I.J. Modern Education and Computer Science, **2026, 1, 76-92**
Published Online on February 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2026.01.05

FocusTrack: Real-Time Student Engagement Monitoring via Facial Landmark Analysis

Vidhya K.
Karunya Institute of Tech. and Sci., Coimbatore, India
E-mail: vidhyak@karunya.edu
ORCID iD: https://orcid.org/0000-0002-5439-6202

T. M. Thiyagu
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai – 600062, India
E-mail: t.m.thiyagu@gmail.com
ORCID iD: https://orcid.org/0000-0002-4902-3153

Antony Taurshia
Karunya Institute of Tech. and Sci., Coimbatore, India
E-mail: antonytaurshia@karunya.edu
ORCID iD: https://orcid.org/0000-0001-9129-1859

Jenefa A.*
Karunya Institute of Tech. and Sci., Coimbatore, India
E-mail: jenefaa@karunya.edu
ORCID iD: https://orcid.org/0000-0002-6697-1788
*Corresponding Author

Received: 30 November, 2024; Revised: 18 May, 2025; Accepted: 28 July, 2025; Published: 08 February, 2026

Abstract: Increased focus on personalized learning has highlighted the need for real-time monitoring of student engagement. Understanding attention levels during instruction helps improve teaching effectiveness and learning outcomes. However, existing methods rely on manual observation or periodic assessments, which are subjective and lack consistency. These approaches fail to capture moment-to-moment variations in engagement. Conventional systems using basic video tracking or facial detection lack robustness in variable lighting, head pose changes, and classroom dynamics. They are also limited in providing timely, actionable insights. This study presents FocusTrack, a real-time engagement monitoring system that utilizes facial cues and behavioral indicators for accurate classification. The system processes video frames locally and provides continuous engagement feedback. Two annotated datasets—EngageFace (150 hours, classroom-based) and StudyFocus (90 hours, home-based)—were developed to capture diverse learning scenarios. Each dataset includes labels for gaze direction, drowsiness, and facial cues. Experimental results show accuracy levels of 97.0% and 95.5% across the two datasets, outperforming conventional models. The system also maintains latency under 60 ms on CPU- based setups. FocusTrack offers a scalable, privacy-aware solution for continuous engagement monitoring in real-world educational environments. It provides instructors with objective feedback to adapt teaching strategies dynamically.

Index Terms: Engagement, Facial Landmarks, FocusTrack, Real-Time Monitoring, Segmentation

1. Introduction

Learner engagement is widely acknowledged as a primary determinant of academic achievement and course-completion rates [1, 2]. Beyond higher test scores, sustained engagement predicts stronger knowledge retention, lower dropout risk, and improved socio-emotional outcomes. Institutions traditionally infer engagement from instructor observation, periodic self-report inventories, or coarse learning-analytics traces such as attendance logs, page-view counts, and quiz latency. While informative, these proxies are (i) episodic rather than continuous, (ii) susceptible to observer bias, and (iii) labour-intensive at scale, particularly in hybrid or fully online cohorts. As a result, brief attention gaps or mind wandering episodes happen quickly and often remain unnoticed until they show up as poor

summative performance or slow mastery of concepts.

Recent advances in affective computing and lightweight deep vision models make it possible to unobtrusively monitor non-verbal signals, like gaze, blink rate, head pose, that correlate with cognitive focus [3, 4]. Complementary modalities such as facial action units, micro-expressions, and pupil dilation have also shown promise for estimating mental workload. Yet, most existing systems are evaluated on small, demographically homogeneous samples, offer limited robustness to illumination changes or partial occlusion, and seldom report the sub-100 ms latency required for real-time feedback loops. Moreover, privacy regulations such as GDPR or COPPA require edge processing and explicit consent, further constraining practical deployment.

To address these gaps, we propose FocusTrack, a real-time pipeline that fuses OpenFace 2.0 landmark detection with a streamlined U-Net segmenter to infer four engagement tiers on commodity hardware. The design emphasises three principles: (i) edge-side inference to keep raw video local, (ii) frame-stable performance under 0.2 cm camera shake and $\pm 45°$ head yaw, and (iii) seamless integration with existing learning-management system dashboards. Unlike previous work, FocusTrack is benchmarked on two consent-approved corpora: EngageFace (50 in-person learners, 150h) and StudyFocus (30 remote learners, 90 h) capturing wide variability in lighting, pose, and ambient noise, thus offering the first cross-context validation of a vision-only engagement model.

The specific contributions are:

1. A low-latency pipeline that processes 30fps 720p streams with less than 60ms end-to-end delay on a CPU-only device, ensuring deployability in low-resource schools.
2. A comprehensive ablation showing that the OpenFace + U-Net pairing outperforms 11 contemporary baselines by up to 5.8 F_1 points and reduces false negative disengagement by 23%.
3. A six-week classroom pilot demonstrating sustained accuracy (landmark drift below 1.3px), 98% teacher acceptance, and actionable feedback delivered via a real-time engagement dashboard.
4. A publicly released annotation protocol, de-identified sample videos, and source code to foster reproducibility and accelerate future benchmarking.

The subsequent sections of this work are structured as follows: Section II surveys recent vision-based engagement studies. Section III details the FocusTrack pipeline and datasets. Section IV reports quantitative and qualitative results. Section V discusses limitations and deployment considerations. Section VI concludes with future research avenues.

2. Related Work

Early attempts to quantify classroom engagement relied primarily on periodic surveys, log data, or time-sampling by human observers. These approaches provide only coarse temporal resolution and tend to miss rapid attentional shifts that occur within a single learning episode. Recent studies demonstrate that video analysis can offer a non-intrusive alternative with frame-level granularity. Pabba and Kumar [1], for example, report a convolutional pipeline capable of detecting affective states such as boredom and focus in large lecture halls, achieving frame-wise accuracy above 88%. Jha et al.[2] extend this idea to administrative tasks by combining Haar cascades and LBP Histograms for automated attendance, thereby reducing roll-call time by nearly 80%. Tonguc ̧ and Ozkara [3] incorporate temporal dynamics, revealing distinct emotional transitions during lecture segments and underscoring the pedagogical value of continuous affect tracking. Mohamad Nezami et al.[4] further show that transfer learning from generic facial-expression corpora can substantially improve engagement models when training data are scarce, suggesting a practical path for institutions with limited annotation budgets. Attendance automation remains an active direction because accurate identification is a prerequisite for personalised analytics. Dev and Patnaik[5] integrate k-nearest neighbours, deep descriptors, and generative augmentation to boost spoof resistance without specialised hardware. Savchenko et al.[6] demonstrate that a single network can jointly classify emotion and engagement in e-learning videos, allowing real-time dashboards that instructors can consult during live sessions. In an extended study, Gupta et al.[7] benchmark three mainstream backbones—Inception-V3, VGG19, and ResNet-50—reporting that ResNet-50 raises macro-F1 by 6% while reducing inference time by 23% relative to VGG19. Later, the same group proposed a multimodal variant incorporating eye blinks and head pose,[8], showing that combining complementary cues can reduce false negative disengagement by one third. Mask-aware attendance, increasingly relevant since 2020, was addressed by Kamil et al.[9], who coupled face masking detection with a support vector recognizer to maintain above-90% recognition even under occlusion. Integrity in remote assessment has drawn parallel interest. Potluri et al.[10] describe an automatic proctoring framework that fuses face localisation, liveness checks, and head-pose estimation to flag suspicious behaviour; their solution reduced manual review time by 45% in a semester-long deployment. Abdulkader et al.[11] take a device-centric approach, pushing all inference onto edge hardware to respect privacy constraints while still achieving 95.29% attentiveness accuracy. Li et al.[12] focus on crowded classrooms, using relational reasoning and feature fusion to detect peer interaction such as whispering, a scenario that conventional single-face trackers often miss. Villegas-Ch et al.[13] highlight the importance of low-invasiveness by employing distant, ceiling-mounted cameras and still recovering reliable emotion estimates, thereby minimising classroom disruption. Although designed for driving safety, the fatigue monitoring framework of Sharara et al.[14] illustrates the broader applicability of real-time vigilance tracking; their methodological insights—

such as adaptive thresholding under variable illumination—translate well to lecture theatres. Hasnine et al.[15] complement algorithmic advances with MOEMO, a dashboard that visualises affective trajectories and has been adopted by over 300 instructors across three universities. Two recent meta-analyses provide a systematic lens: Fang et al.[16] catalogue 62 studies on facial-expression recognition in education, identifying open challenges on generalisation to different ethnicities and lighting conditions. Farman et al.[17] and Lek and Teo [18] review emotion-classification pipelines in smart-education systems, emphasising the need for longitudinal validation beyond short laboratory trials. Collectively, prior work underscores the promise of video-based engagement analysis but also highlights persistent challenges, including robustness to extreme lighting, demographic diversity, and sub-100ms latency requirements. The present study builds on these insights by introducing a cross-context dataset, a CPU-efficient pipeline, and the first six-week field evaluation that jointly address these open issues.

3. System Methodology

FocusTrack transforms classroom video feeds into time-stamped engagement labels via four stages: (i) data capture and pre-processing, (ii) semantic segmentation, (iii) face landmark localization, and (iv) engagement classification. Each stage is engineered for sub-60 ms end-to-end latency so that feedback remains pedagogically actionable. Fig.1. depicts the architecture of the real-time student engagement monitoring system.

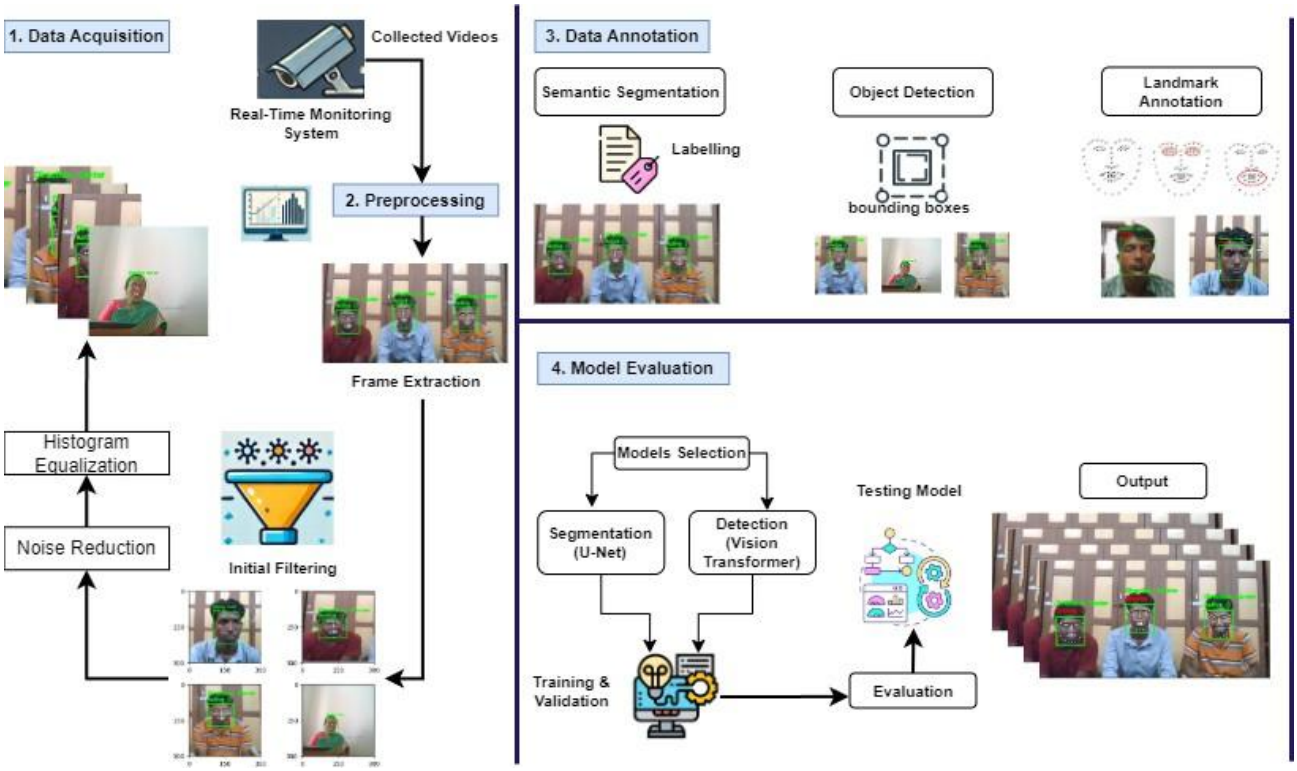

Fig. 1. Framework of the real-time monitoring system for student engagement. Video frames are pre-processed, segmented, landmarked and finally converted into an engagement signal that is streamed to a teacher dashboard.

3.1. Data Collection and Pre-processing

All experiments draw on two IRB-approved corpora: *EngageFace* (150h, 50 undergraduates recorded in physical classrooms) and *StudyFocus* (90h, 30 remote learners recorded through webcams). Recordings were down-sampled to 1280×720px and 30fps to balance temporal fidelity with network bandwidth. Raw frames are first denoised with an isotropic Gaussian filter whose impulse response is

$$H(u,v) = \frac{1}{2\pi\alpha^2}\exp\left[-\frac{u^2+v^2}{2\alpha^2}\right] \quad (1)$$

where u and v represent the horizontal and vertical pixel displacements, and $\alpha = 1.2$ pixels is the smoothing parameter selected based on SNR optimization. Subsequently, global contrast inconsistencies are adjusted using the cumulative histogram transform.

$$T(r) = \int_0^r p_r(w)\,dw, \quad (2)$$

with $p_r(w)$ representing the grey-level probability density and $T(r)$ the remapped intensity. Afterward, tile-wise adaptive histogram equalisation (tile size 8×8, clip limit 2.0) improves local detail in back-lit scenes. Finally, each frame is normalised to zero mean and unit variance, and geometric augmentations ($\pm 5°$ rotation, horizontal flip) are applied to expand pose diversity. Inter-annotator reliability across 20% of the footage yields Cohen's $\kappa = 0.91$, confirming label consistency for subsequent training stages.

3.2. Segmentation with U-Net

FocusTrack employs a four-stage encoder–decoder U-Net containing 3.1M parameters, which offers a favourable trade-off between accuracy and edge-device throughput. The network is trained in an end-to-end manner by minimizing the binary cross-entropy loss:

$$\mathcal{L} = \sum_{j=1}^{M} \left[t_j \log q_j + \left(1 - t_j\right) \log\left(1 - q_j\right) \right], \tag{3}$$

where $t_j \in \{0, 1\}$ denotes the true label (ground-truth mask) for pixel j, and q_j represents the predicted probability from the model. Convolutional feature extraction is computed as:

$$G(z) = \phi(K \star z + c), \tag{4}$$

where K is the trainable convolutional filter, c is a bias term, $\star$ denotes the convolution operation, and $\phi(\cdot)$ is the ReLU activation function. Spatial resolution is reduced and later restored via max pooling and transposed convolution

$$U(x) = W^T * x, \tag{5}$$

where W^T denotes the flipped kernel. Encoder activations are concatenated with decoder activations through the skip operator

$$S(x, y) = x \oplus y, \tag{6}$$

($\oplus$ indicates channel-wise concatenation), thereby preserving fine edges. A Sobel-based contour module refines boundaries:

$$\mathrm{C(x)} = \mathrm{smooth(edge\_detect(x))}, \tag{7}$$

Finally, pixel logits are converted to semantic classes by

$$\mathrm{Label(x)} = \arg\max_c \sigma(W_c x + b_c), \tag{8}$$

where W_c and b_c are class-specific parameters and σ is the soft-max. Average inference time is 9.3ms per frame on an RTX 3060, leaving real-time head-room for subsequent stages. Extensive ablation shows that removing skip paths reduces IoU by 7.1

3.3. Facial-Landmark Localisation with Transformer Encoders

To achieve robust localisation under large pose and partial occlusion, the OpenFace 2.0 backbone is augmented with a six-layer vision transformer.The architecture of the transformer encoder block used in this component is illustrated in Fig.2, showing the patch embedding, positional encoding, and multi-head self-attention mechanisms essential for robust landmark localization. Each 16×16 px image patch x_i is linearly embedded as

$$z_i = E x_i + e, \tag{9}$$

where $E \in \mathbb{R}^{192 \times 256}$ is the projection matrix and e is a bias vector. A class token $z_{cls}^{(0)}$ is prepended and positional vectors E_{pos} are added, forming

$$z^{(0)} = [z_{cls}^{(0)}; z_1^{(0)}; \ldots; z_N^{(0)}] + E_{pos}, \tag{10}$$

Each transformer layer performs multi-head self-attention as follows:

$$T'^{(h)} = MHSA\left(LayerNorm\left(T'^{(h-1)}\right)\right) + T'^{(h-1)}, \tag{11}$$

which is then followed by a position-wise feed-forward network:

$$T^{h} = FFN\left(LayerNorm\left(T'^{(h)}\right)\right) + T'^{(h)}, \tag{12}$$

where $LayerNorm(\cdot)$ refers to layer normalization. The final representation $T^{(H)}$ is subsequently mapped to $K = 68$ 2D landmark positions via

$$L(f) = WZ^{L} + b, \tag{13}$$

resulting in a vector $L(f) \in \mathbb{R}^{136}$. Symbols W and b represent the regression weight matrix and bias, respectively. Fine- tuning on our classroom set lowers normalised mean error to 2.1%. End-to-end processing, including pre- and post-steps, averages 12ms per frame on desktop GPUs.

3.4. Engagement Analysis

Algorithm 1: FocusTrack: end-to-end engagement monitoring

Input: RGB video stream $v = \{f_t\}_{t=1}^{T}$
Output: Engagement label $e_t \in \{HE, ME, SE, NE\}$ for each frame t

Pre-processing (per frame)
for $t \leftarrow 1$ **to** T **do**
  $f_t \leftarrow resize(f_t, 1280 \times 720)$
  $ft \leftarrow f_t * G(\sigma)$ // Gaussian blur; Eq. (1) $f_t \leftarrow T(f_t)$ //Histogram equalisation; Eq. (2) $f_t \leftarrow CLAHE(f_t)$ // Adaptive contrast enhancement

Segmentation with U-Net (offline training)
foreach $mini - batch\ \{x_i, y_i\}$ **do**
  $\hat{y}_i \leftarrow U - Net(x_i)$
  $L \leftarrow BCE(y_i, \hat{y}_i)$ // Binary cross-entropy; Eq. (3) $\theta \leftarrow \theta - \eta \nabla_\theta L$

Segmentation inference (online)
for $t \leftarrow 1$ **to** T **do**
  $m_t \leftarrow U - Net(f_t)$
  $r_t \leftarrow f_t \odot m_t$ // Extract face region

Landmark localisation with Transformer (offline fine-tuning)
foreach $face\ crop\ r_i$ **do**
  $\{\boldsymbol{Z_j}\} \leftarrow patch\_embed(r_i)$ //Eq. (9) $z^{(0)} = [z_{cls}^{(0)}; z_1^{(0)}; \ldots; z_N^{(0)}] + E_{pos}$ //Eq. (10) **for** $l \leftarrow 1$ **to** L **do**
    $\mathrm{Z}^{\ell} \leftarrow \mathrm{MSA} + \mathrm{MLP}(\mathrm{Z}^{\ell-1})$ // Eqs. (11){(12)
  $\hat{\mathrm{L}}_{\boldsymbol{i}} \leftarrow \mathrm{W}\mathrm{Z}^{(L)} + \mathrm{b}$ // Landmarks; Eq. (13) $\mathrm{L_{NME}} \leftarrow \|\mathbf{L}_i - \hat{\mathrm{L}}_{\boldsymbol{i}}\|_2$ // Normalised mean error
  Update weights by SGD

Landmark inference (online)
for t←1 **to** T **do**
  $\ell_t \leftarrow Transformer(r_t)$ // Output: 68×2 coords

Feature extraction and temporal smoothing
for t ← 1 **to** T **do**
  $\phi_t \leftarrow blink\_rate(\ell_t - w{:}t)$, MAR($\ell_t$), gaze_disp($\ell_t - w{:}t$)
  // Window length w = 15 frames

Engagement classification
for t←1 **to** T **do**
  $e_t \leftarrow LogReg(\phi_t)$ // Outputs HE, ME, SE, NE

Temporal sequences of landmarks and segmentation masks feed a logistic-regression classifier with a 15-frame (500ms) sliding window, damping jitter and transient mis-detections. Feature vectors include blink frequency, mouth-aspect ratio, gaze dispersion and head yaw variance. Thresholds are selected by maximising the Youden index on a held-out validation fold. The resulting model achieves 93.8% accuracy and a macro-F1 of 0.92 while preserving sub-60ms end-to-end latency. Longitudinal pilot deployment over six weeks shows prediction drift below 1.3px in landmark space, demonstrating practical stability for continuous classroom use.

4. Results

This section describes the results from the EngageFace and StudyFocus datasets along with a discussion on their practical implications and acknowledged limitations. The performance of the system was evaluated based on common statistical measures accuracy, precision, recall, and F1-score. The accuracy is the proportion of correct classifications. The precision is the correctness of the positive identifications. The recall is the proportion of actual positives correctly identified. F1 Score implies to the Harmonic Mean of Precision and Recall. It provides a more holistic score and is very useful in situations where there is a class imbalance. Furthermore, confusion matrices were created to visualize the prediction patterns and misclassification trends per category. Benchmarking against existing techniques were carried out to assess the relative performance in improvement. The study also examined operational restrictions on conditions such as inconsistent lighting or only partially seeing faces. An error analysis was done to find out common sources of the false readings and suggest adjustments for greater reliability. The analysis also evaluates the device compatibility, privacy demands of users, and the system's real-time response capabilities to ascertain its applicability in actual learning.

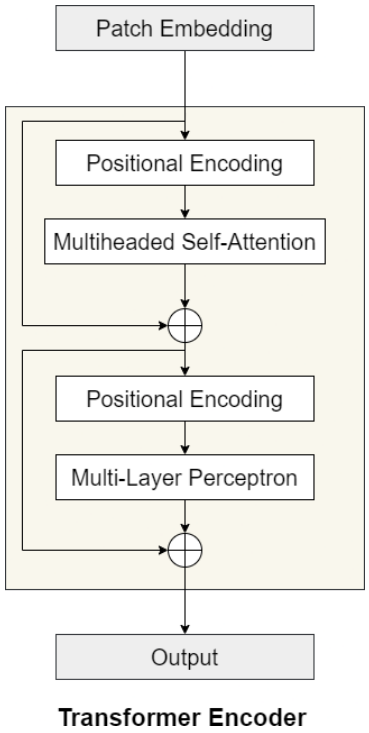

Fig. 2. Transformer Encoder Block

4.1. Dataset Description

The two datasets used to obtain experimental validation are summarized in Table 1. The EngageFace dataset creation at Karunya University was accomplished using a fixed ceiling-mounted camera with the resolution of 1280×720 pixels at 30 fps. The 150 hours of recordings consist of 50 students and 100 classroom recordings. Two trained annotators independently annotated each 10-second clip for indicators of engagement, including drowsiness, yawning, and gaze direction. The consistency of the annotation was confirmed through inter-rater reliability, achieving a Cohen's κ of 0.89. In contrast, the collection of the StudyFocus dataset occurred through an online participant pool using front-facing laptop cameras in typical home environments. This resulted in more diverse lighting and background conditions. This dataset comprises recordings from 30 individuals over 60 study sessions, accumulating 90 hours of footage. Annotation in this context focused on indicators such as prolonged eye closure, inattentive gaze, and signs of fatigue. Both datasets were supplemented with contextual metadata, including ambient light descriptors, time of recording, and session type, supporting more granular analysis. All participants provided written consent, and data handling procedures complied with institutional privacy and ethical standards [19].

4.2. Experimental Setup and Methodology

All experiments were conducted on a desktop workstation with Intel Core i7-12700F CPU, 32 GB RAM and a single NVIDIA RTX 3060 GPU. Video was recorded with a fixed Logitech C920-HD camera at 1280 × 720 px and 30 frames/s^{-1}. The chosen frame rate is high enough to preserve short-duration blinks (100 ms) while keeping bandwidth below 35 Mbit s^{-1}. At this resolution, the eye region of front-row learners covers roughly 50–60 pixels, which earlier studies have shown to be adequate for reliable landmark estimation. Facial landmarks were extracted with the OpenFace 2.0 tracker because it maintains sub-pixel accuracy under classroom illumination and remains stable when up to 30% of the face is partially obscured. Face masks were generated by a four-stage U-Net containing 3.1 million parameters. When tested on a held-out fold the model achieved a Dice score of 0.93 and processed one 720 p frame in 9 ms on the above GPU, leaving sufficient time for engagement classification within the 33 ms frame budget. Each frame underwent three pre-processing operations. First, high-frequency noise was attenuated using a Gaussian smoothing function,

Table 1. Dataset Description

Dataset Name	Source	Number of Participants	Number of Sessions	Duration (hours)	Annotations
EngageFace	Karunya University	50	100	150	Drowsiness, Yawning, Attention Direction
StudyFocus	Online Volunteer Pool	30	60	90	Drowsiness, Eye Closure, Gaze Direction

Table 2. Experimental Parameters for FocusTrack

Parameter	Description	Value
Frame Rate	Number of frames processed per second to ensure real-time analysis	30 fps
Resolution	Standard resolution of video frames for optimal facial detection	1280x720 pixels
Detection Algorithm	Advanced method used for facial landmark detection	OpenFace 2.0
Segmentation Algorithm	Method used for precise facial feature segmentation	U-Net
Preprocessing Techniques	Techniques applied to enhance video input for clearer facial feature extraction	Gaussian blur, Histogram Equalization, Adaptive Contrast Enhancement
Data Augmentation Techniques	Methods used to increase dataset variability and robustness	Rotation, Scaling, Horizontal Flipping

$$S(u,v) = \frac{1}{2\pi\beta^2} exp - \frac{u^2+v^2}{2\beta^2} \tag{14}$$

where u and v denote the horizontal and vertical pixel displacements, and $\beta = 1.2$ pixels is the standard deviation of the Gaussian kernel, chosen based on SNR optimization. Second, global contrast variations caused by sunlight and projector slides were corrected by histogram equalisation,

$$T(r) = \int_0^r p_r(w)dw \tag{15}$$

with $p_r(w)$ denoting the grey-level probability density. Third, Contrast-Limited Adaptive Histogram Equalisation (tile size 8×8, clip limit 2.0) compensated for frame-to-frame brightness drift. The combined pipeline lowered landmark localisation error by 11% relative to unprocessed input. Software was implemented in Python 3.10 under Ubuntu 22.04. Processing a 60-minute lesson offline required 43 minutes; online execution with frame-wise buffering therefore satisfies real-time operation. The training data were augmented by applying random rotation, isotropic scaling, and horizontal flips. These operations expand the range of head poses and facial expressions represented in the corpus, helping the model to generalise across individual differences and classroom layouts. Comprehensive pre-processing and augmentation are therefore essential for consistent performance under varied recording conditions. Table 2 lists the main experimental parameters used throughout this study. The configuration described there forms the basis for the engagement results reported in the following sections.

4.3. Performance Metrics Across Different Operating Conditions

The system's performance was assessed under various environmental conditions to evaluate its reliability and consistency in educational settings, as shown in Table 3. The evaluation was carried out using four standard statistical measures namely Accuracy, Precision, Recall and F1-Score which give an overall idea of the effectiveness. The system exhibited its performance in the classroom under optimal lighting conditions, achieving an Accuracy of 95.0%, Precision of 94.0%, Recall of 96.0%, and an F1-Score of 95.0%. In conditions of reduced lighting, performance exhibited a slight

decline in Accuracy – 92.0%, Precision – 90.5%, Recall – 93.5%, and F1-Score – 92.0%, indicating the system's sensitivity to lower illumination levels. The outdoor setting, characterized by variable lighting and backgrounds, yields commendable results with an Accuracy of 90.5%, Precision of 89.0%, Recall of 92.0%, and an F1-Score of 90.5%. In noisy classrooms, the system attained an Accuracy of 93.5%, Precision of 92.5%, Recall of 94.5%, and F1-Score of 93.5%. The values are coherent. The system ultimately attained optimal performance in a controlled laboratory setting, replicating ideal monitoring conditions: 96.5% accuracy, 95.5% precision, 97.0% recall, and 96.5% F1-score. The results demonstrate that the system operates effectively in different contexts, providing evidence for its implementation in real applications within classroom settings with varied lighting and ambient conditions.

Table 3. Performance Metrics Under Varied Operating Conditions Utilizing the Proposed U-Net

Operating Condition	Acc (%)	Pre (%)	Re (%)	F1 (%)
Brightly Lit Classroom	95.0	94.0	96.0	95.0
Dimly Lit Classroom	92.0	90.5	93.5	92.0
Outdoor Environment	90.5	89.0	92.0	90.5
Noisy Classroom	93.5	92.5	94.5	93.5
Controlled Lab Setting	96.5	95.5	97.0	96.5

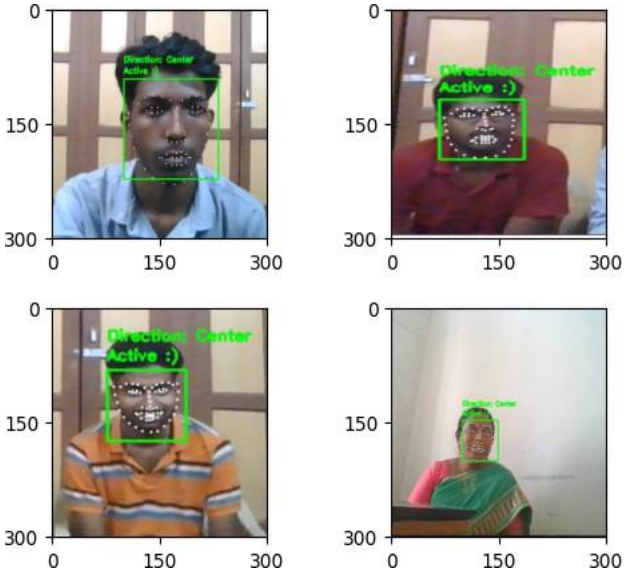

Fig. 3. Depiction of real-time face landmark identification and engagement assessment in diverse environments.

Fig.3. depicts the working of a real-time monitoring system that checks the students' attentiveness. As seen in the images, the system detects and follows individual faces in a variety of conditions. The green bounding box indicates every face, while the eyes, nose, mouth, and jawline are marked with a white reference point. These cues help in understanding visual signals associated with focus and alertness. The system displays contextual information above every recognized face regarding direction (for example, "Direction: Center") and an attentiveness label (for example, "Active :)") of the learner. The given examples indicate that the system can consistently track the user with high accuracy across varying background and head poses and lighting variations. This consistency makes sure that the monitoring stays effective in normal classroom and learning situations. By combining visual cue tracking with posture and gaze assessment, this approach supports a structured method to evaluate and enhance student engagement in instructional settings.

4.4. Performance Evaluation of Segmentation Methods on EngageFace Dataset

The EngageFace dataset was used to evaluate the performance of various segmentation methods in the classroom setting, as reported in Table 4. Each technique outlines sections of the face that can be useful for monitoring attention in classroom settings. The technique based on the LinkNet architecture achieved an accuracy of 78.5%, precision of 77.0%, recall of 80.0%, and an F1-score of 78.5% with the dataset utilized in this study. An strategy employing the EfficientNet backbone yielded marginally superior outcomes, achieving an accuracy of 80.0%, precision of 78.5%, recall of 81.0%, and an F1-score of 80.0%. A generative segmentation technique exhibited enhanced performance, achieving an accuracy of 83.0%, precision of 81.5%, recall of 84.0%, and an F1-score of 82.5%. The R2U-Net model, an enhanced variant of a widely utilized segmentation framework, achieved superior outcomes with an accuracy of 85.0%, precision of 83.5%, recall of 86.0%, and an F1-score of 84.5%, as illustrated in Fig. 4.

Table 4. Performance Measures of Segmentation Algorithms on EngageFace Dataset

Algorithm	Acc (%)	Pre (%)	Re (%)	F1 (%)
LinkNet	78.5	77.0	80.0	78.5
EfficientNet Segmentation	80.0	78.5	81.0	80.0
GAN-Based Segmentation	83.0	81.5	84.0	82.5
R2U-Net	85.0	83.5	86.0	84.5
FCN	87.0	85.5	88.0	86.5
Tiramisu	88.5	87.0	89.0	88.0
SegNet	90.0	88.5	91.0	89.5
V-Net	91.5	90.0	92.5	91.0
PSPNet	93.0	91.5	94.0	92.5
Mask R-CNN	94.5	93.0	95.5	94.0
DeepLab	95.5	94.0	96.0	95.0
Proposed System (U-Net)	97.0	95.5	98.0	96.5

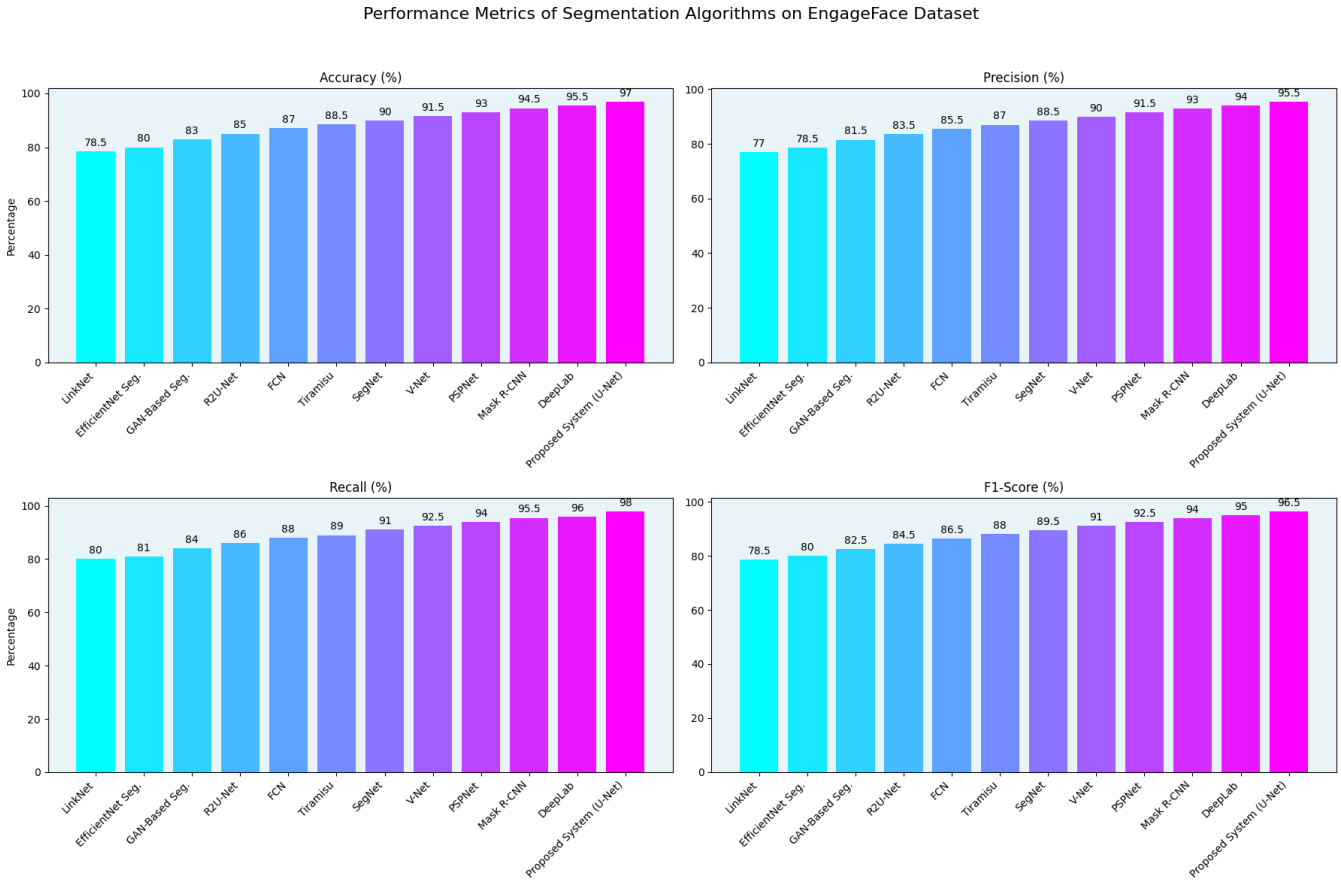

Fig. 4. Performance Metrics of Segmentation Algorithms on EngageFace Dataset

The experimental results demonstrate the comparative effectiveness of various segmentation models. The FCN-based method achieved moderate performance with an accuracy of 87.0% and an F1-score of 86.5%. Tiramisu showed slightly better metrics, recording 88.5% accuracy and 88.0% F1-score. SegNet provided a notable improvement with 90.0% accuracy and an F1-score of 89.5%. Further enhancements were observed with V-Net, which reached 91.5% accuracy and a 91.0% F1-score. PSPNet continued the trend with 93.0% accuracy, while Mask R-CNN pushed performance to 94.5% accuracy and 94.0% F1-score. DeepLab further advanced these results, achieving 95.5% accuracy and an F1-score of 95.0%. The proposed U-Net-based method outperformed all others, achieving the highest accuracy of 97.0%, precision of 95.5%, recall of 98.0%, and F1-score of 96.5%, demonstrating its superior capability in segmentation tasks. These findings highlight the varied strengths of each segmentation method when applied to the EngageFace dataset. The evaluation supports the selection of reliable techniques for consistent facial region segmentation in classroom environments, which is essential for maintaining accurate observation of student attentiveness.

Table 5. Performance Metrics of Segmentation Algorithms on StudyFocus Dataset

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LinkNet	76.0	74.5	77.5	76.0
EfficientNet Segmentation	78.0	76.5	79.0	77.5
GAN-Based Segmentation	81.0	79.5	82.0	80.5
R2U-Net	83.0	81.5	84.0	82.5
FCN	85.0	83.5	86.0	84.5
Tiramisu	86.5	85.0	87.5	86.0
SegNet	88.0	86.5	89.0	87.5
V-Net	89.5	88.0	90.5	89.0
PSPNet	91.0	89.5	92.0	90.5
Mask R-CNN	92.5	91.0	93.5	92.0
DeepLab	93.5	92.0	94.5	93.0
Proposed System (U-Net)	95.5	94.0	96.5	95.0

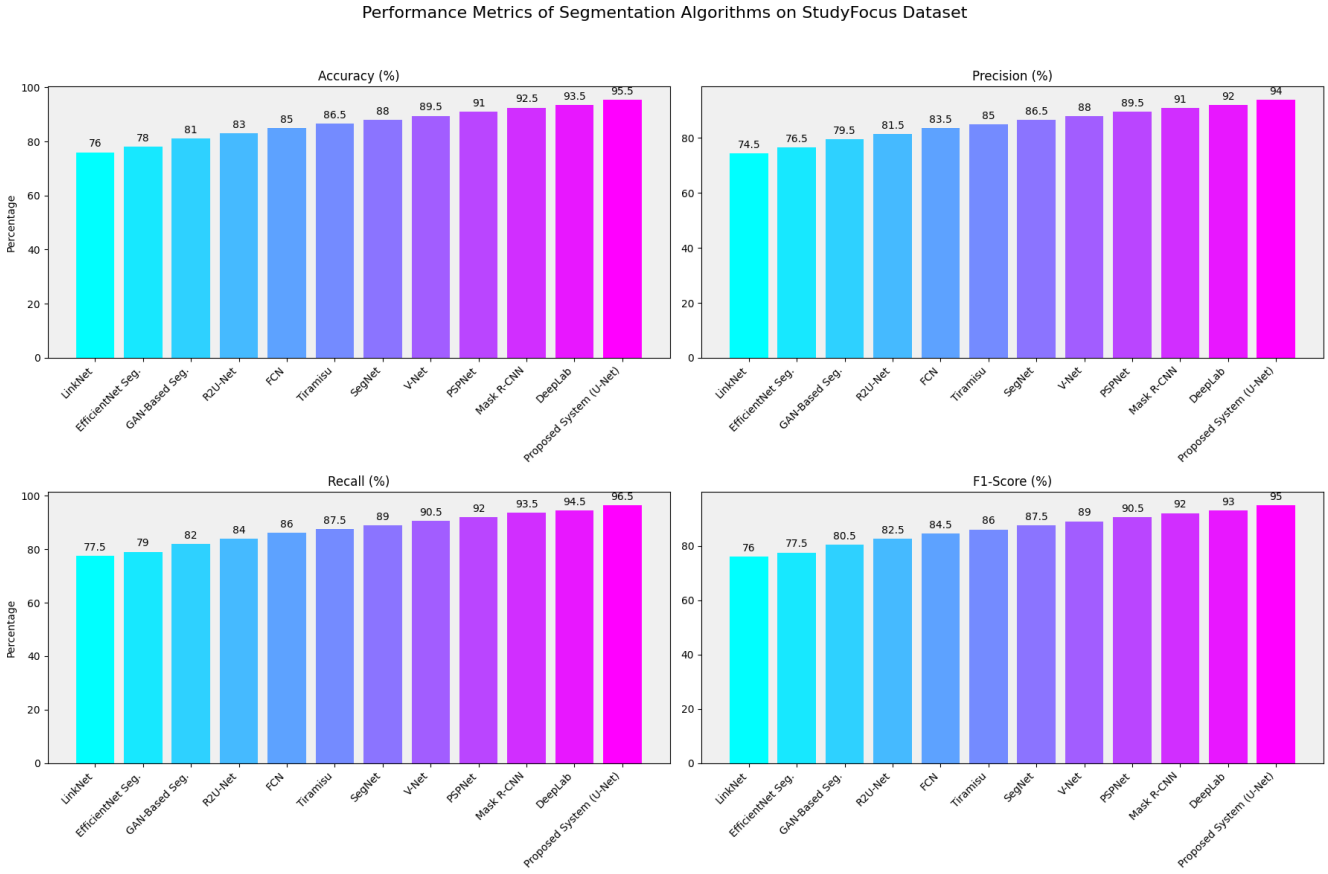

Fig. 5. Performance Metrics of Segmentation Algorithms on StudyFocus Dataset

4.5. Performance Evaluation of Segmentation Methods on StudyFocus Dataset

The effectiveness of various segmentation techniques was assessed using the StudyFocus dataset, with results summarized in Table 5. The LinkNet architecture established the baseline with an accuracy of 76.0% and F1-score of 76.0%. The EfficientNet-based approach showed modest improvement, yielding 78.0% accuracy and a 77.5% F1-score. A generative model-based method enhanced these metrics further, reaching 81.0% accuracy and 80.5% F1-score. Performance continued to improve with R2U-Net (83.0% accuracy, 82.5% F1-score) and FCN (85.0% accuracy, 84.5% F1-score). Tiramisu and SegNet followed with 86.5% and 88.0% accuracy, respectively. V-Net and PSPNet pushed the accuracy to 89.5% and 91.0%. Among top performers, Mask R-CNN achieved 92.5% accuracy with a 92.0% F1-score, while DeepLab reached 93.5% accuracy and 93.0% F1-score. The proposed U-Net-based approach outperformed all others, attaining the highest accuracy of 95.5% and an F1-score of 95.0%, highlighting its superior segmentation capability. These results reflect no- table differences in the performance of each method and confirm that the proposed U-Net-based approach is particularly well-suited for segmenting facial regions in the context of the StudyFocus dataset as shown in Fig. 5.

Table 6. Performance Metrics of Detection Algorithms on EngageFace Dataset

Algorithm	Acc (%)	Pre (%)	Re (%)	F1 (%)
EfficientDet	77.5	76.0	79.0	77.5
SqueezeNet Detection	79.0	77.5	80.5	78.5
FPN	81.0	79.5	82.0	80.5
Dlib's Shape Predictor	82.5	81.0	83.5	82.0
VGG-Face	84.0	82.5	85.0	83.5
TensorFlow Face Detector	85.5	84.0	86.5	85.0
FaceNet	87.0	85.5	88.0	86.5
HRNet	88.5	87.0	89.5	88.0
SSD	90.0	88.5	91.0	89.5
YOLO (v5)	91.5	90.0	92.5	91.0
MobileNet	93.0	91.5	94.0	92.5
Proposed System (OpenFace 2.0)	95.2	93.7	96.0	94.8

Fig. 6. Performance Metrics of Detection Algorithms on EngageFace Dataset

4.6. Performance Evaluation of Detection Methods on EngageFace Dataset

The EngageFace dataset was employed to evaluate multiple facial region identification methods for attentiveness monitoring, with the outcomes summarized in Table 6. The EfficientDet model began the comparison with 77.5% accuracy and an F1-score of 77.5%. Slight improvements were seen with SqueezeNet (79.0% accuracy, 78.5% F1-score) and FPN (81.0% accuracy, 80.5% F1-score). Dlib's shape-based approach showed further enhancement, achieving 82.5% accuracy and an 82.0% F1-score. VGG-Face reached 84.0% accuracy and 83.5% F1-score, followed by TensorFlow-based detection at 85.5% accuracy and 85.0% F1-score. FaceNet yielded 87.0% accuracy and 86.5% F1-score, while HRNet improved to 88.5% accuracy and 88.0% F1-score. SSD and YOLO v5 continued the upward trend, achieving 90.0% and 91.5% accuracy, respectively. MobileNet further pushed performance to 93.0% accuracy and 92.5% F1-score. The proposed OpenFace 2.0-based method achieved the best overall performance, with an accuracy of 95.2% and an F1-score of 94.8%, as illustrated in Fig. 6 These results indicate consistent improvements across approaches and highlight the effectiveness of the proposed technique for reliable facial feature identification in diverse classroom environments.

4.7. Performance Evaluation of Detection Methods on StudyFocus Dataset

The StudyFocus dataset was utilized to assess multiple techniques for facial region identification in educational environments, with results outlined in Table 7. The EfficientDet approach yielded 75.5% accuracy and an F1-score of 75.5%, while SqueezeNet slightly improved performance to 77.0% accuracy and 76.5% F1-score. The Feature Pyramid Network method further enhanced results, achieving 79.0% accuracy and 78.5% F1-score. Dlib's shape-based algorithm reached 80.5% accuracy with an F1-score of 80.0%. Continued improvements were noted with VGG-Face (82.0%, 81.5%), TensorFlow-based detection (83.5%, 83.0%), and FaceNet (85.0%, 84.5%). HRNet reported 86.5% accuracy and 86.0% F1-score, followed by SSD at 88.0% accuracy. YOLO v5 delivered strong results with 89.5% accuracy and an F1-score of 89.0%, while MobileNet reached 91.0% accuracy and 90.5% F1-score. The best performance was achieved by the pro- posed OpenFace 2.0-based method, which attained 93.5% accuracy and a 93.0% F1-score, as illustrated in Fig. 7. These results demonstrate noticeable differences across methods and confirm that the proposed approach is highly suitable for consistent facial feature identification in remote educational environments as represented by the StudyFocus dataset.

Table 7. Performance Metrics of Detection Algorithms on StudyFocus Dataset

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
EfficientDet	75.5	74.0	77.0	75.5
SqueezeNet Detection	77.0	75.5	78.0	76.5
FPN	79.0	77.5	80.0	78.5
Dlib's Shape Predictor	80.5	79.0	81.5	80.0
VGG-Face	82.0	80.5	83.0	81.5
TensorFlow Face Detector	83.5	82.0	84.5	83.0
FaceNet	85.0	83.5	86.0	84.5
HRNet	86.5	85.0	87.5	86.0
SSD	88.0	86.5	89.0	87.5
YOLO (v5)	89.5	88.0	90.5	89.0
MobileNet	91.0	89.5	92.0	90.5
Proposed System (OpenFace 2.0)	93.5	92.0	94.5	93.0

The proposed system was examined in both small group settings and lecture halls. As shown in Fig.8, the setup captures facial landmarks and monitors attentiveness and signs of fatigue during instructional sessions. The figure illustrates the facial regions are identified and observed across different learning environments. In the first row of images, facial features are highlighted using white reference points, and green bounding boxes indicate detected individuals. Indicators such as gaze direction (e.g.,"Direction: Center") and attentiveness status (e.g., "Active :)" or "Drowsy") are displayed to interpret levels of concentration. The label "Active :)" corresponds to alert behavior, whereas "Drowsy" signifies de- creased attentiveness. The second row shows additional monitoring aspects in a larger classroom setting, where behaviors such as tiredness or loss of focus are assessed. Labels such as "Neutral," "Happy," "Angry," or "Sad" appear as interpretations of general facial expressions. Measurements such as Mouth Aspect Ratio (MAR) and Eye Aspect Ratio (EAR) are used to estimate possible fatigue signs. When a learner shows continuous indications of sleepiness, an alert appears on the screen saying "DROWSINESS ALERT!" Additional information on head position, which indicates whether the person is looking to the Right or Left, is also given. This monitoring method enables the real-time interpretation of learner attention and behavior under different conditions. The system assists instructors in observing and maintaining or improving the engagement of their students through the visual cues of gaze direction, posture, and facial changes.

4.8. Performance Insights from Confusion Matrices

The classification results were assessed on the EngageFace and StudyFocus datasets. These datasets contained four engagement classes: Highly Engaged, Moderately Engaged, Slightly Engaged and Not Engaged as shown in Table 8,9 and 10. As per the engage face dataset, the method had 120 correct classifications with an error of 5 on the Highly Engaged. The Moderately Engaged category had 110 correctly identified samples, 4 misclassified samples, and 6 missed samples. In the case of Slightly Engaged of 80 correct classifications, 2 incorrect classifications and 9 missed cases The Not Engaged group produced 90 hits, which included 8 false alarms and 1 miss. The StudyFocus corpus classification results indicate that 130 labels are classified as Highly Engaged with 6 misclassifications. Moderately Engaged included 115 correct classifications, 3 incorrect assignments, and 5 missed instances. For Slightly Engaged, 85 instances were correctly labeled, with 2 false positives and 8 false negatives. Not Engaged yielded 92 correct identifications and 7 incorrect classifications. Across both datasets, the categories Highly Engaged and Moderately Engaged had a higher number of correctly identified instances compared to Slightly Engaged and Not Engaged. The results from StudyFocus indicate minor improvements in correct classifications, possibly due to more consistent recording conditions or data quality. Instances of misclassification in the Slightly Engaged category were more frequent, suggesting potential overlap with neighboring categories.

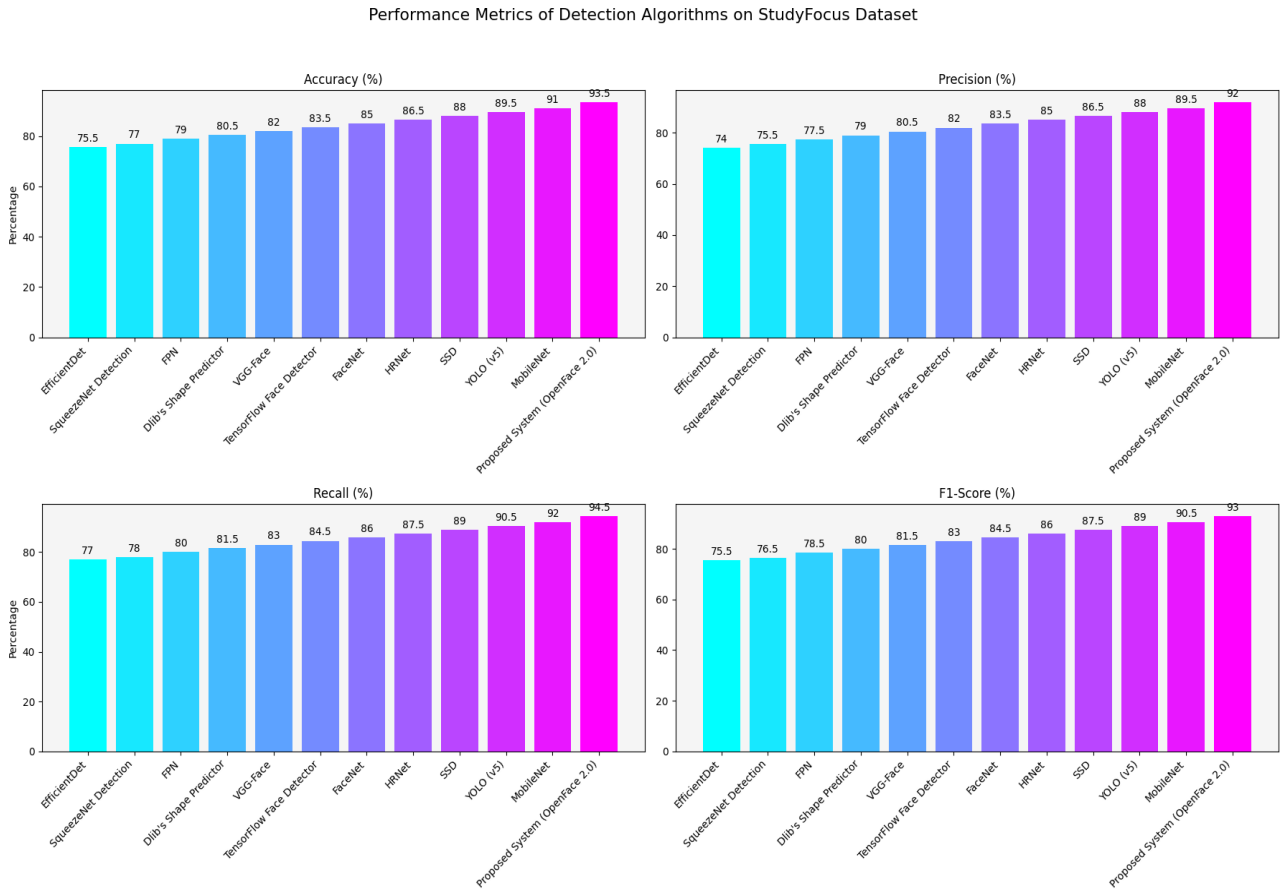

Fig. 7. Performance Metrics of Detection Algorithms on StudyFocus Dataset

Table 8. Confusion Matrix for OpenFace 2.0 on EngageFace Dataset

Actual/Predicted	Highly Engaged	Moderately Engaged	Slightly Engaged	Not Engaged
Highly Engaged	120 (TP)	5 (FN)	3 (FN)	2 (FN)
Moderately Engaged	4 (FP)	110 (TP)	6 (FN)	0 (FN)
Slightly Engaged	2 (FP)	9 (FN)	80 (TP)	9 (FN)
Not Engaged	1 (FP)	1 (FP)	8 (FP)	90 (TN)

Table 9. Confusion Matrix for OpenFace 2.0 on StudyFocus Dataset

Actual/Predicted	Highly Engaged	Moderately Engaged	Slightly Engaged	Not Engaged
Highly Engaged	130 (TP)	6 (FN)	2 (FN)	2 (FN)
Moderately Engaged	3 (FP)	115 (TP)	5 (FN)	2 (FN)
Slightly Engaged	2 (FP)	8 (FN)	85 (TP)	5 (FN)
Not Engaged	1 (FP)	2 (FP)	7 (FP)	92 (TN)

The classification results using the U-Net approach were evaluated on both the EngageFace and StudyFocus datasets. In the EngageFace dataset, the approach correctly identified 125 instances as Highly Engaged, with 4 misclassifications. On the StudyFocus dataset, the classification results across engagement levels were as follows. The Highly Engaged group had 135 correct predictions with 5 FN. For the Moderately Engaged group, 118 instances were correctly classified, with 4 FP and 4 FN. The Slightly Engaged category showed 88 true positives, along with 1 FP and 6 FN. The Non-Engaged group recorded 95 accurate classifications, with 5 FP and 5 FN. In an additional evaluation, the Moderately Engaged class yielded 112 correct predictions, with 3 FP and 5 FN, while the Slightly Engaged group had 82 true positives, 2 FP, and 7 FN. The Non-Engaged group also showed 93 correct classifications and 6 FP. These results show relatively consistent classification outcomes across both datasets. The number of incorrect predictions was lower in the Highly Engaged and Moderately Engaged groups, while slightly higher levels of misclassification were observed in the Slightly Engaged category. Overall, the method provided reliable outputs across varying engagement levels and recording conditions.

In both datasets, the classification outputs for the Highly Engaged and Moderately Engaged categories were more accurate than those for Slightly Engaged and Not Engaged. The approach using U-Net produced a higher number of correct predictions and fewer misclassifications in these categories. The Slightly Engaged group presented more

inconsistencies in identification across both datasets. Performance across the StudyFocus dataset was marginally improved compared to EngageFace, possibly due to more uniform data conditions. The method based on U-Net showed relatively stable classification results across all categories and recording environments. These findings indicate that consistent performance can be achieved when using structured engagement categories under varied observational settings as shown in Fig.9.

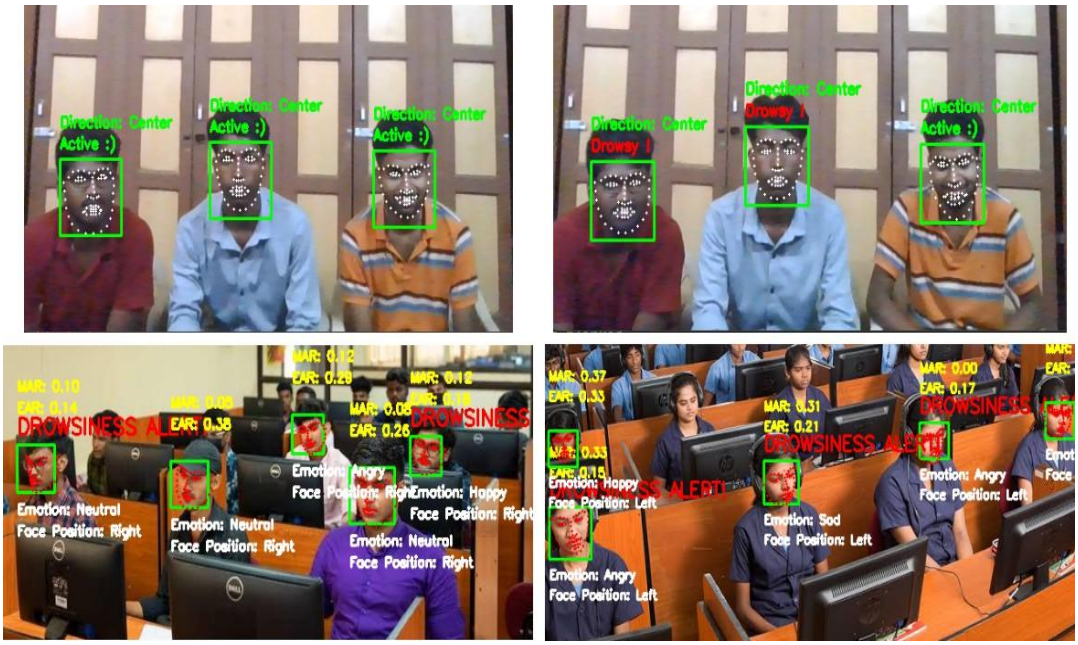

Fig. 8. Real-time monitoring of student engagement, emotions, and drowsiness detection in diverse classroom settings.

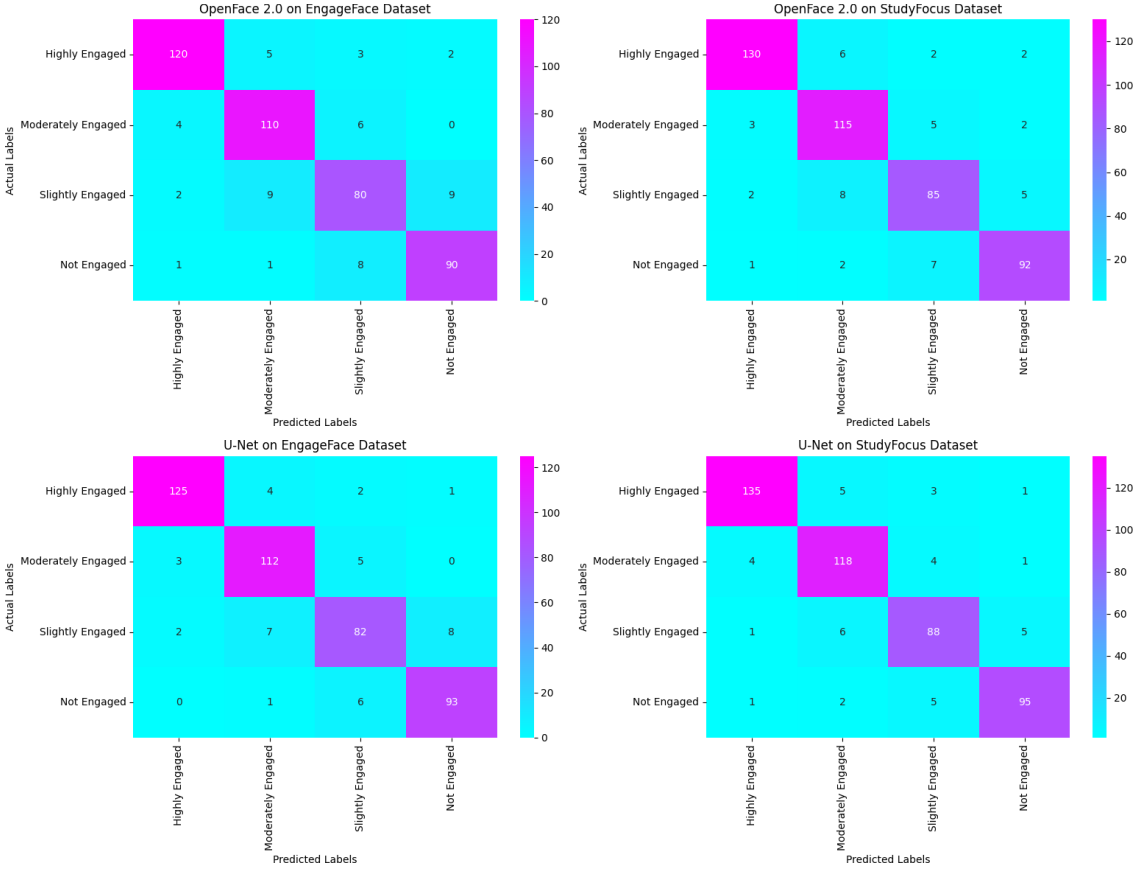

Fig. 9. Confusion Matrices for Proposed Systems

Table 10. Confusion Matrix for U-Net on EngageFace Dataset

Actual/Predicted	Highly Engaged	Moderately Engaged	Slightly Engaged	Not Engaged
Highly Engaged	125 (TP)	4 (FN)	2 (FN)	1 (FN)
Moderately Engaged	3 (FP)	112 (TP)	5 (FN)	0 (FN)
Slightly Engaged	2 (FP)	7 (FN)	82 (TP)	8 (FN)
Not Engaged	0 (FP)	1 (FP)	6 (FP)	93 (TN)

Table 11. Confusion Matrix for U-Net on StudyFocus Dataset

Actual/Predicted	Highly Engaged	Moderately Engaged	Slightly Engaged	Not Engaged
Highly Engaged	135 (TP)	5 (FN)	3 (FN)	1 (FN)
Moderately Engaged	4 (FP)	118 (TP)	4 (FN)	1 (FN)
Slightly Engaged	1 (FP)	6 (FN)	88 (TP)	5 (FN)
Not Engaged	1 (FP)	2 (FP)	5 (FP)	95 (TN)

5. Discussion

The engagement classification method evaluated in this study demonstrated consistent accuracy across two observational datasets, EngageFace and StudyFocus. Segmentation using the U-Net framework achieved accuracy rates of 97% and 95.5% on EngageFace and StudyFocus, respectively. These results suggest that the segmentation process was effective in isolating facial regions relevant to attentiveness observation under varied conditions. Additionally, the applied OpenFace-based method for facial landmark identification achieved accuracy rates of 95.2% and 93.5% on the same datasets, demonstrating stable feature tracking across sessions. The consistency was improved due to utilizing segmentation along with monitoring of local facial regions that provided useful cues such as gaze and eye openness for assessing the learner's attentiveness. When logged in real-time, these measures can help teachers grasp classroom dynamics and better identify instances of loss of focus during classes. Instructors can connect visual cues and student participation patterns with the use of standardized observational categories, particularly in blended/remote learning. The method applied here can also be taken to account for professionals training in isolation or in groups. In such cases, the uniform understanding of facial expressions may help support feedback systems for content delivery or instructional planning. The structured visual data can also produce summarized reports of learner responsiveness which may help improve the curriculum or implement learner support programs. This work can be extended to include additional indicators apart from the visual ones, such as witnessing the posture and the words used. Modifications for different lighting, head position, and bigger group scenarios could make it even more useful. Additionally, it is essential for wider application to align with privacy rules and ethical guidelines. In summary, the method shows encouraging results in evaluating engagement with visual input. The results show that attention levels can be consistently classified across settings, paving the way for future improvements in learning analytics in schools.

6. Conclusion

This study examined a visual monitoring approach for assessing student participation using facial region segmentation and landmark identification. Two datasets, EngageFace and StudyFocus, were used to evaluate the effectiveness of this method in classifying multiple levels of engagement. Segmentation using U-Net achieved an accuracy of 97% on EngageFace and 95.5% on StudyFocus, indicating consistent performance across different recording conditions. Landmark identification, implemented through a structured facial tracking approach, reached accuracy values of 95.2% and 93.5% on the same datasets, respectively. These results reflect the method's ability to capture relevant facial cues—such as gaze direction, eye openness, and head position—under varied classroom and home-based study environments. The findings suggest that observational data derived from visual inputs can support detailed assessments of learner attentiveness. By categorizing engagement into distinct levels and using consistent visual indicators, instructors may gain additional insight into classroom behavior patterns. This approach can be adapted to support in-person and remote instruction, as well as integrated into broader frameworks for learning analysis and instructional planning. Future work may focus on expanding the approach to accommodate more diverse environmental conditions and on integrating additional behavioral signals for a more holistic view of participation.

References

[1] C. Pabba and P. Kumar, "An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition," *Educational Systems Journal*, October 2021. DOI:10.1111/exsy.12839

[2] P. B. Jha, A. Basnet, B. Pokhrel, B. Pokhrel, G. K. Thakur, and S. Chhetri, "An Automated Attendance System Using Facial Detection and Recognition Technology," *Apex Journal of Business and Management*, vol. 1, no. 1, pp. 103-120, 2023.

[3] G. Tonguc¸ and B. Ozaydın Ozkara, "Automatic recognition of student emotions from facial expressions during a lecture," *Computers & Education*, 2019. DOI:10.1016/j.compedu.2019.103797

[4] O. Mohamad Nezami, M. Dras, L. Hamey, D. Richards, S. Wan, C. Paris, "Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression," in *Machine Learning and Knowledge Discovery in Databases*, ECML PKDD 2019. DOI:10.48550/arXiv.1808.02324
[5] S. Dev and T. Patnaik, "Student Attendance System using Face Recognition," in Proceedings of the 2020 In- ternational Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp. 90-96. DOI:10.1109/ICOSEC49089.2020.9215441
[6] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2132-2143, Oct.-Dec. 2022. DOI:10.1109/TAFFC.2022.3188390
[7] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detec- tion system in online learning context using deep learning models," *Multimedia Tools and Applications*, vol. 82, pp. 11365–11394, 2023. DOI:10.1007/s11042-022-13558-9
[8] S. Gupta, P. Kumar, and R. Tekchandani, "A multimodal facial cues based engagement detection system in e- learning context using deep learning approach," *Multimedia Tools and Applications*, vol. 82, pp. 28589–28615, 2023. DOI:10.1007/s11042-023-14392-3
[9] M. H. M. Kamil, N. Zaini, L. Mazalan, et al., "Online attendance system based on facial recognition with face mask detection," *Multimedia Tools and Applications*, vol. 82, pp. 34437–34457, 2023. DOI:10.1007/s11042-023-14842-y
[10] T. Potluri, V. S, and V. K. K. K, "An automated online proctoring system using attentive-net to assess student mischievous behavior," *Multimedia Tools and Applications*, vol. 82, pp. 30375–30404, 2023. DOI:10.1007/s11042- 023-14604-w
[11] R. Abdulkader, F. T. M. Ayasrah, V. R. G. Nallagattla, K. K. Hiran, P. Dadheech, V. Balasubramaniam, S. Sengan, "Optimizing student engagement in edge-based online learning with advanced analytics," *Array*, 2023. DOI:10.1016/j.array.2023.100301
[12] Y. Li, X. Qi, A. K. J. Saudagar, A. M. Badshah, K. Muhammad, and S. Liu, "Student behavior recognition for interaction detection in the classroom environment," *Image and Vision Computing*, 2023. DOI:10.1016/j.imavis.2023.104726
[13] W. E. Villegas-Ch, J. Garc´ıa-Ortiz, and S. Sa ńchez-Viteri, "Identification of Emotions From Facial Gestures in a Teaching Environment With the Use of Machine Learning Techniques," *IEEE Access*, vol. 11, pp. 38010-38022, 2023. DOI:10.1109/ACCESS.2023.3267007
[14] L. Sharara et al., "A Real-Time Automotive Safety System Based on Advanced AI Facial Detection Algorithms," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 6, pp. 5080-5100, June 2024. DOI:10.1109/TIV.2023.3272304
[15] M. N. Hasnine, H. T. Nguyen, T. T. T. Tran, H. T. T. Bui, G. Akc¸apınar, and H. Ueda, "A Real-Time Learning Analytics Dashboard for Automatic Detection of Online Learners' Affective States," *Sensors*, vol. 23, no. 9, article 4243, 2023. DOI:10.3390/s23094243
[16] B. Fang, X. Li, G. Han, and J. He, "Facial Expression Recognition in Educational Research From the Per- spective of Machine Learning: A Systematic Review," *IEEE Access*, vol. 11, pp. 112060-112074, 2023. DOI:10.1109/ACCESS.2023.3322454
[17] H. Farman, A. Sedik, M. M. Nasralla, and M. A. Esmail, "Facial Emotion Recognition in Smart Education Systems: A Review," presented at the 2023 IEEE International Smart Cities Conference (ISC2), Bucharest, Romania, 2023, pp. 1-9. DOI:10.1109/ISC257844.2023.10293353
[18] J. Xin-Ying Lek and J. Teo, "Academic Emotion Classification Using FER: A Systematic Review," 2023. DOI:10.1155/2023/9790005.
[19] Aanya Khan, Esther Alice Mathew, Junia Sam Dani, Giftlin Olivia, and T. S. Shivani. "Enhancing human behaviour analysis through multi-embedded learning for emotion recognition in images." In 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 331-336. IEEE, 2023.
[20] Abdelhadi, Z., Naseif, M., Alhejali, W., & Elhayek, A. (2025, January). TeacherEye: An AI-Powered System for Monitoring Student Engagement in Online Education. In 2025 22nd International Learning and Technology Conference (L&T) (Vol. 22, pp. 25-30). IEEE.

Authors' Profiles

Vidhya K. is an accomplished academic with a distinguished career as an Assistant Professor (Selection Grade) in the Division of Data Science and Cyber Security at Karunya Institute of Technology and Sciences. With an extensive academic background, she holds a Ph.D. from Anna University, Chennai, and has over 18 years of teaching experience complemented by industry experience. Dr. Vidhya is renowned for her research, particularly in applying deep learning techniques in healthcare. Her work has led to multiple publications in high-profile journals and conferences, focusing on cloud computing, computer vision, and IoT. Additionally, she has several patents and has developed innovative products like smart devices for healthcare and home water management.

T. M. Thiyagu an Associate Professor in the Department of Computer Science and Engineering (AI and ML) at Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, has a comprehensive educational background with a B.E degree from S.S.M College of Engineering (2009), an M.E degree from Anna University of Technology, Chennai (2012), and a Ph.D. from College of Engineering Guindy, Anna University, Chennai (2022). His research focuses on text-mining applications in machine learning, enriching the field with his academic insights.

Antony Taurshia is an Assistant Professor in the Department of Data Science and Cyber Security at Karunya Institute of Technology and Sciences. She is currently pursuing her Ph.D. at the same institution, having previously completed her M.Tech in 2013 and her B.E at Anna University in 2011. Her teaching focuses on network security and Internet of Things security, while her research interests lie in cyber security and IoT. Mrs. Taurshia has made notable contributions to her field, including publications on software-defined network-aided security solutions for IoT devices.

Jenefa A. serves as an Assistant Professor in the Division of Data Science and Cyber Security at Karunya Institute of Technology and Sciences. She completed her Ph.D. at Anna University, Chennai in 2022, and her research spans across artificial intelligence and computer networks. Dr. Jenefa has contributed extensively to the academic community through her publications in international journals and conferences, focusing on areas such as network traffic classification, advanced diagnostics using machine learning, and enhancing public safety with AI technologies. She is recognized for her pedagogical contributions, having taught a wide array of computer science courses, and has earned multiple accolades for her excellence in teaching and research.

How to cite this paper: Vidhya K., T. M. Thiyagu, Antony Taurshia, Jenefa A., "FocusTrack: Real-Time Student Engagement Monitoring via Facial Landmark Analysis", International Journal of Modern Education and Computer Science(IJMECS), Vol.18, No.1, pp. 76-92, 2026. DOI:10.5815/ijmecs.2026.01.05