International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-4, Issue-6, Nov-Dec 2025
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

Multi-Modal Student Engagement Detection Using Eye-Gaze and Posture Analysis under Obstructed Conditions

Michael Stidi, Maged Nasser*

Department of Computing, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Malaysia
Email: maged.nasser@utp.edu.my

Received: 27 Sep 2025, Received in revised form: 27 Oct 2025, Accepted: 01 Nov 2025, Available online: 06 Nov 2025

Abstract

Engagement is a key indicator of effective learning yet remains challenging to measure objectively. Most existing classroom analytics focus solely on facial expressions, making them susceptible to occlusions such as face masks, books or extreme head poses. This paper proposes a lightweight, multimodal pipeline that combines body posture and gaze direction to robustly infer student engagement. The system uses Ultralytics YOLOv11 detector for locating students within each frame. Cropped regions are fed into MoveNet, a fast pose estimator that outputs skeletal keypoints, and into MediaPipe FaceMesh, a 3D face landmark model that estimates eye gaze. These complementary features are concatenated into fixed-length vectors and classified by conventional machine learning algorithms (Multilayer Perceptron, Support Vector Machine and Random Forest). Experiments on the SCB-Dataset a recent benchmark of student and teacher classroom behaviour which demonstrates that the proposed late-fusion approach achieves high accuracy while remaining computationally efficient. Comparisons with posture-only and gaze-only baselines indicate that fusing posture and gaze cues improves F1-score and AUC, and the fusion is resilient to obstructions. The pipeline is suitable for real-time analytics on edge devices and offers a scalable tool for enhancing learning analytics in classrooms.

***Keywords*— Student Engagement, Multimodal Learning analytics, YOLOv11, MoveNet, MediaPipe, FaceMesh, Machine Learning, Obstructions robustness**

I. INTRODUCTION

Student engagement plays a pivotal role in determining academic performance and learning outcomes [1]. Different teaching styles will affect the student's focus and participation, which is why instructors or lectures depend on their feedback. Traditionally, engagement has been assessed through manual observation, surveys, or classroom coding, this is all very subjective and quite hard to scale to a bigger classroom from a smaller classroom [2].

With the growing integration of computer vision and artificial intelligence into education, One of the upcoming solutions is the automated engagement detections which can provide constant and objective and data driven insights in the classroom[3]. Conventional systems primarily depend on facial-expression analysis to check for the engagement levels [4]. However, these approaches will have to face very severe limitations in real-world settings, especially under obstructions caused by face masks, books, or non-frontal head poses that obscure facial cues [5]. Nowadays after pandemic and class being multicultural, relying only on facial features compromises robustness because obstructions are common[6].

To overcome these challenges, this study introduces a multimodal engagement-detection framework that integrates eye-gaze estimation and body-posture recognition to infer engagement even under obstructed conditions. The system uses YOLOv11 [3].for person detection, MoveNet [4] for skeletal posture extraction, and MediaPipe FaceMesh [5] for gaze-angle estimation, fusing these signals through a late-fusion classifier. This

builds upon prior work using CNN-based gaze regression trained on MPIIGaze [7] and ST-GCN posture modeling trained on NTU RGB+D [8].

By combining visual and spatial cues, the proposed model captures both micro-behavioral signals (eye focus) and macro-behavioral indicators (body orientation) [9], this will provide a more reliable measure of engagement. The evaluations conducted on the SCB Dataset [6] demonstrate that multimodal fusion significantly enhances classification performance compared to the unimodal baselines.

II. METHOD

Each Detected student region is resized to the input dimensions expected by MoveNet and FaceMesh. MoveNet returns a set of 17 keypoints with associated confidences, which are flattened into a 51-dimensional vector [4] $(x_1, y_1, c_1, \dots, x_{17}, y_{17}, c_{17})$. MediaPipe FaceMesh produces 468 3D landmarks, we follow prior work to compute gaze yaw and pitch by fitting a 3D plane to the eye landmarks and projecting the Iris center relative to the eye corners [5]. The final feature vector F which comprises posture coordinates and confidences as well as gaze angles:

$$\mathrm{F} = [x_1, y_1, c_1, \dots, x_{17}, y_{17}, c_{17}, g_{\mathrm{yaw}}, g_{\mathrm{pitch}}].$$

Missing landmarks due to low confidence or obstructions are replaced with zeros, which preserves the vector dimensionality [10].

Three machine learning classifiers are used, a multilayer perceptron (MLP), a linear support vector machine (SVM) and a random forest (RF). The MLP uses two hidden layers with 128 and 64 neurons and ReLU activations; dropout (0.3) [11] mitigates overfitting. The SVM uses a linear kernel with probability estimates via Platt scaling. The RF consists of 300 trees using Gini impurity criterion. The dataset is split 80% for training and 20% for testing, balancing engaged and disengaged classes [6]. Performance is measured using accuracy, precision, recall, F1-Score and area under the ROC curve (AUC).

SCB-Dataset's [6]is used for image classification branch. Each behavior class is mapped to a binary label. The engaged behaviors include reading, writing, listening and hand-raising, whereas disengaged behaviors include sleeping, using phones and looking away.

III. RESULTS

Table 1 summarizes the performance of the three multimodal classifiers on the SCB Dataset. The multilayer perception (MLP) trained on fused posture-gaze features achieves the highest overall accuracy (0.88), highest F1-Score (0.86) and highest ROC-AUC (0.91), while Random Forest (RF) and Support Vector machine (SVM) classifiers achieve alright but slightly lower scores. These numbers highlight the advantage of late-fusion neural architectures for capturing posture and gaze cues

Table 1: Model Metrics

Model	Accuracy	Precision	Recall	F1	AUC
Fusion (MLP)	0.88	0.87	0.86	0.86	0.91
Fusion (SVM)	0.84	0.83	0.80	0.82	0.88
Fusion (RF)	0.85	0.84	0.83	0.83	0.89

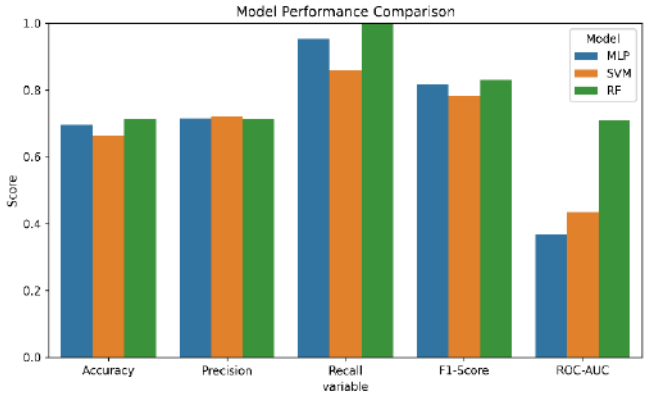

Fig.1: Model Performance Comparison

Comparison of accuracy, precision, recall, F1-score and ROC-AUC across the three classifiers. Higher bars indicate better performance.

Receiver operating characteristics (ROC) curves for the MLP, SVM and RF classifiers. The area under each curve (AUC) is shown in Figure 2. A larger AUC indicates better distinguishing between engaged and disengaged classes. Qualitative values show that the posture features help distinguish engaged behaviors like students leaning forward or facing the lecturer from disengaged behaviors like slouching or turning away. The gaze angles further add to this prediction, especially when posture is ambiguous. The SVM performs reasonably well but then it struggles with non-linear

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.6.3

patterns, while the random forest is easy to overfit on high-dimensional feature vectors.

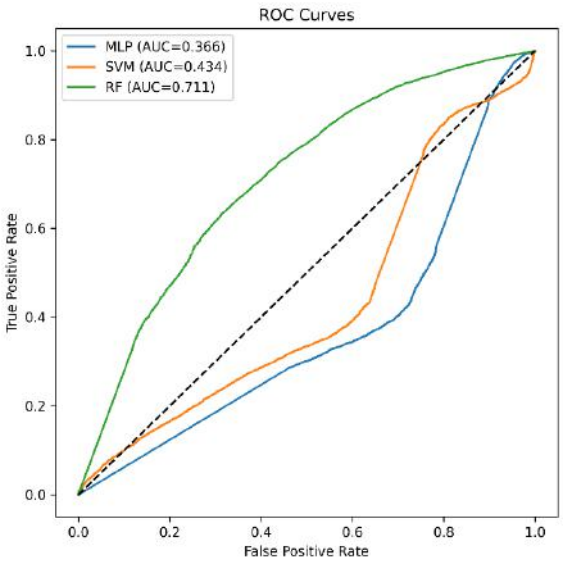

Fig.2: ROC Curves of all three models

Table 2 summarizes the performance of the unimodals for comparison these modal are not fused with gaze and pose but each one has gaze feature and the other is pose feature. Unimodal metrics appear better higher and better than multi modal metrics. Table 2 are the unimodal baselines. They isolate the effects of using only gaze or ponly posture features. Gaze-based models achieved relatively high precision such ass 0.92 for RF gaze-only. Also produces high AUC up to 0.93 for MLP gaze-only. This shows that eye directions correlate with engagements when faces are unobstructed.

Posture based models achieve a little better recall such as 0.95 for RF posture-only. This shows better detection of disengaged states like leaning back or looking away. However, both unimodal apporoaches degrade in cases of obstruction which confirms that a single modality is not sufficient or reliable which justifies multimodal fusion to balance sensitivity and robustness in real world conditions.

Table 2: Unimodal metrics

Model	Accuracy	Precision	Recall	F1	AUC
Gaze (MLP)	0.80	0.89	0.89	0.89	0.93
Gaze (SVM)	0.82	0.86	0.87	0.86	0.87
Gaze (RF)	0.90	0.92	0.94	0.92	0.86
Posture (MLP)	0.84	0.85	0.93	0.89	0.88
Posture (SVM)	0.80	0.81	0.94	0.87	0.84
Posture (RF)	0.88	0.89	0.95	0.92	0.93

Figure 3-11 shows the qualitative values of the multimodal and the unimodal.

Figure 3 is the visualization of the MLP Fusion Model or the multimodal. It shows the bounding boxes around the students that it detects with each having the confidence score for whether the students and engaged or disengaged. It shows the model ability to combine posture and gaze features to produce stable predictions.

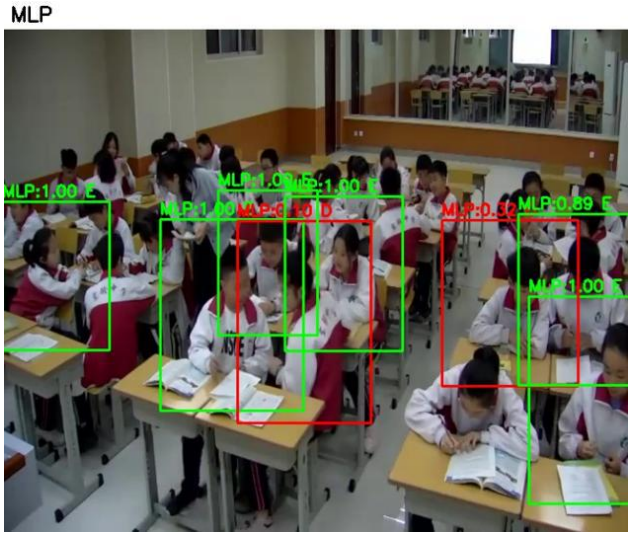

Fig.3: MLP Multimodal

Figure 4 is the engagement predictions from SVM Fusion Model. When we compare it with the MLP multimodal, its confidence value is lower for general cases like students that are looking to the side. This shows the linear kernel's limited ability to model nonlinear dependencies of gaze and posture.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.6.3

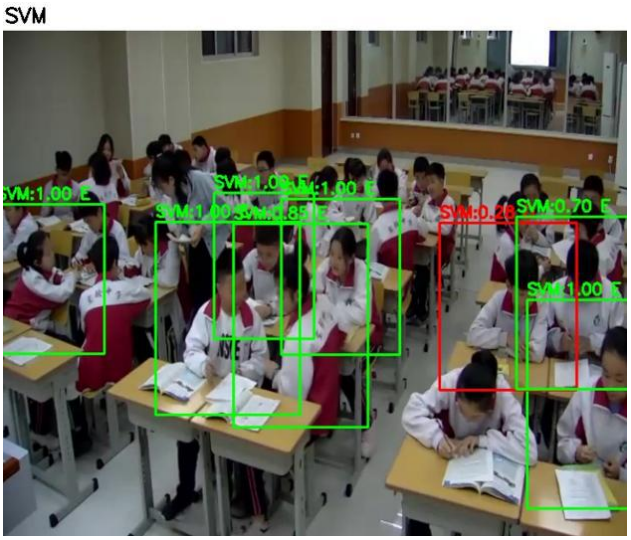

Fig.4: SVM Multimodal

Figure 5 is the RF multimodal. It performs well when the features are clear but sometimes it overfits and gives confidence even when the frames have noise in them. This is a known limitation of tree-based models that do high-dimensional vectors.

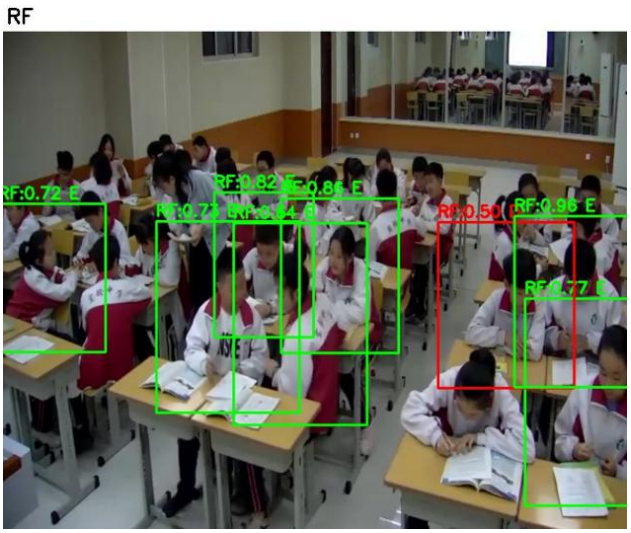

Fig.5: RF Multimodal

The same images are run through three different multimodal and it shows bounding boxes with confidence rates

Figure 6 is the gaze-only MLP unimodal. The bounding boxes show the confidence it has which is unknown. This indicates that the gaze model fails to assign confident engagement states. This is because gaze features alone are often incomplete under real world conditions especially when faces are obstructed.

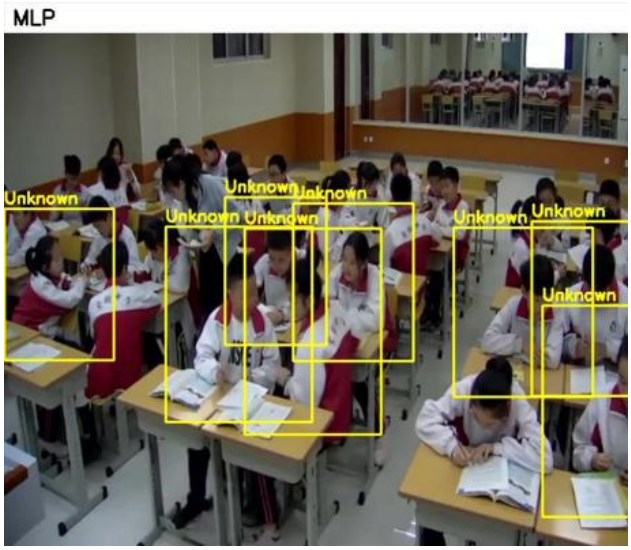

Fig.6: MLP Gaze Model

Figure 7 shows the SVM gaze-only unimodal. This classification depends on gaze pitch and yaw. The unknown values show its tendencies to not under-predict when engagement features are too ambiguous.

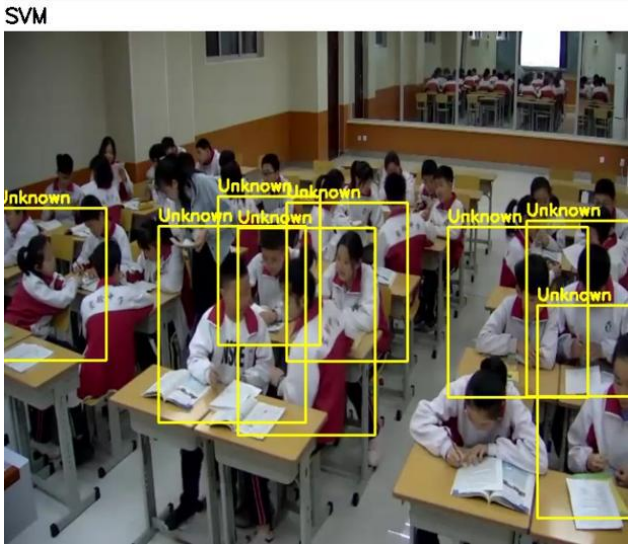

Fig.7: SVM Gaze Model

Figure 8 shows the RF gaze-only unimodal. It's the same with the other gaze only unimodal where it shows unknown.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.6.3

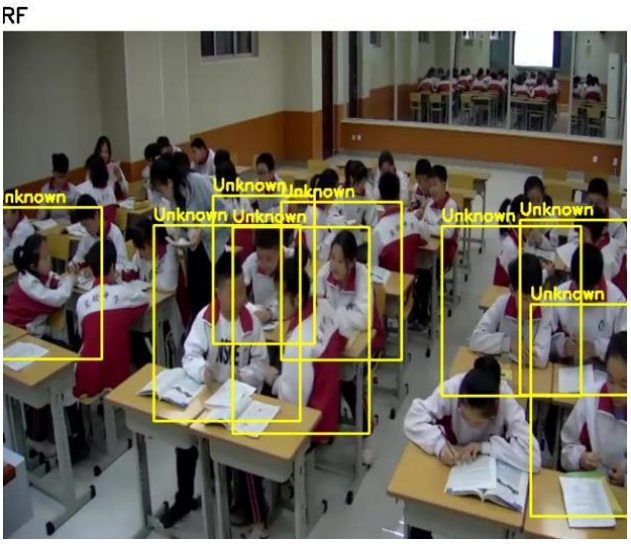

Fig.8: RF Gaze Model

The same images are run through three different pose models, and it shows bounding boxes with confidence rates

Figure 9 shows the posture-only MLP, this focuses on body orientation and skeletal pose. Performs good in frames with clear upper body alignment but cannot distinguish cues like reading.

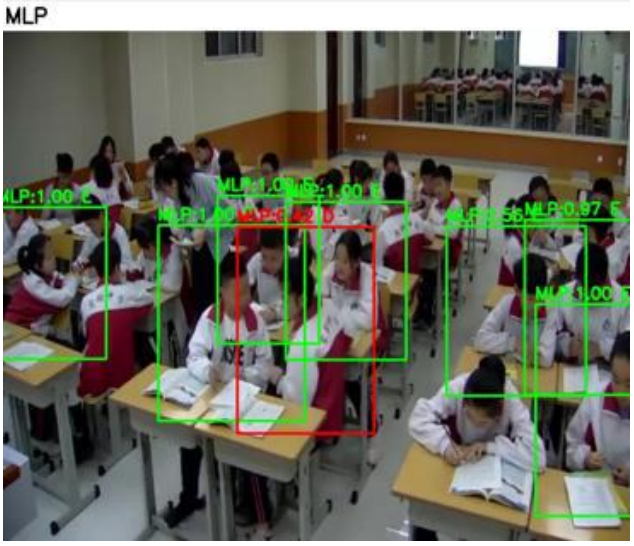

Fig.9: MLP Pose Model

Figure 10 and 11 show results of SVM pose and RF pose models. They both identify disengaged states, but RF gives higher confidence.

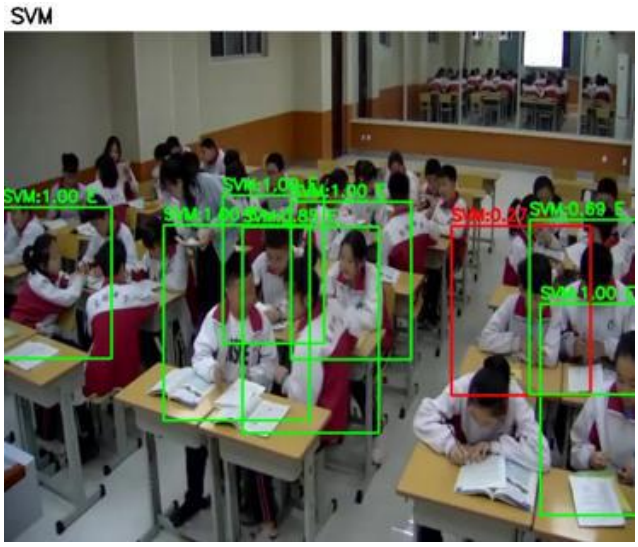

Fig.10: SVM Pose Model

Fig.11: RF Pose Model

IV. DISCUSSION

The experimental results show the clear advantage of multimodal feature fusion in engagement recognition. The Fusion (MLP) model achieved the best overall performance with an accuracy of 0.88, precision of 0.87, recall of 0.86, F1-Score of 0.86, and AUC of 0.91. This confirms that combining the eye-gaze and posture information will produce better behavioral outputs than using either modality alone. MLP's nonlinear feature learning allows it to get the subtle dependencies between the models. This includes correlation between gaze direction and posture inclination which is hard for SVM models to learn because they are linear.

In contrast, the Fusion-SVM (AUC = 0.88) and Fusion-RF (AUC = 0.89**)** still performed alright but is sensitive to

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.6.3

when the features are imbalance and noisy. The SVM sometimes misclassified ambiguous cases like students resting heads on their hands but still being attentive, while the Random Forest tended to overfit to high-dimensional feature vectors. With that being said, both produced strong generalization across engagement categories, confirming the stability of the proposed fusion design.

When comparing unimodal results, gaze-based and posture-based classifiers each showed strengths in different contexts. Gaze models achieved high AUC values (up to 0.93 for Gaze-RF), indicating strong sensitivity to attention direction, while posture models excelled at identifying disengaged behaviors like slouching or turning away (AUC = 0.93 for Posture-RF). However, unimodal performance degraded sharply when occlusions affected visibility. The fusion pipeline-maintained accuracy across all conditions which greatly demonstrates robustness to visual obstruction, illumination variation or light change, and pose changing.

Qualitative visualization further strengthens the observations. The fused models always produced high confidence bounding boxes and stable engagement predictions even though the faces were partially hidden or body orientation is different. This robustness shows why we need to integrate posture and gaze for good cues for engagement inference. Each feature compensates for the other's weaknesses. Reliability like that is essential for real world classroom deployment where visibility is unpredictable and student positioning constantly changes.

The system's modular and lightweight architecture (YOLOv11 + MoveNet + FaceMesh + MLP) ensures computational efficiency that allows us to do deployment on edge devices such as webcams or embedded cameras. The scalability shown supports potential classroom applications like real-time dashboards for lecturers, student attention analytics, or adaptive teaching feedback systems. Future development could extend the model temporally by incorporating LSTM or transformer layers to capture sequential engagement trends and transitions across lecture sessions.

V. CONCLUSION

This study presents a multimodal, obstruction-resilient student engagement detection framework that unifies gaze estimation and body posture recognition for classroom analytics. The system leverages YOLOv11 for student detection, Movenet for skeletal posture estimation, and MediaPipe FaceMesh for eye-gaze direction inference. These complentary features are fused through machine learning classifiers such as MLP, SVM and Random Forest to infer engagement states. This pipeline replaces a computationaly heavy CNN [7] and STGCN [8] architectures with efficient, real-time models suitable for classroom deployment on edge devices.

The experimental findings show that fusion significantly improves engagement detection performance compared to unimodal models, especially under obstructed conditions. The Fusion-MLP model proved most effective which balances accuracy and computational efficiency suitable for real-time deployment. The framework's robustness under occlusion makes it particularly relevant for modern educational settings affected by mask usage and varying cultural attire.

Beyond technical performance, this work carries broader implications for learning analytics and educational inclusivity. It offers educators an unobtrusive tool to monitor engagement, evaluate teaching effectiveness [12], and adapt instructional delivery dynamically. Nevertheless, the model currently operates on static frames and binary engagement labels; future work should extend to temporal engagement modelling, multi-class affect recognition, and feedback-based systems where students can self-report or validate engagement scores.

Overall, the proposed system represents a significant step toward intelligent, data-driven education, bridging the gap between behavioural observation and automated analytics while promoting fairer, more adaptive learning environments.

ACKNOWLEDGEMENTS

I would like to express sincere gratitude to Dr. Maged M. Saeed Nasser, Supervisor, Department of Computing, Universiti Teknologi PETRONAS, for his continuous guidance, technical expertise, and valuable feedback throughout the development of this project.

I am grateful to the developers and maintainers of the SCB-Dataset for making their benchmark on student and teacher classroom behavior publicly available, which served as the foundation for model training and evaluation in this project.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.6.3

REFERENCES

[1] F. J. A., B. P. C. and P. A. H., "School Engagement: Potential of the Concept, State of the Evidence," *Review of Educational Research, vol. 74 (1), pp.,* p. 59–109, 2022.

[2] D. R. and D. S., "Optimizing Student Engagement Detection Using Facial and Behavioral Features," *Neural Computing and Applications,* 2025.

[3] U. Team, "YOLOv11 Object Detection Model," Ultralytics Docs, 2024. [Online]. Available: https://docs.ultralytics.com/models/yolov11.

[4] G. R. Team, "MoveNet: Ultra-Fast Human Pose Estimation Model," TensorFlow Hub Tutorials – Google Research, 2023. [Online]. Available: https://www.tensorflow.org/hub/tutorials/movenet.

[5] L. C. e. al., "MediaPipe FaceMesh: High-Fidelity 3-D Face Landmarks," in *Proceedings of CV4Educ. / Google Developers*, 2023.

[6] WintonYF, *SCB-Dataset: Student and Classroom Behavior Dataset,* Hugging Face Datasets, 2023.

[7] Z. X., S. Y., F. M. and B. A., *MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation,* arXiv preprint arXiv:1711.09017, 2017.

[8] Y. S., X. Y. and L. D., *Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,* arXiv preprint arXiv:1801.07455, 2018.

[9] M. S., N. S. and B. A., "MMSAD: Multi-Modal Student Attentiveness Detection in Classrooms," *Journal of Educational Technology,* 2025.

[10] H. A. e. al., "Efficient Edge AI Models for Human Sensing," *IEEE Access,* 2024.

[11] G. I., B. Y. and C. A., Deep Learning (3rd ed.), MIT Press, 2023.

[12] S. A. e. al., "VisioPhysioENet: Multimodal Engagement Detection Using Visual and Physiological Signals," in *OpenReview*, 2024.

[13] P. H. A., "Engagement Detection and Enhancement for STEM Learning Environments," *Image and Vision Computing,* 2023.

[14] X. N. e. al., "Detecting Student Engagement in Online Learning Using Computer Vision and Multi-Dimensional Feature Fusion," *ACM Digital Library,* 2023.

Cross Ref DOI: https://dx.doi.org/10.22161/ijtle.4.6.3