EngageSense: A Hybrid Approach for Real Time Engagement Detection for Virtual Classrooms

1st Muhammad Irfan
School of Computing and Digital Media
London Metropolitan University
166-220 Holloway Road
London, UK
mui0316@my.londonmet.ac.uk

2nd Preeti Patel
School of Computing and Digital Media
London Metropolitan University
166-220 Holloway Road
London, UK
p.patel@londonmet.ac.uk

3rd Bilal Hassan
School of Computing and Digital Media
London Metropolitan University
166-220 Holloway Road
London, UK
b.hassan@londonmet.ac.uk

Abstract—Advancements in digital education have revolutionized traditional learning environments, driving the widespread adoption of virtual and hybrid classrooms. Engagement, a vital factor for effective learning, necessitates continuous monitoring and assessment to optimize outcomes. This study introduces EngageSense, a hybrid real-time engagement detection system leveraging facial biometrics, computer vision, and deep learning. First, a new dataset is created via user eye images taken from webcam of laptop. Then, Dlib's HOG + Linear SVM for face detection, a CNN model trained on 4,453 eye images dataset(classified into left, right, and center gaze directions), and OpenPose MobileNetV1 for body pose estimation are used. By fusing gaze direction (99.50% accuracy) and pose features, EngageSense classifies engagement into three levels: fully engaged, partially engaged, and not engaged with an accuracy of 90%. By providing actionable real-time insights, EngageSense empowers educators to foster meaningful interactions and enhance learning experiences in virtual environments.

Index Terms—Student Engagement, Real-Time Monitoring, Virtual Classroom, Online Student Monitoring, EngageSense

I. Introduction

In recent years, higher education and corporate training sectors have undergone significant transformations, leading to the adoption of virtual and hybrid classroom environments. While these shifts offer flexibility and accessibility, they also pose challenges, particularly in monitoring and sustaining learner engagement, which is closely linked to academic performance and learning outcomes [1]. The absence of physical presence in virtual classrooms makes it challenging for educators to effectively gauge student engagement in real time.

Research has demonstrated that learners are more engaged in physical classrooms [2], where instructors can directly observe behaviors and interactions. In virtual settings, however, maintaining and detecting engagement is more complex due to the lack of direct physical interaction. Studies have explored various machine learning (ML) approaches to address this issue. For instance, [3] highlights the importance of realtime data in reliably detecting student engagement and its correlation with academic success. Similarly, [4] shows that facial expressions, head poses, and gaze movements can serve as indicators of engagement, using long short-term memory (LSTM) networks for analysis. Traditional techniques for assessing engagement, such as interactive discussions, quizzes, and feedback reviews, are less effective in virtual settings, where disengagement often occurs. Advanced methods, including eye movement tracking, body posture analysis, and facial expression detection, have shown promise in virtual classrooms [5]. Translating these methods to virtual environments necessitates real-time engagement detection systems capable of accurately analyzing behaviors and focus.

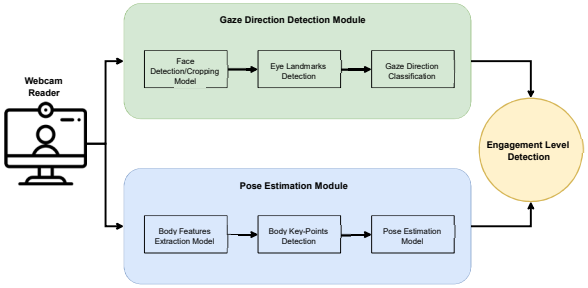


Fig. 1. Schematic Overview of EngageSense

Numerous studies have explored student engagement detection in educational settings, making use of computer vision to automate what educators traditionally assess visually. The complexity and occlusion in a real classroom environment make it challenging to accurately detect each student, even with high-definition cameras. Relying on a single modality to assess learning engagement is insufficient in such conditions. [6] explored emotion recognition and its relationship to engagement but noted that relying solely on emotion detection is insufficient for comprehensive engagement analysis. [7] explored classifying engagement by training Attention-Net for head pose estimation and Affect-Net for facial expression recognition using facial videos and found out that in a virtual classroom, the unobtrusive learning engagement can be effectively recognized using multiple nonverbal cues, such as facial expressions, hand gestures, and

body postures. As a result, the current trend focuses on recognizing and analyzing student engagement using multiple nonverbal cues.

This paper aims to address the challenges of single modality and computationally ineffective system by proposing a new system that prioritizes multiple nonverbal cues to evaluate engagement levels in virtual classrooms. Figure 1 gives the schematic overview of the proposed system which provides a comprehensive, hybrid and real-time solution. Our primary contributions are:

- Introducing a system capable of detecting learner engagement levels in virtual environments in real time.
- Designing a system using Lightweight models that are computationally efficient models to ensure suitability for real-time processing.
- Developing a CNN based eye gaze detection model using custom dataset with an accuracy of 99.50%.
- Implementing a hybrid model to classify learner engagement into three categories: fully engaged, partially engaged, and not engaged with an accuracy above 90%.

Research highlights the importance of combining multiple behavioural data points. [8] proposed a cost-effective framework using a webcam and a CNN to detect engagement levels in real-time. Similarly, [9] introduced a Mean Engagement Score (MES) derived from facial emotion detection using CNN. [10] emphasized the integration of emotional and cognitive components, utilizing OpenPose and deep learning for pose classification. [11] enhanced detection accuracy by 3.9% using a hybrid method with ResNet and Temporal Convolutional Networks (TCN). Similarly, [5] fused affective and attention features, achieving an Under the Curve (AUC) of 0.720 for engagement classification using trained Attention-Net and Affect-Net. [12] utilized webcam data with a haar-cascade algorithm and CNN, correlating high engagement with better academic performance. [13] developed a lightweight FER model combining multitask CNN and EfficientNet, but noted challenges with low-resolution images, limiting detection effectiveness. Instructor behavior and presentation style also influence engagement. [3] achieved 92.23% accuracy using the CATBoost model, surpassing prior approaches. [14] developed a system combining facial recognition, head pose estimation, and Eye Aspect Ratios, achieving 72.4% engagement classification accuracy. Similarly, [15] proposed a multi-cue vision-based approach estimating engagement through affect, attention, and head movement, achieving a 75-80% correlation with self-reports. The integration of IoT devices has introduced novel approaches for monitoring engagement. [16] proposed an IoT-based framework using wireless emotional signals and LSTM models, achieving 95% accuracy and ultra-low latency. However, challenges remain in mitigating signal interference and environmental variability. These studies underscore the critical role of computer vision, IoT, and deep learning in addressing the complexities of engagement detection in virtual and hybrid classrooms. Table I list down past references with the details of datasets used, methodology applied and the results achieved.

[23] propose a aggregated deep CNN model ApparelNet, which is designed for person verification in border control environment. It provides single-based-image detection using OpenPose model and also evaluated on Front-View-Gait(FVG) dataset with shows an accuracy of 98% on training and validation. Model verification achieve 96% prediction accuracy when tested on 12 randomly selected individuals. [24] investigates the selection of soft-biometrics and their relations for efficient verificaion. They developed a multi non-linear regression based framework named as RSFS, for selection of higly supportive soft-biometric features using very large collection of soft-biometrics. [25] proposed a OneDetect, a federated learning architecture that uses intrusive features of human body to detect three most common global soft biometrics features gender, age and ethicity. [26] study shows that invention of Motion Sensing Camera opened new domain of research and people are using them for recognition of human activities. Mircosoft Kinect for XBOX provides two types of information RGB and Depth and many vendors are developing tools that works with Microsoft Kinect.

In Section 2, we present the EngageSense framework, detailing the processes for face detection and the extraction of

TABLE I

LIST OF PAST REFERENCES WITH THEIR DATASETS, METHODOLOGY, AND RESULTS

Ref	Year	Dataset	Methodology	Results
[17]	2024	100 images to 1000 images from UPNA Head Database	OpenCV, CNN	Obtained 99% accuracy of CNN based face recognition model.
[18]	2024	FER 2013	Haar Cascade Algorithms, MobileNetV2, CNN	Obtain 73.4% accuracy of emotion recognition model
[19]	2023	Used own Face Emotion Dataset, CK+, FER-2013, RAF-DB	ResNet-50, VGG-19, Inception-V3	Achieved 92.3% accuracy of ResNet-50 on own Face Emotion Dataset used for engagement detection.
[20]	2023	Uses own Face Dataset	Dlib, VGG-19	Achieved an accuracy of 95.25% for two engagement states (Focused or Fatigued).
[21]	2022	Kaggle Eye-Image Dataset (Around 14500 eye images was used for real time detection)	Haar Cascade Algorithms, CNN, MobileNet	Achieved an accuracy of 99% of eye gaze detection for two engagement states (Engaged or Not Engaged).
[22]	2022	ClassX, LectureVideoDB, IIIT-AR-13K	RetinaNet, VGG-16, Kmeans, TCN	Obtain presentation style model accuracy 86% and student engagement model accuracy 76%.

Sensing Camera opened new domain of research and people are using them for recognition of human activities. Mircosoft Kinect for XBOX provides two types of information RGB and Depth and many vendors are developing tools that works with Microsoft Kinect.

In Section 2, we present the EngageSense framework, detailing the processes for face detection and the extraction of

body features for eye gaze direction and pose estimation. Section 3 focuses on the experimental outcomes and their analysis across the various modules of EngageSense. Finally, Section 4 concludes the study and highlights potential directions for future research and further investigation.

II. Experimental Design

This section outlines the EngageSense framework for realtime engagement detection in virtual classrooms, detailing its implementation steps: face detection, eye gaze tracking, pose estimation, and engagement classification. The integration of deep learning (DL) and computer vision (CV) techniques ensures system accuracy and efficiency within a virtual learning environment.

A. EngageSense Framework

The EngageSense framework, illustrated in Figure 2, combines DL and CV methodologies. For face and eye detection, the system utilizes Dlib's library, known for its accuracy and efficiency. Eye gaze direction is determined using a pretrained Convolutional Neural Network (CNN), enabling real-time tracking of gaze movements, which serves as a key indicator of visual engagement. Pose estimation is performed using the OpenPose MobileNet V1 pretrained model, which identifies 19 key body points, including the head, shoulders, and neck, to track posture with precision. Engagement levels are classified by aggregating the outputs from these modules over a one-second duration of real-time video captured via a webcam. This integrated approach, combining facial, gaze, and pose analysis, provides a robust and accurate solution for detecting learner engagement in real-time.

B. Dataset Preparation

The dataset used in this research comprises images collected via a webcam, focusing on the eye region for further analysis. Each participant contributed ten images, which were systematically cropped to extract the eye region. To enhance the dataset and improve model generalization, various augmentation techniques were applied, including rotation and quality adjustments. Furthermore, the images were converted to grayscale and resized to a uniform dimension of 56x64 pixels to ensure consistency in the dataset. As a result of these preprocessing steps, the final augmented dataset consists of 4,453 images, providing a robust foundation for model training. The dataset preparation flow is shown in Figure 3

C. Face Detection

Dlib is a widely utilized Python library for face recognition, known for its accuracy and speed compared to alternatives like OpenCV's HAAR Cascade [21]. The library offers two face detection methods: the Histogram of Oriented Gradients (HOG) + Linear Support Vector Machine (SVM) face detector [5] and a deep learning-based MMOD CNN face detector. For this study, the HOG + Linear SVM detector was selected due to its ease of integration with Dlib's shape predictor. Table II highlights the comparision of performance between two Dlib face detection models.

HOG is a feature descriptor designed to extract critical information from images, making it essential in computer vision tasks such as object or shape detection. It focuses on capturing the structural features and shapes within an image by analyzing edge gradients and orientations. HOG's ability to detect edge directions contributes to its effectiveness in face detection. After extracting the HOG feature vector, passed to a Linear SVM, which identifies the face and provides a bounding box around the detected face. Figure 4 illustrates the face detection process using Dlib's frontal face detector based on the HOG + Linear SVM method.

TABLE II
Performance Comparision of Face Detection Models

Performance Feature	HOG+Linear SVM	MMOD CNN
Accuracy	85%-95%	96%-98%
Computational Efficiency	Provide 25-35 FPS on a standard CPU without requiring a GPU	Provides 10-15 FPS on CPU and 30-50 FPS on GPU
Model Size	<100 MB	>100 MB to GB
Training Efficiency	Easily train on small dataset with less training time	Requires large dataset and more training time

TABLE III
Eye Gaze Detection CNN Model Summary

Layers	Output Shape	Para #
conv2d 54 (Conv2D)	(None,56,64,16)	160
max _pooling2d 54 (MaxPooling2D)	(None,28,32,16)	0
conv2d 55 (Conv2D)	(None,28,32,32)	4640
max _pooling2d 55 (MaxPooling2D)	(None,14,16,32)	0
conv2d 56 (Conv2D)	(None,14,16,64)	18,496
max _pooling2d 56 (MaxPooling2D)	(None,7,8,64)	0
flatten 18 (Flatten)	(None, 3584)	0
dense _35 (Dense)	(None, 256)	917,760

activation 35 (Activation)	(None, 256)	0
dense _36 (Dense)	(None, 3)	771
activation 36 (Activation)	(None, 3)	0

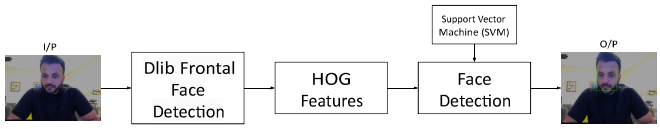


Fig. 4. HOG and Linear SVM based Face Detection

D. Eye Gaze Detection and Classification

Various methods are available for detecting facial features to track gaze movement. In this study, a gaze tracking system was developed using Dlib's facial landmarks detector for eye landmark detection and a trained CNN model for classifying eye states as looking center, left, or right in real time. Summary of each layer of CNN model and its hyperparameters are shown in Table III and Table IV.

Dlib's facial landmarks detector, pretrained within the library, predicts 68 landmarks corresponding to facial structures. For gaze detection, the right and left eyes are extracted based on the facial landmark indices: [42–47] for the left eye and [36–41] for the right eye. These extracted eye regions are then fed into the pretrained CNN model, which classifies the gaze direction as looking center, right, or left, as shown in Figure 5.

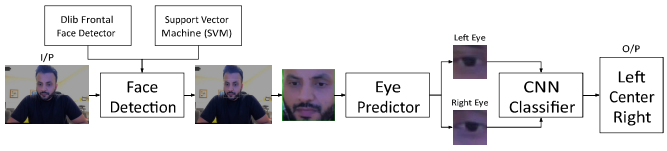


Fig. 5. Eye Landmarks and CNN based Eye Gaze Detection

TABLE IV
LIST OF HYPERPARAMETERS OF CNN MODEL

Hyperparameter	Value
Batch Size	32
Optimizer	Adam
Activation Function	ReLU, Softmax
Filters	16,32,64
Kernal Size	(3,3)
Pool Size	(2,2)
No. of training Epochs	10

E. Pose Estimation Using OpenPose

Human pose estimation involves identifying the positions of key points on the human body, effectively enabling computers to recognize and analyze human movements. OpenPose is a real-time pose estimation system that uses a skeleton-based model to map body movements. To enhance its efficiency, pretrained models are available, and for this study, we utilized

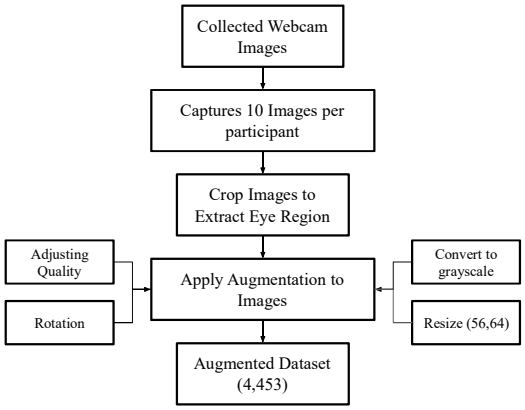


Fig. 3. Dataset Preparation and Augmentation

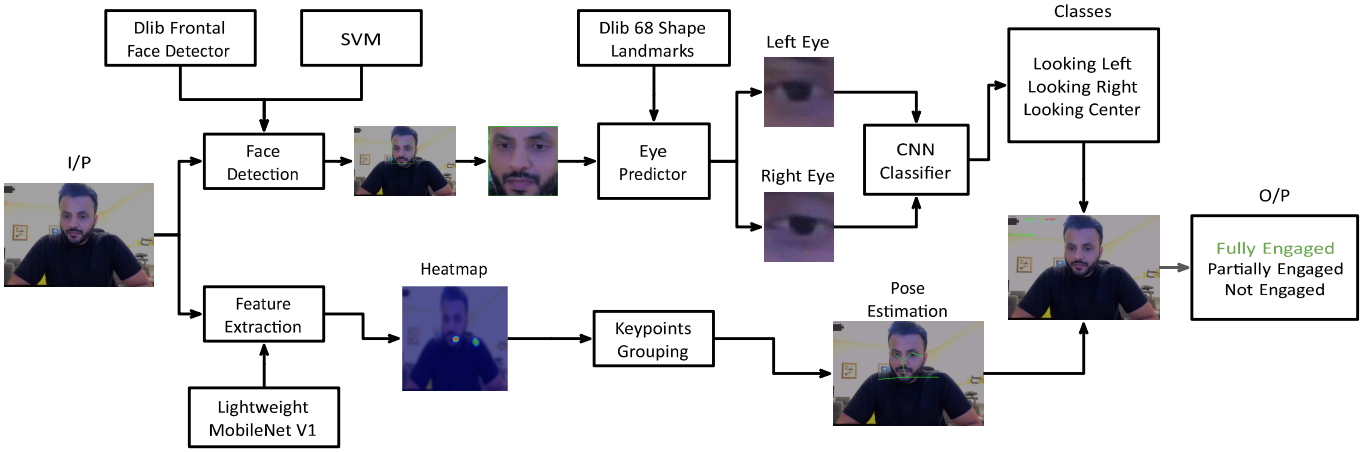


Fig. 2. EnagageSense Framework

the MobileNetV1 model trained on the COCO (Common Objects in Context) dataset. Table V provides comparision of MobileNetV1 with InceptionV3 and ResNet50 models based on number of parameters and size. It shows that MobileNetV1 is more efficient in terms of parameters and size as compared to other pretrained models. MobileNetV1 is 18 times and 44 times smaller than InceptionV3 and ResNet50. Due to having low weights, MobileNetV1 also provide easily integration on low resources system also on android and IOS after its conversion in Tensorflow Lite version [23].

TABLE V
MOBILENET SIZE AND PARAMETERS COMPARISION WITH OTHER PRETRAINED MODELS

Model Name	Total Parameters	Model Size
MobileNetV1	4.2 millions	16 MBs
InceptionV3	74.2 millions	717 MBs
ResNet50	26.2 millions	314 MBs

The process, presented in Figure 6, begins by inputting an RGB image into MobileNetV1 to extract body features. A heatmap is then generated from these features to improve the localization of key points. The required key points are grouped from the heatmap, and OpenCV is used to map the skeleton model of these grouped key points onto the input image, providing a visual representation of the detected human pose.

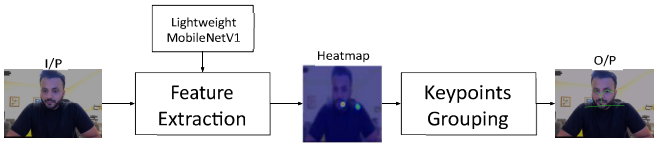


Fig. 6. Human Pose Estimation Using OpenPose

F. Engagement Level Classification

The process begins with pose estimation using OpenPose, which provides coordinates for various body parts such as the ears, eyes, and shoulders. If any of these parts are not detected, a counter for the missing part is incremented. If more than two parts are consistently undetected during pose estimation, the pose is marked as not detected. Simultaneously, eye movements are tracked. If the gaze is centered, it is marked as detected; otherwise, it is not detected. The engagement level is then predicted based on the combined results of pose estimation and eye gaze detection as shown in Equation 3.

The Gaze is determined as:

$$classes = [center, left, right]$$

$$Gaze = \begin{cases} 1, & \text{if } argmax(p_{eye}) = center \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $p_{eye} = [p_{center}, p_{left}, p_{right}]$ represents the probability distribution over gaze directions. The function selects the class with the highest probability.

The Pose is calculated as:

$$Features = [l_{shoulder}, r_{shoulder}, neck, r_{ear}, l_{ear}]$$

$$Pose = \begin{cases} 1, & \text{if at least two features are detected} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Classification of Engagement Level:

$$classes = [Fully_{Engaged}, Partially_{Engaged}, Not_{Engaged}]$$

Getting Pose and Gaze value from Equation 2 and 1, we will calculate V_1 and V_2

$$V_1 = min(Pose, Gaze)$$

$$V_2 = max(Pose, Gaze)\ class_{index} = len(classes - 1) - V_1 - V_2$$

$$Engagement_{level} = classes[class_{index}] \quad (3)$$

III. OUTCOMES AND ANALYSIS

This section presents the experimental results and analysis validating the effectiveness of the EngageSense system. The findings show that combining pose estimation and gaze detection enhances engagement classification significantly, emphasizing the importance of integrating multiple features for real-time engagement detection.

A. Outcomes of Eye Gaze Detection

The eye gaze detection model effectively identifies gaze directions (center, left, right), playing a crucial role in engagement level classification. Figure 7 illustrates the performance of the proposed model with gaze directions labelled as ”centre,” ”right,” or ”left.” While the model demonstrates high accuracy in most cases, misclassification is observed in Test Image (2), where the ground truth direction is ”right,” but the system predicts ”center.” This error is likely due to limited training data or insufficient feature extraction.

Test_Image (3) Test_Image (4)

Fig. 7. Detection of Eye Gaze on different participants

Figure 8 depicts the training process of the eye gaze detection model over 10 epochs, showing a steady increase in accuracy, ultimately reaching 99.50% on training data and 97.50% on validation data, while the loss decreases to 0.0162 on training data and 0.0533 on validation data; highlighting the model's effective learning, with consistent improvements observed in both training and validation metrics.

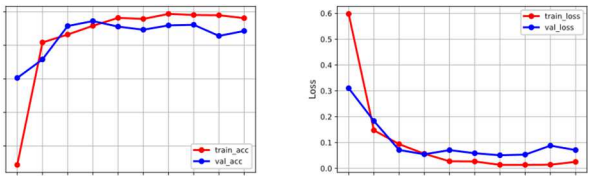


Fig. 8. Accuracy and Loss of Eye Gaze Detection Model

Table VI provides the classification report, showing the performance of the eye gaze detection model across three classes: "centre," "left," and "right." Figure 9 demonstrate confusion matrix that shows the model's strong overall accuracy, with minimal misclassifications. Figure 10 shows the curve closer to the top-left corner of the plot indicates better performance of eye gaze detection model on each class.

TABLE VI
CLASSIFICATION REPORT OF EYE GAZE DETECTION MODEL

Gaze Direction	Precision(%)	Recall(%)	F1-Score(%)	Samples
center	0.97	1.00	0.98	243
left	1.00	1.00	1.00	273
right	1.00	0.97	0.99	315

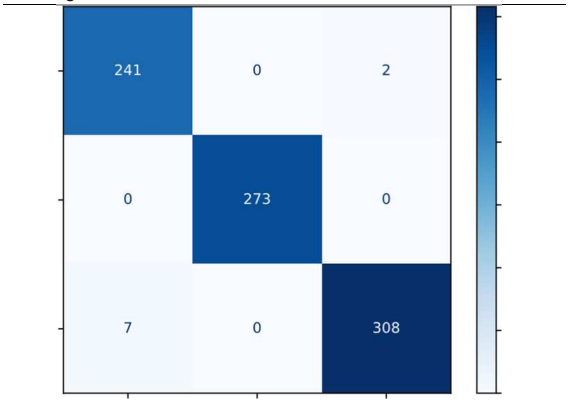


Fig. 9. Confusion Matrix on test data of Eye Gaze Detection Model

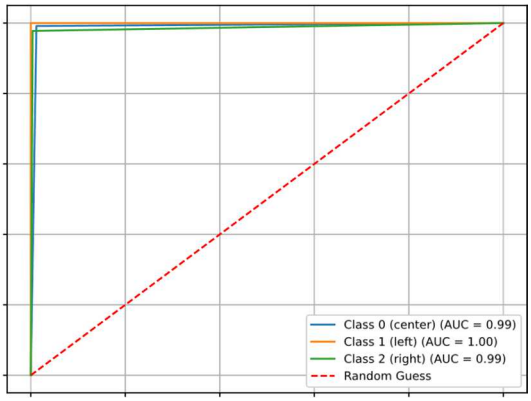


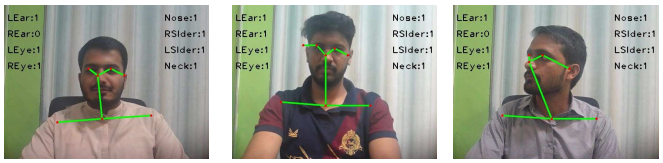
Fig. 10. ROC Curve of 3 Classes (center,left,right) of Eye Gaze Detection Model

B. Outcomes of Pose Estimation (OpenPose)

Figure 11 illustrates the results of pose estimation for three participants during the engagement detection process. The model successfully identifies key landmarks, including the nose, eyes, ears, shoulders, and neck, and maps them with corresponding feature counts.

- Participant 1: The pose estimation captures a frontal position with all key landmarks visible, except for the right ear.
- Participant 2: A slightly tilted pose is observed, indicated by asymmetry in the left and right-side landmark counts.
- Participant 3: A side view is depicted, where the algorithm accurately detects visible landmarks while marking occluded ones (e.g., the right ear) as absent.

These results confirm the robustness of the pose estimation model in handling different head orientations and positions, which are crucial for determining engagement levels.



Participant (1) Participant (2) Participant (3)

Fig. 11. Pose key points estimation on different participants

C. Outcomes of Hybrid Model (Eye Gaze Detection and OpenPose)

The hybrid model combines pose estimation and gaze detection to classify learner engagement into three categories: Fully Engaged, Partially Engaged, and Not Engaged. This integration enhances the model's ability to assess engagement levels by capturing details of learner behaviour. Figure 12 illustrates the classification outcomes for test images.

The hybrid model effectively detects partial engagement when a learner's gaze or head orientation deviates. For example, in Figure 12, Test _Image (2), the learner's gaze shifts to the right, leading the model to classify the state as "Partially Engaged." Similarly, in Test Image (6), where the head is turned to the left, the model also identifies "Partially Engaged." In Test Images (1) and (3), where the learner faces forward with a direct gaze, the model accurately classifies the state as "Fully Engaged," reflecting complete focus. Conversely, in Test Image (5), where the head is tilted downward, the model correctly identifies the posture as "Not Engaged," indicating disengagement.

Almost 50 samples are taken on real time of 10 participants to see how EngageSense classification accuracy. Figure 13 shows confusion matrix which shows that EngageSense classify 90% samples accurately. These results highlight the hybrid model's robustness in capturing both pose and gazebased cues to determine engagement levels.

D. Comparative Analysis of Related Works

We present, in Table VII, a comparative analysis of engagement detection approaches, focusing on the key components

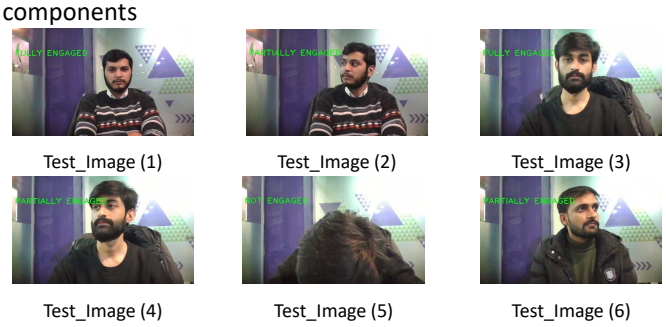


Fig. 12. Results of Hybrid Model (using both Pose and Gaze Detection)

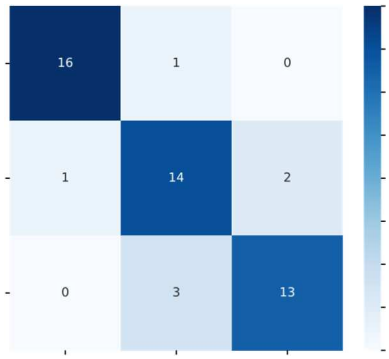


Fig. 13. Results of Hybrid Model (using both Pose and Gaze Detection)

of face detection, pose estimation, and eye gaze detection accuracy.

The proposed EngageSense method integrates deep learning and computer vision to combine face detection, pose estimation, and eye gaze detection. This comprehensive approach achieves a notable improvement, with an eye gaze detection accuracy of 99.50% and engagement level detection accuracy of around 90%. The enhanced accuracy highlights the collective effect of integrating pose estimation and gaze detection, significantly improving the model's robustness in capturing engagement level across diverse scenarios. Additionally, the use of advanced pretrained models and the integration of multiple features ensures that EngageSense offers a more reliable and accurate prediction of engagement compared to the other methods reviewed.

TABLE VII

COMPARISION OF PAST PAPER FEATURES SELECTION AND ACCURACY

Ref	Model Used	Face Detection	Pose Estimation	Eye Gaze Detection (Accuracy)
[24]	Deep Learning	Yes	No	70%
[25]	Haar Cascade	Yes	No	95.25%
[17]	Deep Learning	Yes	Yes	No
[18]	Deep Learning + Computer Vision	Yes	Yes	73.4%
Engage Sense	Deep Learning + Computer Vision	Yes	Yes	99.50%

IV. CONCLUSION AND FUTURE WORK

This paper presents EngageSense, a real-time learner engagement detection system for virtual classrooms, utilising computer vision and deep learning techniques. The system utilizes Dlib for face detection, CNN for gaze classification, and OpenPose for pose estimation to classify engagement into three levels: fully engaged, partially engaged, and not engaged. Experimental results demonstrate that combining eye gaze detection with an accuracy level of 99.50

A limitation of this paper is the small dataset, as only ten participants agreed to provide data due to privacy concerns. Consequently, the eye gaze dataset is customized based on this limited number of participants. In future work, we aim to expand the dataset by increasing the number and diversity of participants. During the training of the eye gaze detection model, the small dataset led to overfitting and underfitting when a larger number of epochs were selected. This shortcoming can be addressed by increasing the size of the dataset and applying regularization techniques. Future work could explore these strategies to enhance the model's robustness and applicability. Another limitation is the challenge of face detection in lowlight conditions and the presence of occlusions. While there is no perfect solution for these issues, they can be mitigated by training the model on a dataset with varying light conditions or by using an occlusion-aware CNN model.

Another important aspect is that eye and pose cues may not always reflect the participant's mental and physiological state. Future work will focus on enhancing engagement analysis by incorporating additional behavioural cues, physiological signals and hand-gestures . Real-time feedback mechanisms will be developed to help educators address disengagement promptly. Future work is aimed at integrating a lightweight engagement system based on MobileNetV1 on IoT devices and wearables. Sensor data such as hand movements, body movements or pulse rate can be used to detect engagement levels. However, the proposed system EngageSense has so far demonstrated reliable results in face detection, eye gaze detection, pose estimation, and engagement detection. Therefore, this study has the potential to provide support to the development of innovative educational technologies.

References

[1] H. P. Bui, "L2 teachers' strategies and students' engagement in virtual classrooms: a multidimensional perspective," in Micro-Electronics and Telecommunication Engineering: Proceedings of 6th ICMETE 2022. Springer, 2023, pp. 205–213.

[2] S. Fabriz, J. Mendzheritskaya, and S. Stehle, "Impact of synchronous and asynchronous settings of online teaching and learning in higher education on students' learning experience during covid-19," Frontiers in psychology, vol. 12, p. 733554, 2021.

[3] N. Alruwais and M. Zakariah, "Student-engagement detection in classroom using machine learning algorithm," Electronics, vol. 12, no. 3, p. 731, 2023.

[4] P. Buono, B. De Carolis, F. D'Errico, N. Macchiarulo, and G. Palestra, "Assessing student engagement from facial behavior in on-line learning," Multimedia Tools and Applications, vol. 82, no. 9, pp. 12 859–12 877, 2023.

[5] Z. A. T. Ahmed and M. E. Jadhav, "A review of early detection of autism based on eye-tracking and sensing technology," in 2020 International Conference on Inventive Computation Technologies (ICICT). IEEE, 2020, pp. 160–166.

[6] M. N. Hasnine, H. T. T. Bui, T. T. Thu Tran, H. T. Nguyen, G. Akc¸apınar, and H. Ueda, "Students' emotion extraction and visualization for engagement detection in online learning," Procedia Computer Science, vol. 192, pp. 3423–3431, Jan. 2021, doi: https://doi.org/10.1016/j.procs.2021.09.115.

[7] O. Sumer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein, and E. Kasneci, "Multimodal Engagement Analysis from Facial Videos in the Classroom," IEEE Transactions on Affective Computing, pp. 1–1, 2021, doi: https://doi.org/10.1109/taffc.2021.3127692.

[8] Y. Wang, A. Kotha, P. Hong, and M. Qiu, "Automated Student Engagement Monitoring and Evaluation during Learning in the Wild," IEEE Xplore, Aug. 01, 2020. https://ieeexplore.ieee.org/abstract/document/9170958 (accessed May 07, 2023).

[9] P. Bhardwaj, P. K. Gupta, H. Panwar, M. K. Siddiqui, R. Morales-Menendez, and A. Bhaik, "Application of Deep Learning on Student Engagement in e-learning environments," Computers Electrical Engineering, vol. 93, p. 107277, Jul. 2021, doi: https://doi.org/10.1016/j.compeleceng.2021.107277.

[10] P. Vanneste et al., "Computer Vision and Human Behaviour, Emotion and Cognition Detection: A Use Case on Student Engagement," Mathematics, vol. 9, no. 3, p. 287, Feb. 2021, doi: https://doi.org/10.3390/math9030287.

[11] A. Abedi and S. S. Khan, "Improving state-of-the-art in Detecting Student Engagement with Resnet and TCN Hybrid Network," May 2021, doi: https://doi.org/10.1109/crv52889.2021.00028.

[12] P. Sharma et al., "Student Engagement Detection Using Emotion Analysis, Eye Tracking and Head Movement with Machine Learning," Communications in Computer and Information Science, pp. 52–68, 2022, doi: https://doi.org/10.1007/978-3-031-22918-3 5.

[13] "Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network — IEEE Journals Magazine — IEEE Xplore," ieeexplore.ieee.org. https://ieeexplore.ieee.org/abstract/document/9815154

[14] M. U. Uc¸ar and E. Ozdemir,¨ "Recognizing Students and Detecting Student Engagement with Real-Time Image Processing," Electronics, vol. 11, no. 9, p. 1500, May 2022, doi: https://doi.org/10.3390/electronics11091500.

[15] Chakradhar Pabba and P. Kumar, "A vision-based multi-cues approach for individual students' and overall class engagement monitoring in smart classroom environments," Multimedia Tools and Applications, Nov. 2023, doi: https://doi.org/10.1007/s11042-023-17533-w.

[16] M. Awais et al., "LSTM-Based Emotion Detection Using Physiological Signals: IoT Framework for Healthcare and Distance Learning in COVID-19," IEEE Internet of Things Journal, vol. 8, no. 23, pp. 16863–16871, Dec. 2021, doi: https://doi.org/10.1109/JIOT.2020.3044031.

[17] Nuha Alruwais and M. Zakariah, "Student Recognition and Activity Monitoring in E-Classes Using Deep Learning in Higher Education," IEEE access, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3354981.

[18] A. Sukumaran and A. Manoharan, "Multimodal Engagement Recognition From Image Traits Using Deep Learning Techniques," IEEE Access, vol. 12, pp. 25228–25244, 2024, doi: https://doi.org/10.1109/access.2024.3353053.

[19] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," Multimedia Tools and Applications, Sep. 2022, doi: https://doi.org/10.1007/s11042-022-13558-9.

[20] Rasheed Abdulkader et al., "Optimizing student engagement in edge-based online learning with advanced analytics," vol. 19, pp. 100301–100301, Sep. 2023, doi: https://doi.org/10.1016/j.array.2023.100301.

[21] Ahmed, Z.A., Jadhav, M.E., Al-madani, A.M., Tawfik, M., Alsubari, S.N. and Shareef, A.A.A., 2022. Real-Time Detection of Student Engagement: Deep Learning-Based System. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 1 (pp. 313-323). Springer Singapore.

[22] C. Thomas, K. A. V. Puneeth Sarma, S. Swaroop Gajula, and D. B. Jayagopi, "Automatic prediction of presentation style and student engagement from videos," Computers and Education: Artificial Intelligence, vol. 3, p. 100079, 2022, doi: https://doi.org/10.1016/j.caeai.2022.100079.

[23] B. Hassan and E. Izquierdo, "ApparelNet: Person Verification Encompassing Auxiliary Attachments Variation," pp. 1–6, Oct. 2021, doi: https://doi.org/10.1109/mmsp53017.2021.9733523.

[24] B. Hassan and E. Izquierdo, "Rsfs: A soft biometrics-based relative support features set for person verification," in Fourteenth International Conference on Digital Image Processing (ICDIP 2022), vol. 12342. SPIE, 2022, p. 1234202.

[25] B. Hassan and E. Izquierdo, "OneDetect: A Federated Learning Architecture for Global Soft Biometrics Prediction," May 2022, doi: https://doi.org/10.1109/iscv54655.2022.9806101.

[26] M. Shoaib Farooq, B. Hassan, M. Naseer, A. Abid, Y. D. Khan, N. S. Khan, M. Usman Akram, S. Ullah et al., "Studio applications and software development kits for microsoft kinect: A survey," Journal of Applied Environmental and Biological Sciences, vol. 4, pp. 398–402, 2014.

[27] H. Zia et al., "Plastic Waste Management through the Development of a Low Cost and Light Weight Deep Learning Based Reverse Vending Machine," Recycling, vol. 7, no. 5, p. 70, Sep. 2022, doi: https://doi.org/10.3390/recycling7050070.

[28] I. Lasri, A. R. Solh, and M. E. Belkacemi, "Facial Emotion Recognition of Students using Convolutional Neural Network," 2019 Third International

Conference on Intelligent Computing in Data Sciences (ICDS), Oct. 2019, doi: https://doi.org/10.1109/icds47004.2019.8942386.

[29] D. Yang, A. Alsadoon, P. W. C. Prasad, A. K. Singh, and A. Elchouemi, “An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment,” Procedia Computer Science, vol. 125, pp. 2–10, 2018, doi: https://doi.org/10.1016/j.procs.2017.12.003.