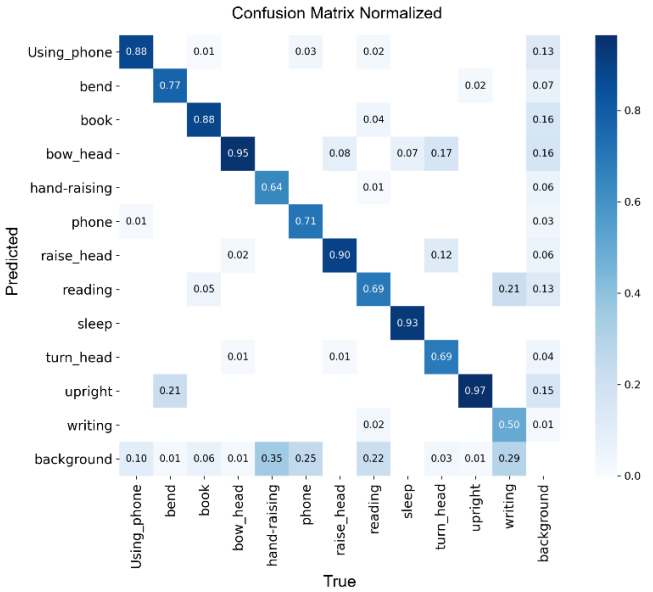

Figure 6. Normalized confusion matrix of YOLOv10 predictions, indicating strong overall accuracy with notable misclassification trends.

YOLOv8 and YOLOv9, using the classroom behavior dataset. As shown in Table III, YOLOv8s serves as a baseline with moderate detection accuracy and inference efficiency [21], [22]. YOLOv9s demonstrated incremental improvements, particularly in precision and recall, attributable to its generalized efficient layer aggregation network (GELAN) architecture and programmable gradient information. However, despite these enhancements, YOLOv9s maintained higher computational demands and relatively lower inference speed compared to the latest YOLOv10 variants [23], [24].

The YOLOv10 series consistently outperformed both YOLOv8 and YOLOv9 across nearly all evaluation metrics [25], [26]. For instance, YOLOv10s achieved a higher mAP@0.5 and F1 score while sustaining faster inference speed on RTX 2060 GPU, underscoring its suitability for real-time classroom monitoring. Larger variants such as YOLOv10m and YOLOv10x delivered superior accuracy (mAP@0.5:0.95 exceeding 50%) without compromising reliability under dense and occluded scenes, although at the expense of higher computational cost [24]. These results reinforce that YOLOv10 establishes a new benchmark for balancing accuracy, robustness, and efficiency in student behavior detection tasks [27].

These findings confirm that YOLOv10 not only surpasses prior YOLO architectures in accuracy and inference efficiency but also demonstrates practical readiness for deployment in real-world classroom monitoring systems. By delivering a superior balance of precision, robustness, and speed, YOLOv10 sets a new benchmark for state-of-the-art behavior

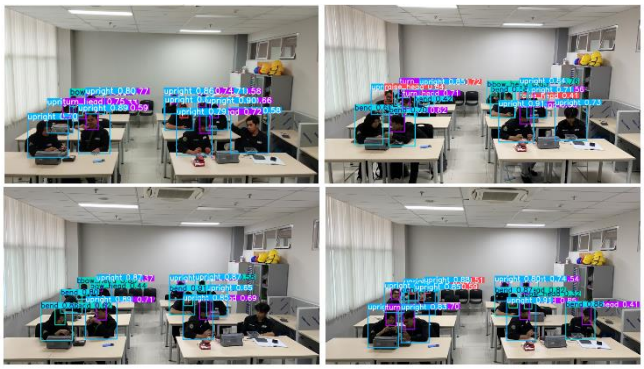

Figure 7. Robustness evaluation of YOLOv10 under environmental variations. The examples show consistent detection performance across differences in illumination, seating arrangements, and camera perspectives, highlighting the model's adaptability to real-world classroom conditions.

detection and provides a scalable foundation for future intelligent education technologies.

IV. CONCLUSION

This study presented a YOLOv10-based framework for real-time student behavior detection to support classroom engagement analysis. Experimental results demonstrated that YOLOv10 consistently outperformed previous YOLO variants in terms of detection accuracy while maintaining real-time inference capability. Among the evaluated models, YOLOv10s and YOLOv10m provided the most balanced trade-off between accuracy and speed, achieving reliable detection performance under diverse classroom conditions, including variations in lighting, camera angles, and student postures. Distinct behaviors such as upright attention and sleeping were detected with high accuracy, while visually similar actions, such as writing and hand raising, remained more challenging due to subtle posture differences and occlusion.

Overall, the proposed framework demonstrates strong potential for practical deployment in intelligent classroom monitoring systems. Future work will focus on expanding the dataset across different educational environments, incorporating temporal information to better capture continuous behavior dynamics, and integrating multimodal data sources to further improve the recognition of subtle student behaviors.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest related to the content or findings of this study.

AUTHORS' CONTRIBUTIONS

Conceptualization, Resa Pramudita, Mochamad Rizal Fauzan, and Jaja Kustija; methodology, Resa Pramudita and Mochamad Rizal Fauzan; software, Mochamad Rizal Fauzan and Muhammad Adli Rizqulloh; validation, Resa Pramudita, Ilyasa Nafan Faza, and Ibnu Hartopo; formal analysis, Resa Pramudita and Mochamad Rizal Fauzan; investigation, Resa Pramudita and Ilyasa Nafan Faza; resources, Universitas Pendidikan Indonesia and National Taipei University of Technology; data curation, Resa Pramudita and Muhammad Adli Rizqulloh; writing—original draft preparation, Resa Pramudita and Mochamad Rizal Fauzan; writing—reviewing and editing, Jaja Kustija and Ibnu Hartopo; visualization, Mochamad Rizal Fauzan; supervision, Jaja Kustija and Ibnu Hartopo; project administration, Resa Pramudita; funding acquisition, none.

ACKNOWLEDGMENT

The authors would like to sincerely thank the Department of Industrial Automation and Robotics Engineering Education and the Department of Electrical Engineering Education, Faculty of Technology and Vocational Education, Universitas Pendidikan Indonesia, for their continuous guidance and academic support. Special appreciation is also extended to the Department of Electrical Engineering, National Taipei University of Technology, Taiwan, and the Department of Computer Engineering, King Fahd University of Petroleum & Minerals, Saudi Arabia, for providing valuable insights and facilities that greatly contributed to the successful completion of this research.

REFERENCES

[1] L. Li *et al.,* "ET-YOLOv5s: Toward deep identification of students' in-class behaviors," *IEEE Access*, vol. 10, pp. 44200–44211, Apr. 2022, doi: 10.1109/ACCESS.2022.3169586.

[2] D. Zhou *et al.,* "MFDS-STGCN: Predicting the behaviors of college students with fine-grained spatial-temporal activities data," *IEEE Trans. Emerg. Top. Comput.,* vol. 12, no. 1, pp. 254–265, Jan.-Mar. 2024, doi: 10.1109/TETC.2023.3344131.

[3] Y. Shi, F. Sun, H. Zuo, and F. Peng, "Analysis of learning behavior characteristics and prediction of learning effect for improving college students' information literacy based on machine learning," *IEEE Access*, vol. 11, pp. 50447–50461, May 2023, doi: 10.1109/ACCESS.2023.3278370.

[4] S.A. Amoudi *et al.,* "Click-based representation learning framework of student navigational behavior in MOOCs," *IEEE Access*, vol. 12, pp. 121480–121494, Aug. 2024, doi: 10.1109/ACCESS.2024.3450514.

[5] N. Ruiz *et al.,* "ATL-BP: A student engagement dataset and model for affect transfer learning for behavior prediction," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 5, no. 3, pp. 411–424, Jul. 2023, doi: 10.1109/TBIOM.2022.3210479.

[6] S. Kim *et al.,* "Characteristic behaviors of elementary students in a low attention state during online learning identified using electroencephalography," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 619–628, Jun. 2024, doi: 10.1109/TLT.2023.3289498.

[7] X.M. Zhao *et al.,* "Classroom student behavior recognition using an intelligent sensing framework," *IEEE Access*, vol. 13, pp. 49767–49776, Mar. 2025, doi: 10.1109/ACCESS.2025.3550921.

[8] H. Liu, R. Hu, H. Dong, and Z. Liu, "SFC-YOLOv8: Enhanced strip steel surface defect detection using spatial-frequency domain-optimized YOLOv8," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–11, Mar. 2025, doi: 10.1109/TIM.2025.3548193.

[9] S. Tao *et al.,* "MIS-YOLOv8: An improved algorithm for detecting small objects in UAV aerial photography based on YOLOv8," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–12, Mar. 2025, doi: 10.1109/TIM.2025.3551917.

[10] K. Xu *et al.,* "RMT-YOLOv9s: An infrared small target detection method based on UAV remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, Oct. 2024, doi: 10.1109/LGRS.2024.3484748.

[11] X. Yu *et al.,* "FEL-YoloV8: A new algorithm for accurate monitoring soybean seedling emergence rates and growth uniformity," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–17, Jun. 2025, doi: 10.1109/TGRS.2025.3578800.

[12] M. Hussain and R. Khanam, "In-depth review of YOLOv1 to YOLOv10 variants for enhanced photovoltaic defect detection," *Solar*, vol. 4, no. 3, pp. 351–386, Sep. 2024, doi: 10.3390/solar4030016.

[13] H. Sun *et al.,* "SOD-YOLOv10: Small object detection in remote sensing images based on YOLOv10," *IEEE Geosci. Remote Sens. Lett.*, vol. 22, pp. 1–5, Jan. 2025, doi: 10.1109/LGRS.2025.3534786.

[14] H. Fu *et al.,* "MSOAR-YOLOv10: Multi-scale occluded apple detection for enhanced harvest robotics," *Horticulturae*, vol. 10, no. 12, pp. 1–24, Dec. 2024, doi: 10.3390/horticulturae10121246.

[15] D. Wang *et al.,* "Real-time detection and identification of fish skin health in the underwater environment based on improved YOLOv10 model," *Aquac. Rep.*, vol. 42, pp. 1–11, Jul. 2025, doi: 10.1016/J.AQREP.2025.102723.

[16] W. Tu *et al.,* "YOLOv10-UDFishNet: Detection of diseased Takifugu rubripes juveniles in turbid underwater environments," *Aquac. Int.*, vol. 33, pp. 1–27, Jan. 2025, doi: 10.1007/s10499-024-01798-5.

[17] M. Mao, A. Lee, and M. Hong, "Efficient fabric classification and object detection using YOLOv10," *Electronics*, vol. 13, no. 19, pp. 1–23, Oct. 2024, doi: 10.3390/electronics13193840.

[18] C. Zhang *et al.,* "A novel YOLOv10-DECA model for real-time detection of concrete cracks," *Buildings*, vol. 14, no. 10, pp. 1–24, Oct. 2024, doi: 10.3390/buildings14103230.

[19] Q. Du, S. Zhang, and S. Yang, "BLP-YOLOv10: Efficient safety helmet detection for low-light mining," *J. Real-Time Image Process.*, vol. 22, pp. 1–11, Nov. 2024, doi: 10.1007/s11554-024-01587-6.

[20] R. Chai *et al.,* "Automated detection of early-stage osteonecrosis of the femoral head in adult using YOLOv10: Multi-institutional validation," *Eur. J. Radiol.*, vol. 184, pp. 1–9, Mar. 2025, doi: 10.1016/J.EJRAD.2025.111983.

[21] L. Zheng, T. Hu, and J. Zhu, "Underwater sonar target detection based on improved ScEMA-YOLOv8," *IEEE Geosci. Remote Sens. Lett.,* vol. 21, pp. 1–5, May 2024, doi: 10.1109/LGRS.2024.3397848.

[22] J. Wang *et al.,* "DPH-YOLOv8: Improved YOLOv8 based on double prediction heads for the UAV image object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, Oct. 2024, doi: 10.1109/TGRS.2024.3487191.

[23] Y. Xing *et al.,* "MAM-YOLOv9: A multiattention mechanism network for methane emission facility detection in high-resolution satellite remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–16, Feb. 2025, doi: 10.1109/TGRS.2025.3545034.

[24] A.S. Geetha, M.A.R. Alif, M. Hussain, and P. Allen, "Comparative analysis of YOLOv8 and YOLOv10 in vehicle detection: Performance metrics and model efficacy," *Vehicles*, vol. 6, no. 3, pp. 1364–1382, Sep. 2024, doi: 10.3390/vehicles6030065.

[25] W. Kong *et al.,* "A shadow-robust pavement damage detection framework based on RACycle-GAN and DDE-YOLOv8 *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 8, pp. 11342–11355, Aug. 2025, doi: 10.1109/TITS.2025.3556941.

[26] Y. Long, Y. Yang, J. Hu, and X. Huang, "Operating mechanism detection in aluminum electrolysis workshops via YOLOv8-MIE," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–15, Jan. 2025, doi: 10.1109/TIM.2024.3522436.

[27] L. Zhang *et al.,* "Intelligent psyllid monitoring based on DiTs-YOLOv10-SOD," *IEEE Trans. AgriFood Electron.*, vol. 3, no. 1, pp. 286–294, Mar./Apr. 2025, doi: 10.1109/tafe.2025.3551072.