Student Emotion Recognition System for Virtual Class Engagement Using Deep Learning and Multimodal Fusion

*P. Sai Leela[1], Shriraj Laxman Gore[2] and Bharath Setturu[3]**

[1]School of Engineering, Chanakya University, Bengaluru-562165
[2]School of Engineering, Chanakya University, Bengaluru-562165
[3]Associate Professor, School of Engineering, Chanakya University, Bengaluru-562165

[1]leelasai023@gmail.com, [2]goreshriraj@gmail.com, [3]bharath.s@chanakyauniversity.edu.in

Abstract

This paper proposes an AI-based Student Emotion Recognition and Engagement Monitoring System, which analyzes the facial emotions and other relevant factors to determine the level of student engagement during online classrooms. This system uses deep learning algorithms such as EfficientNet-B0 and MobileNetV3 to recognize facial emotions and then uses a regression-based fusion model to determine the level of student engagement by analyzing the facial features and other relevant factors such as sleep time, stress level, motivation level, and screen time. This system operates in the background and provides a summary of the level of engagement of students to the teacher only, indicating the level of class and student disengagement. From the experimental results, it is clear that the fusion of the proposed system enhances the robustness of the student engagement prediction system.

Keywords: *Student Engagement, Facial Emotion Recognition, Deep Learning, EfficientNet, MobileNetV3, Multimodal Fusion, Online Learning*

1. Introduction

The growth of virtual learning environments has created a greater number of opportunities for students to find their way into education but has also presented challenges for teachers by making it more challenging to gauge their students' attentiveness and emotions while physically present in the classroom. In traditional classrooms, teachers can use visual cues from students to assess whether students are engaged, bored, confused, etc., and adjust their teaching methods accordingly. Engagement assessment for students in online settings is often completed only at the end of a lesson, thereby limiting the amount of time available for pedagogical interventions to support student success.

Since emotional states like interest, frustration, and exhaustion have a direct impact on attention and comprehension, facial emotions are strongly associated with learning engagement [14], [5].Existing online platforms primarily rely on attendance records or post-class feedback, which fail to capture real-time emotional dynamics. Recent studies have demonstrated that facial expressions are reliable indicators of emotional engagement[14], [9] in synchronous online learning, while also emphasising that engagement is multidimensional [8] and cannot be fully represented by behavioural metrics alone.

Prior research has shown that lightweight deep learning architectures[13], [14], including EfficientNet and MobileNet variants, achieve high facial emotion recognition accuracy while remaining suitable for real-time deployment. However, many existing engagement detection systems either focus solely on emotion recognition or depend on complex multimodal setups that increase computational cost and reduce scalability. Moreover, facial expressions alone may not fully reflect a learner's engagement [8], [10] when influenced by factors such as sleep deprivation, stress, screen fatigue, or health conditions.

To tackle these limitations, this work introduces a **Student Emotion Recognition System for Virtual Class Engagement**. This system converts changes in facial expressions into ongoing engagement scores during live online classes.The system employs CNN-based models— EfficientNet-B0 and MobileNetV3—and integrates facial emotion features with contextual secondary data collected through a pre-class questionnaire. In addition to benchmark datasets, real online class videos are analysed frame-by- frame to correlate visual emotion patterns with contextual factors, enabling interpretable, teacher-oriented engagement summaries that explain not only *who* is disengaged, but also *why*.

Rather than reporting isolated emotion labels, the proposed system transforms facial emotion dynamics into acontinuous engagement score, enabling engagement to be measured as a behavioural trend over time.

2. Related Work

Past research in facial emotion recognition and engagement analytics have established that Convolutional Neural Networks (CNNs) are highly efficient in performing facial emotion recognition (FER) tasks [17], [18]. Models such as EfficientNet, MobileNet, ResNet, VGG, etc., have been successfully employed in transfer learning to enhance the efficiency of these models in performing well on the dataset for FER tasks [13], [14].

Recent research emphasises:

- Lightweight architectures for real-time deployment
- Multi-face detection in classroom environments
- Multimodal approaches combining visual and contextual features [18], [9]

However, many existing systems either focus solely on emotion recognition or lack interpretability for classroom deployment. The aim of this research is to develop a real-time, multimodal, teacher-friendly engagement monitoring system, which is still needed.

After a tremendous amount of research in the domain, Convolutional Neural Networks (CNNs) have been found to be the most widely used deep learning approach for the domain of facial emotion recognition (FER) [18], [17]. Transfer learning-based approaches have been found to yield improvements in the accuracy of emotion recognition using models like EfficientNet, ResNet, and MobileNet [13], [14] for scenarios with limited training data. Among these, EfficientNet models have achieved better results in terms of accuracy with fewer parameters using compound scaling techniques.

Several studies have focused on optimising **lightweight CNN architectures** for real-time FER[14]. Improved versions of MobileNetV3 incorporating attention mechanisms[18], dilated convolutions, and group normalisation have been shown to enhance feature localisation while maintaining low computational complexity. These characteristics make MobileNetV3 particularly suitable for live classroom monitoring systems where inference speed is critical.

Some recent studies have explored the possibility of mapping facial emotions to engagement levels to identify student interest, apart from emotion recognition [14], [9]. Some of these approaches rely on spatiotemporal models such as LSTM/GRU networks to track changes in emotions, while others rely on rule-based approaches to map identified emotions to engagement levels.These techniques enhance engagement prediction, but their real-time utility is limited by their frequent need for large amounts of video data and more processing power.

Multimodal engagement detection systems that combine facial expressions with eye gaze, head movement, or physiological signals [17] have reported higher prediction accuracy than emotion-only models. Nevertheless, such systems may suffer from increased hardware requirements and reduced robustness in unconstrained environments.

Additionally, empirical studies indicate that **facial expressions correlate most strongly with emotional engagement[8]**, while behavioural and cognitive engagement may require supplementary contextual information obtained through self-reports or surveys. These findings indicate that there is still a significant research gap for a lightweight, real-time, and comprehensible engagement monitoring system [14], [9] that incorporates contextual elements and strikes a balance between accuracy and computational efficiency. This work addresses that gap by combining EfficientNet-B0 and MobileNetV3 for facial emotion recognition and integrating secondary contextual data through a fusion-based engagement prediction framework.

3. Proposed System

3.1. System Overview

The proposed system is designed for real-time monitoring of online classes involving multiple students, capturing frames from live sessions, detecting student faces[13] , extracting emotional features, and computing continuous engagement scores. A teacher-only dashboard displays engagement trends and summaries.

A fusion-based regression model is used to estimate engagement and create continuous engagement scores that can be used for real-time classroom analytics[17]

The system operates on live online classes involving multiple students, enabling simultaneous face detection and engagement monitoring.

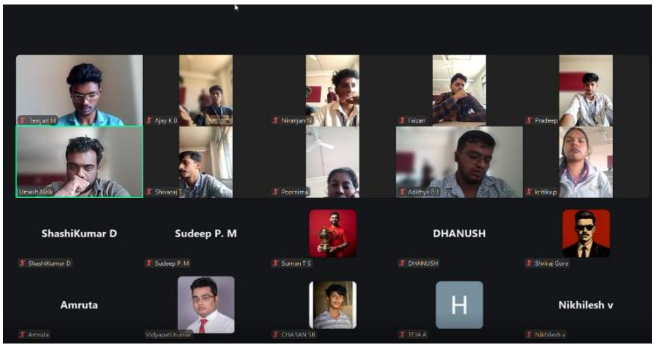

Fig. 1. Live online classroom with multiple students monitored simultaneously for engagement analysis.

3.2. System Workflow

1. Frame capturing from live online class videos
2. Multi-face detection through a deep learning-based detector
3. Feature extraction through EfficientNet-B0 and MobileNetV3
4. Real-time emotion prediction for each detected student
5. Student-wise emotion trends over time
6. Incorporation of pre-class secondary data collected through questionnaires
7. Fusion-based computation of engagement scores
8. Correlation of video-based emotion patterns and other contextual factors
9. Generation of reasoning-based engagement summary and alerts for teachers

All these activities take place behind the scenes without interrupting the class with frequent alerts and notifications.

There are two types of teacher feedback that the system supports, and the monitoring of student engagement takes place in the background. The first type of teacher feedback occurs when, at predetermined time intervals such as every 10 to 15 minutes, teacher-only student engagement notifications are sent to alert the teacher to the students whose student engagement level is consistently or constantly low. The second type of teacher feedback occurs towards the end of the session, or around the 40th minute in a 60-minute session, when a consolidated student engagement dashboard is displayed to the teacher.

Low-frequency engagement alerts are periodically displayed to instructors to indicate students with declining attention.

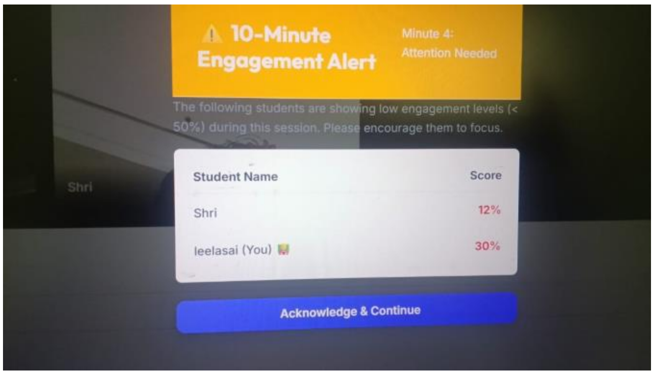

Fig. 2. Periodic teacher-only engagement alert highlighting students with low engagement during the session.

3.3. Emotion Recognition Models

• **EfficientNet-B0**: robust feature extraction with moderate complexity and high accuracy.

• **MobileNetV3:** Real-time inference optimized, lightweight.

Accuracy and computational efficiency are balanced thanks to the combination.

3.4. Multimodal Fusion

Visual features are fused with contextual parameters, including:

- Sleep duration
- Stress level
- Motivation
- Screen time
- Mood and health condition

A regression model outputs a **continuous engagement score[18], [9]**, enabling fine-grained monitoring instead of discrete labels.

4.Experimental Setup

4.1. Datasets

FER2013[1]
35,887 grayscale images (48×48), 7 emotion classes.

RAF-DB[2]
High-quality real-world facial expression dataset with improved diversity.

The datasets were combined to improve generalisation and robustness to real classroom conditions.

In addition to public datasets, real online class videos and associated secondary student data were used for system-level evaluation and engagement analysis.

4.2. Implementation Details

- Programming Language: Python
- Framework: PyTorch
- Libraries: OpenCV, NumPy, Pandas
- Hardware: GPU-enabled environment

Transfer learning was used for the CNN models. This is in line with the requirements for the use of modern deep learning techniques for a postgraduate research project.

Note: The CNN models use transfer learning. The CNN models were trained with a two-phase transfer learning protocol, and convergence was monitored with respect to the loss and accuracy metrics.

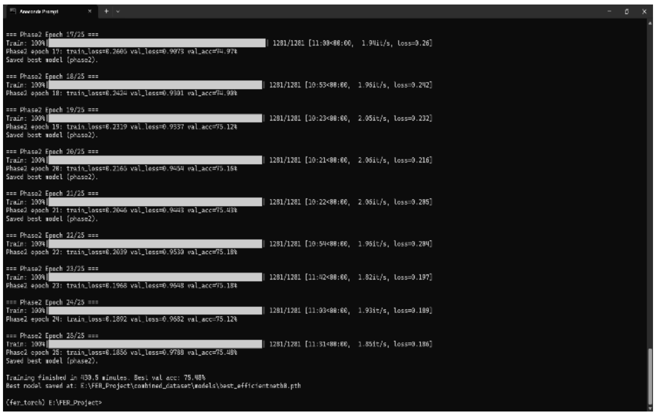

Fig. 3. Validation accuracy (post-model) and loss (post-model) have improved over time with systemic training on EfficientNet-B0 & MobileNetV3 Models throughout multiple epochs.

4.3. Class Video Data Collection Online

In addition to benchmark facial emotion datasets, the suggested system was evaluated in a real-world environment on online class videos collected from actual virtual classes conducted live using webcams and had no restriction on pose or control of lighting. The videos included all students attending the same class session, each having been acquired using a standard webcam. Each of the video streams was processed using OpenCV by extracting a frame from each stream at predetermined intervals.

The EfficientNet-B0 & MobileNetV3 Models trained for facial emotion recognition were then applied across all frames by passing each frame through a multi-facial detection pipeline to identify and classify the facial emotion of each student. Average engagement scores for each student were then obtained by averaging their respective facial emotion classifications over time. The way this was set up allows the system to emulate actual classroom conditions (e.g., student movement, lighting differences, and facial orientation differences).

4.4. Secondary Data Collection and Correlation

To enhance the interpretability of engagement predictions, **secondary contextual data** were collected from students via a brief questionnaire administered before **the online class**. The collected attributes included sleep duration, stress level, screen time prior to the session, willingness to attend, current mood, and health condition.

This secondary data was not used in direct relation to emotion classification, but was later correlated with video-based engagement scores during the analysis phase. Through the association of contextual factors with pre-class emotional trends in online class videos, the system was able to generate explanations for variations in engagement levels.

This correlation mechanism was useful in providing teacher-oriented summaries in instances where low engagement levels were being misinterpreted as student fatigue or stress, or overexposure to screens.

5.Results

5.1. Performance in Emotion Recognition

The EfficientNet-B0 network learned to recognize facial features successfully, while MobileNetV3 was able to be used for efficient and accurate real time inference. The models showed consistent performance for varying lighting conditions and facial orientations.

Figure 4 illustrates real-time facial emotion inference on live video frames using the trained CNN models.

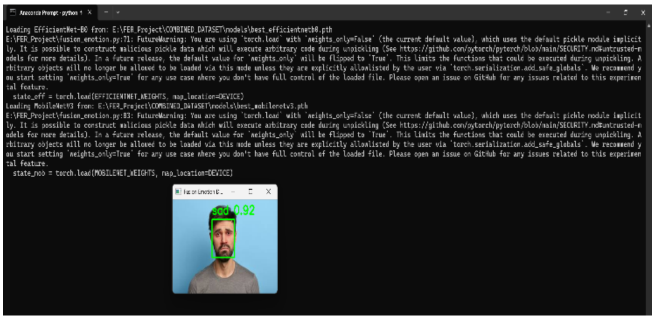

Fig. 4. Real-time facial emotion recognition output displaying detected face and predicted emotion confidence.

5.2. Engagement Score Regression Performance

The fusion-based engagement regression model was trained to predict continuous engagement scores by combining facial emotion features with temporal information. Engagement prediction was formulated as a regression task to reflect varying levels of student attention rather than discrete engagement categories.

A low MSE (Mean Squared Error) and low MAE (Mean Absolute Error) along with a very high R^2 score indicate this model is providing stable, accurate predictions of user engagement.As illustrated in Figure X, the regression model converged very well through the training epochs producing an MAE of 2.24 and an R^2 approaching 0.93. The low values of MAE indicate only slight differences between the predicted and actual engagement scores, while the relatively high R^2 suggests that the model accurately captures the meaningful trends of engagement in terms of facial emotion variability.

As MAE values decrease, engagement prediction accuracy increases, and the closer R^2 values get to 1, the more reliable the model.

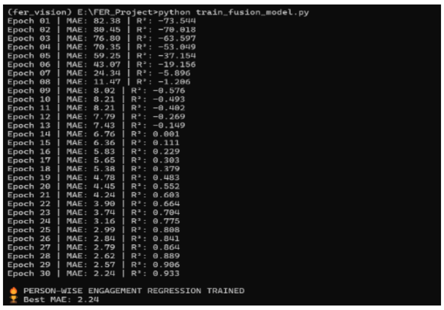

Fig. 5 Training performance of the engagement regression model showing MAE and R^2 values across epochs.

5.3. Real-Time Monitoring

The time series graphs of engagement demonstrated a declining trend in engagement over time, thus validating the practical application of the system.

5.4. Teacher Dashboard

A teacher-only summary includes:

- Overall class engagement
- Engagement levels for individual students
- Engagement distribution
- Possible disengagement causes

Instead of generating interruptions for the students, the system generates low-frequency engagement alerts for the teacher at set intervals during the class and a consolidated engagement dashboard toward the end of the class.

A consolidated engagement dashboard is presented toward the later stage of the class to summarise overall and individual engagement trends.

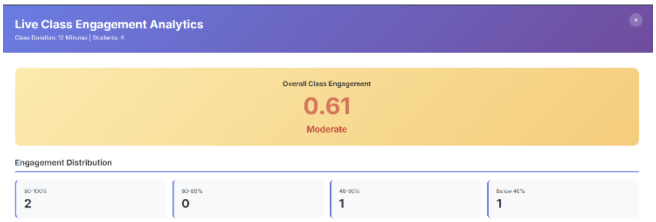

Fig. 6. Final engagement summary dashboard displaying overall class engagement, distribution, and student-wise analytics.

5.5. Engagement Reasoning and Teacher-Oriented Summary

The system generates a **teacher-only engagement summary** by jointly analysing facial emotion patterns extracted from **live online class videos** and **secondary contextual data provided by students before attending the session**. While emotion recognition identifies real-time engagement levels, the secondary data helps explain why certain students may appear disengaged.

For example, students who have continually low levels of engagement background scores have been linked to inadequate amounts of sleep prior to the class, high levels of stress before class, too much time spent on screens before class, or low levels of motivation in the pre-class questionnaire. When those correlations between different types of factors are made with trends in emotion appearance across video frames, the system gave reasonable explanations about the students rather than simply relying on the non-verbal behaviour alone to infer engagement values.

The combined analysis produces actionable items such as identifying students who need assistance and where they lost engagement periods during the period of instruction so that real-time directed support can be provided. The combination of having high-quality data and video capture to investigate variations in student engagement gives the instructor the ability to intervene appropriately and not just based solely on facial expressions.

Having data summaries, distribution, and trend reports and student specific analytics on student engagement provides instructors with a data-based decision support system versus a simple prediction engine. The "equivocal" dashboard also provides actionable items such as points in an instructional period where loss of engagement occurred and recommended instructional strategies to enhance engagement (e.g., quiz or discussion) to allow instructors to make informed real-time instructions. The system explains why students are exhibiting variations in engagement by correlating trends in emotion to student-established contextual variables recorded prior to the class.

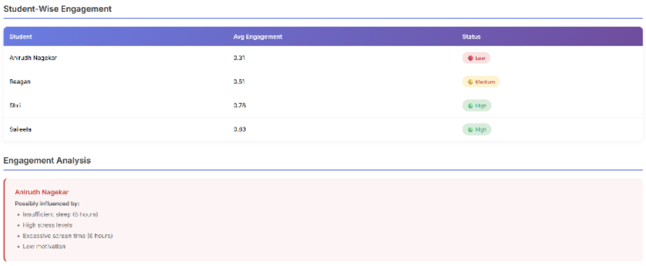

Fig. 7. Engagement reasoning panel correlating low engagement with secondary contextual factors such as sleep and stress.

Based on detected engagement patterns, the system generates actionable teaching recommendations.

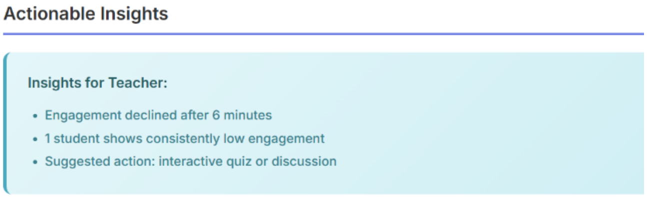

Fig. 8. Instructors receive useful insights. These insights highlight where student engagement drops off and suggest ways to improve it.

6. Discussion

A central contribution of this work is its focus on explaining engagement outcomes by correlating facial emotion trends with pre-class contextual factors, rather than merely predicting engagement scores.

The results reveal that multimodal fusion gives rise to increased reliability compared to emotive-focused systems, thanks in part to EfficientNet-B0 providing a wealth of feature data along with MobileNetV3 providing performance in real-time. The silent monitoring method minimizes disruptions and is beneficial. The limitations of this method include camera quality and lighting.

The future work involves head pose and eye tracking to improve multimodal analysis. Engagement analytics are limited to a teacher-only dashboard, which ensures student privacy and prevents performance pressure during live sessions.

Figure 9 shows how the entire developed system interface looks during an active monitoring session.

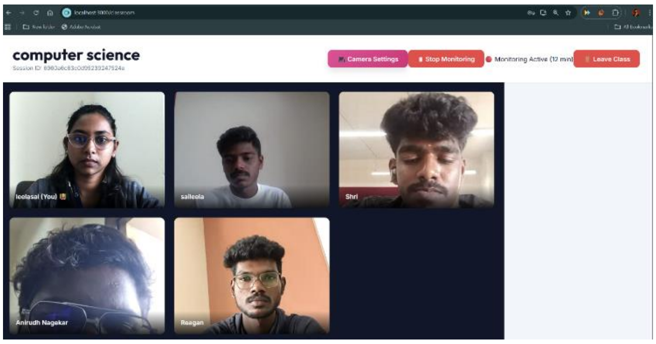

Fig. 9. Developed system interface showing real-time engagement monitoring in an online classroom.

7. Conclusion

This paper presented a deep learning–based **Student Engagement Monitoring System** that combines facial emotion recognition with contextual data fusion. The system also provides teachers with constant engagement scores and a summary. This is very useful in virtual classrooms. The experimental results of AI-based engagement analytics in virtual education are effective and efficient. The results show the potential of the system. By combining the results of the facial emotion analysis of the videos of the online classes with the pre-class results of the secondary students, the system can provide clear and actionable insights. This supports better decision-making for instruction.

The system works in real-time, supports multiple students, uses lightweight models, monitors quietly, and gives clear outputs. This functionality makes the system suitable for use in real-world environments, by providing regular engagement alerts along with an overall dashboard that can promote early intervention and allow more detailed examination following a session.

8. Acknowledgment

The authors thank the School of Engineering at Chanakya University for providing resources and academic support.

9. References

9.1. Journal Articles

[1] I. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D. Lee, Y. Zhou and Y. Bengio, “Challenges in Representation Learning: A Report on Three Machine Learning Contests”, Neural Networks, vol. 64, (2015), pp. 59–63.

[2] S. Li and W. Deng, “Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition”, IEEE Transactions on Image Processing, vol. 28, no. 1, (2019) January, pp. 356–370.

[3] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks”, IEEE Signal Processing Letters, vol. 23, no. 10, (2016) October, pp. 1499–1503.

[4] J. Whitehill, Z. Serpell, Y. Lin, A. Foster and J. R. Movellan, “The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions”, IEEE Transactions on Affective Computing, vol. 5, no. 1, (2014) January–March, pp. 86–98.

[5 S. Bosch and S. K. D’Mello, “The Affective Experience of Learning: Using Emotion Recognition to Improve Instruction”, Learning and Instruction, vol. 45, (2017), pp. 1–12.

[6] R. Jain, A. Aggarwal and P. Kumar, “Facial Expression Recognition Using Transfer Learning with EfficientNet”,Multimedia Tools and Applications, vol. 80, (2021), pp. 1–18.

[7] Y. Liu, H. Zhang and J. Wang, “Facial Expression Recognition Based on Improved MobileNetV3 Network”, IEEE Access, vol. 10, (2022), pp. 43678–43688.

[8] X. D’Mello and A. Graesser, “Dynamics of Affective States During Learning Activities”, Educational Psychology Review, vol. 24, no. 2, (2012), pp. 145–167.

[9] Z. Bosch, K. Chen and S. K. D’Mello, “Detecting Student Engagement from Facial Expressions in Online Learning Environments”, Computers & Education, vol. 163, (2021).

[10] T. Miller, “Explanation in Artificial Intelligence: Insights from the Social Sciences”, Artificial Intelligence, vol. 267, (2019), pp. 1–38.

[11] S. K. D’Mello, A. Olney, C. Williams and P. Hays, “Gaze Tutor: A Gaze-Reactive Intelligent Tutoring System”, International Journal of Human-Computer Studies, vol. 70, no. 5, (2012), pp. 377–398.

9.2. Book

[12] P. Ekman and W. V. Friesen, “Facial Action Coding System: A Technique for the Measurement of Facial Movement”, Consulting Psychologists Press, (1978).

9.3. Conference Proceedings

[13] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”,Proceedings of the International Conference on Machine Learning (ICML), (2019), pp. 6105–6114.

[14] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le and H. Adam, “Searching for MobileNetV3”, Proceedings of the IEEE International Conference on Computer Vision (ICCV), (2019), pp. 1314–1324.

[15] J. Redmon, S. Divvala, R. Girshick and A. Farhadi,“You Only Look Once: Unified, Real-Time Object Detection”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016), pp. 779–788.

[16] R. Gupta, K. Audhkhasi, S. Narayanan and A. V. Dhall, “DAiSEE: Towards User Engagement Recognition in the Wild”, Proceedings of the ACM International Conference on Multimedia, (2016), pp. 1–5.

[17] G. Siemens, “Learning Analytics: Envisioning a Research Discipline and Its Application”, Proceedings of the International Conference on Learning Analytics & Knowledge (LAK), (2013), pp. 1–10.

[18] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel and J. Movellan, “Recognising Facial Expression: Machine Learning and Application to Spontaneous Behavior”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (2005).