IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

A Survey of Live Engagement Detection Using Real-Time Video Processing

[1] Aadhil Muhammed, [2]Adil Bin Anwar, [3]Amith Kurian Joseph, [4] Arun Kumar S, [5] Swathi S

[5] Assistant Professor, [1,2,3,4] Student
[1,2,3,4,5] Department of Computer Science,
[1,2,3,4,5] College of Engineering Karunagappally, Kerala, India

Abstract :
The rapid growth of online education has necessitated the development of tools to monitor and enhance student engagement during virtual classes. This study focuses on predicting student engagement levels using video data from the DAiSEE dataset, which captures students' facial expressions and behaviors. By analyzing these visual cues, the system aims to classify engagement into three distinct categories are Engaged high, Engaged low, and Engaged not listening. The primary objective is to provide real-time feedback to educators, enabling them to address disengagement and improve the overall learning experience.
The proposed system leverages a combination of deep learning and machine learning techniques. Video frames are preprocessed and fed into a pre-trained EfficientNetB0 model for feature extraction, generating a 1280-dimensional feature vector for each frame. These features are then used to train a custom neural network with a fully connected architecture, consisting of an input layer, a hidden layer with 1024 units, and an output layer with softmax activation for multi-class classification. The dataset is balanced using SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance. The entire pipeline is integrated with the Jitsi Meet platform, enabling real-time engagement prediction during online classes.
The model achieved higher accuracy, demonstrating its effectiveness in classifying student engagement levels. These results indicate that the system can reliably identify engagement patterns and provide actionable insights to educators. By integrating this solution into video conferencing platforms, educators can proactively address disengagement, fostering a more interactive and productive learning environment. This study highlights the potential of AI-driven tools in revolutionizing online education and improving student outcomes.

Keywords : **Real Time Engagement Detection, CNN, ResNet, Jitsi Meet, SMOTE, EfficientNetB0, Online Learning, Facial Feature Analysis, DAiSEE, Deep Learning, Machine Learning, Real time**

INTRODUCTION

The shift to online education has transformed the way students and educators interact, bringing both opportunities and challenges. One of the most significant challenges is maintaining student engagement in a virtual environment, where traditional cues like body language and eye contact are harder to interpret. Engaged students are more likely to retain information, participate actively, and achieve better learning outcomes. However, disengagement due to boredom, confusion, frustration can hinder the learning process. To address this, there is a growing need for intelligent systems that can automatically monitor and analyze student engagement in real-time, providing educators with actionable insights to improve the learning experience.

Recent advancements in artificial intelligence and computer vision have opened new possibilities for analyzing student behavior through video data. By leveraging facial expressions and other visual cues, AI-driven systems can classify engagement levels and identify patterns that may indicate disengagement. The DAiSEE dataset, which contains labeled video data of students in online learning environments, serves as a valuable resource for developing such systems. This study focuses on building a robust engagement prediction model using this dataset, with the goal of classifying students into three engagement categories are Engaged high, Engaged low, and Engaged not listening. The system aims to provide real-time feedback to educators, enabling them to intervene promptly and foster a more interactive and productive learning environment.

The proposed system integrates deep learning techniques with video conferencing platforms like Jitsi Meet, making it a practical tool for online education. By combining feature extraction using a pre-trained EfficientNetB0 model with a custom neural network for classification, the system can analyze student engagement in real time. This approach not only addresses the challenges of virtual learning but also highlights the potential of AI-driven tools to enhance educational outcomes.

A. Class Categorization

This logic categorizes engagement into three main classes based on the intensity levels of four engagement-related attributes: Boredom, Engagement, Confusion, and Frustration. If a student has a high Engagement score (≥2) while keeping the other three emotions low (≤1), they are classified as Engaged high, indicating strong focus.

If Engagement is low (≤1) and at least one of Boredom, Confusion, or Frustration is high (≥2), they are classified as Engaged low, representing disengagement.

If Engagement is high (≥2) but at least one of the other three emotions is also high (≥2), they are labeled Engaged not listening, implying divided attention.

In all other cases, the default classification is Engaged low to ensure every sample gets assigned a category.

Table I Distribution of labels in DAiSEE across its affective states

Affective State	Very Low	Low	High	Very High
Boredom	3869	2931	1934	334
Confusion	6024	2191	752	101
Engagement	61	459	4477	4071
Frustration	6986	1649	346	87

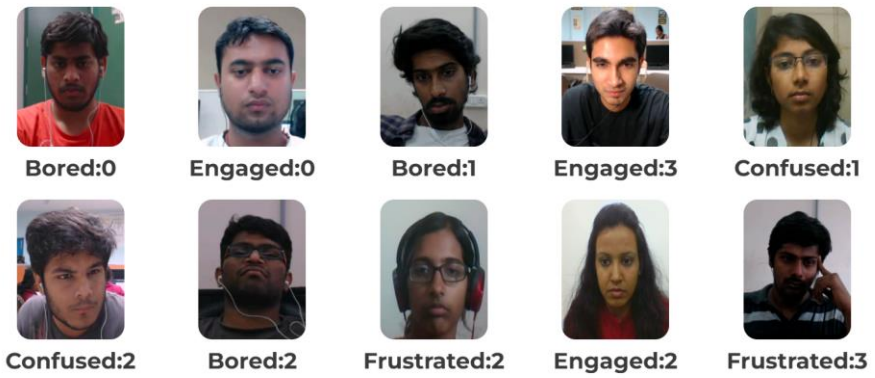

Figure 1 : DAISEE Sample Pictures

B. SMOTE (Synthetic Minority Over-sampling Technique)

Balances the dataset by generating synthetic samples for the minority classes instead of simply duplicating existing ones. It works by identifying the nearest neighbors of a minority class sample and creating new samples along the line segments connecting them in feature space. Here, after encoding the class labels, SMOTE resamples the dataset to ensure that all classes have an approximately equal number of samples, improving model training by reducing class imbalance and preventing the model from being biased toward majority classes.

Table II Before SMOTE

Class	Values
Engaged high	10134
Engaged not listening	3831
Engaged low	655

Table III After SMOTE

Class	Values
Engaged high	10134
Engaged not listening	10134
Engaged low	10134

III. LITERATURE REVIEW

A. Automatic detection of students engagement during online learning: A bagging ensemble deep learning approach

Mayanda Mega Santoni, T. Basaruddin, Kasiyah Junus, and Oenardi Lawanto [1] conducted a study that aimed at improving the detection of participation in real-time on platforms for online learning. Monitoring student participation and engagement has faced challenges, especially in the video-based virtual classrooms. The researchers used a combination of CNNs and architecture of ResNet to exploit facial and body movement for engagement analysis. Their method utilized advanced feature extraction and dimensionality reduction techniques with a dataset of actual real-time video feeds through online sessions. The current model had an accuracy of 97.87%, which was higher compared to existing benchmarks such as traditional models based on SVM. However, some limitations were identified based on high computational demands and lack of balance in the current datasets. This study shows CNN-ResNet's huge potential in addressing the engagement detection challenge in e-learning environments.

B. Assessing student engagement from facial behavior in on-line learning

Paolo Buono et al. [2] studied the engagement of students in online learning via facial behavior analysis. The research examined the problem of automatically forecasting engagement in personalising e-learning environments using LSTM networks. The method used included facial action units, head pose estimation, and gaze tracking to forecast engagement. The model utilized an EmotiW 2019 Challenge dataset and experimental data drawn from 30 undergraduate students with mean squared error (MSE) of 0.077 on the test set that was better than baseline on this test set, mainly under engaging conditions, especially video lectures, underpinning strong correlations of the facial behavior with the levels of perceived emotional engagement. Despite promising results, limitations included a small sample size and variability in subjective engagement levels, indicating the need for further exploration with larger datasets and physiological data integration.

C. Cognitive and emotional engagement while learning with VR: The perspective of multimodal methodology

Ilana Dubovi [3] conducted a study examining cognitive and emotional engagement during virtual reality (VR)-based learning in nursing education. The study addressed the challenge of capturing engagement holistically by combining psycho-physiological measures such as facial expressions, eye tracking, and electrodermal activity (EDA) with self-reports. Sixty-one nursing students participated in a pretest/posttest design using a VR simulation for medication administration. Findings indicated that engagement varies across learning phases, with joy dominating during procedural tasks and cognitive metrics like reduced blink rates signaling high mental effort. The combined modalities explained 51\% of learning outcomes, underlining the emotional and cognitive interplay that characterizes learning achievements. The study illustrates VR's potential to boost procedural and declarative learning while leaving room for further exploration in larger samples and diverse methodologies.

D. Students engagement level detection in online e-learning using hybrid efficientnetb7 together with TCN, LSTM, and Bi-LSTM

Selim et al. [4] carried out a study aimed at the detection of students' engagement levels during online learning. The goal was to address the challenge of automation in the detection of engagement within dynamic e-learning environments. Three hybrid models combining EfficientNetB7 with TCN, LSTM, and Bidirectional LSTM have been proposed by the researchers for spatiotemporal feature extraction and classification. Using the newly introduced VRESEE dataset, comprising 3,525 video snippets from Egyptian students, and the public DAiSEE dataset, their models were able to achieve superior accuracies of 94.47\% and 67.48\%, respectively, outperforming existing benchmarks such as ResNet+TCN. Key innovations included the integration of EfficientNetB7 for efficient spatial feature extraction and advanced temporal modeling techniques, which significantly improved performance despite challenges like dataset imbalance. The study highlights the potential of hybrid deep learning models in addressing engagement detection in online education.

E. Deep learning approach based on dimensionality reduction for designing electromagnetic nanostructures

Kiarashinejad et al. [5] conducted a study focusing on designing and optimizing electromagnetic (EM) nanostructures using deep learning (DL). The study dealt with the challenge of computational inefficiency during the analysis and design of complex

nanostructures, mainly in applications such as metasurfaces optical modulation. The researchers employed dimensionality reduction using autoencoders to convert the problem of many-to-one optimization to a simpler one-to-one type. Their approach used forward and inverse neural network designs with a dataset produced through finite element simulations of optical metasurfaces. The proposed model managed to achieve efficient computational performance by reducing the dimensions of the design and response spaces, overcoming the limitations of traditional brute-force methods. However, some limitations, such as mild non-uniqueness in reduced spaces, are present, but the study demonstrated the potential of DL-based methods in addressing complex nanophotonic design problems while providing insights into light-matter interactions.

F. MDNN: Predicting student engagement via gaze direction and facial expression in collaborative learning

Yi Chen et al. [6] developed a model to predict the engagement of students in collaborative learning environments, mitigating issues such as non-invasive assessment and group interaction. The authors adopted an MDNN that combines facial expression recognition and gaze direction analysis. Their approach used facial expression detection by MTCNN and gaze estimation through a bidirectional LSTM network for evaluating engagement levels. Using classroom video data with 8 cameras, the MDNN obtained a prediction accuracy of 78\%, outperforming models based on gaze alone (74\%) or facial expressions alone (47\%). This innovation reduced dependency on costly gaze-tracking hardware while enhancing engagement evaluation accuracy. Despite its success, some limitations such as occlusions and individual behavior variability were noted. The study highlights MDNN's potential in advancing real-time engagement analysis for collaborative learning.

G. Student engagement detection using emotion analysis, eye tracking, and head movement with machine learning

Prabin Sharma et al. [7] presented a study that emphasized the detection of student engagement through real-time emotion analysis, eye tracking, and head movement. The study addressed the issue of measuring student attention in e-learning, especially in virtual classroom settings. The authors used a system that integrates Haar Cascade Algorithm for facial and eye detection and CNNs for emotion classification and engagement prediction. A total of 15 students captured with their own laptop web cameras formed the dataset used for computing a concentration index from the integral approach combining dominant emotion probability and movement analysis. A highly accurate and correlated model that categorized engagement into "very engaged," "nominally engaged," and "not engaged at all" was achieved in the proposed model. Limitations include occlusion of faces and fluctuating performance due to different external variables such as glasses. The results confirm the feasibility of applying webcam-based systems to monitor and adapt learning experiences in a dynamic way.

H. Integration of EfficientNetB0 and Machine Learning for Fingerprint Classification

Jenan A. Alhijaj and Raidah S. Khudeyer [8] conducted a study focusing on fingerprint classification by integrating EfficientNetB0 for deep feature extraction with machine learning classifiers. The research aimed to improve classification accuracy and efficiency by leveraging pre-trained deep learning models alongside traditional classification techniques. The authors used EfficientNetB0 to extract high-dimensional feature representations from fingerprint images, which were then refined using Principal Component Analysis (PCA) to reduce feature size while preserving critical information. These optimized feature vectors were subsequently fed into a Random Forest (RF) classifier, achieving an outstanding 99.91\% accuracy, outperforming previous fingerprint classification models. The study highlighted the benefits of combining deep learning-based feature extraction with classical machine learning algorithms, demonstrating superior efficiency and classification performance. However, potential limitations include dataset-specific dependencies and the need for further testing on larger and more diverse datasets. The findings emphasize the effectiveness of EfficientNetB0-based feature extraction in biometric authentication systems, setting a benchmark for future research in fingerprint recognition.

I. Comparative approach for facial expression recognition in higher education using hybrid-deep learning from students

Muhammed Usame Abdullah and Ahmet Alkan [9] conducted a study that focused on recognizing students' facial expressions to improve online education. The study solved the issue of monitoring engagement in virtual lectures by using facial expressions in real-time. They applied transfer learning with six pre-trained convolutional neural networks, namely AlexNet, GoogleNet, ResNet18, ResNet50, MobileNetV2, and VGG16, with K-fold cross-validation to classify eight expressions, among which "attention" was a new category for measuring engagement. Their dataset included 6720 images of 70 university students, augmented to balance eight expression categories. ResNet18 achieved the highest accuracy of 99.8\% and F1-score of 99\% compared to existing models. This work introduced custom creation of a dataset and fine-tuned CNNs to adapt facial expression recognition, surmounting the limitations as such expressions are imbalanced and there is hardware confinement. Such findings show possibilities of using CNN-based methodologies to improve feedback and adaptiveness in an e-learning environment.

J. Automatic recognition of student engagement using deep learning and facial expression

Omid Mohamad Nezami et al.[10] designed a study into the use of deep learning and facial expressions for automatic identification of student engagement. Such research challenges the problematics of recognizing engagement in a learning context, especially intelligent tutoring systems. The researchers proposed a two-step deep learning model, which initially used pretraining on the FER-2013 dataset for basic facial expression recognition and then fine-tuned it further on a newly collected ER dataset. This approach

also used CNNs and techniques like data augmentation and transfer learning. ER had 4,627 labeled images of engaged and disengaged students captured in the process of interaction with VLE. The proposed model achieved a classification accuracy of 72.38\% on the test set, outperforming baseline methods such as HOG+SVM and traditional CNNs, demonstrating the benefits of leveraging pre-trained facial expression features. Despite its effectiveness, the study acknowledged challenges like annotator subjectivity and data imbalance. This work underscores the potential of deep learning in enhancing adaptive educational technologies.

K. Enhancing SMOTE for imbalanced data with abnormal minority instances

Surani Matharaarachchi, Mike Domaratzki, and Saman Muthukumarana [11] proposed a study focusing on enhancing the Synthetic Minority Over-sampling Technique (SMOTE) for imbalanced datasets containing abnormal minority instances, such as outliers. The paper introduced four novel SMOTE extensions Distance ExtSMOTE, Dirichlet ExtSMOTE, FCRP SMOTE, and BGMM SMOTE which leverage weighted averages of neighboring instances to generate more representative synthetic samples and mitigate the impact of outliers. These methods were evaluated using simulated and real-world datasets, with performance metrics including F1 score, PR-AUC, and MCC. The results demonstrated that Dirichlet ExtSMOTE outperformed most existing SMOTE variants, particularly in improving classification performance for imbalanced datasets with abnormal instances. The study highlighted the importance of addressing class imbalance and the challenges posed by outliers in minority classes, offering robust solutions for real-world applications such as medical diagnosis and fraud detection. However, the paper acknowledged limitations in handling extremely sparse datasets and the computational complexity of some proposed methods. The findings underscore the effectiveness of advanced SMOTE extensions in enhancing classifier performance and provide valuable insights for future research in imbalanced data handling.

Table IV Summary Of Engagement Detection Studies

DL Method and Ref.	Dataset Used	Pros	Cons	Performance Metrics
CNN + ResNet [1]	Real-time video feeds from online sessions	High accuracy (97.87%),effective feature extraction	Computational demands, imbalanced datasets	Accuracy:97.87%
LSTM Networks [2]	EmotiW 2019 Challenge dataset, data from 30 students	Captures temporal engagement, correlates facial behavior with engagement	Small sample size, variability in engagement levels	Mean Squared Error (MSE):0.077
Multimodal (Facial expressions, EDA, eye tracking) [3]	VR simulation data from 61 nursing students	Captures both cognitive and emotional engage-ment, explains 51% of learning outcomes	Requires diverse data collection tools, limited scalability	Explained variance in learning outcomes:51%
EfficientNetB7 + TCN, LSTM, Bi-LSTM [4]	VRESEE dataset (3,525 video snippets), DAiSEE dataset	High spatiotemporal feature extraction, robust performance	Dataset imbalance, resource-intensive computation	Accuracy: 94.47% (VRE-SEE), 67.48% (DAiSEE)
Autoencoders + Neural Networks [5]	Finite element simulations of optical metasurfaces	Simplifies optimization problem, efficient computation	Non-uniqueness in reduced spaces	Computational efficiency improvement
MDNN (MTCNN + Bidirectional LSTM) [6]	Classroom video data with 8 cameras	Reduces hardware dependency, improves engagement prediction accuracy	Occlusions, individual behavior variability	Accuracy:78%
Haar Cascade + CNNs [7]	Dataset of 15 students captured via webcams	Effective emotion-based engagement classification	Face occlusion, external variables like glasses	Correlation between engagement and emotion classification
EfficientNetB0 + PCA + Random Forest [8]	Fingerprint dataset (various sources)	High classification accuracy (99.91%), reduces feature size with PCA	Dependency on dataset quality, computational cost	Accuracy: 99.91%

Transfer Learning (ResNet18, AlexNet, VGG16) [9]	6720 images of 70 students	Fine-tuned CNNs, custom dataset creation	Hardware constraints, imbalanced expressions	Accuracy:99.8% (ResNet18)
Two-Step CNN Model [10]	FER-2013 and Engagement Recognition (ER) dataset	Combines pre-trained features with fine-tuned models	Annotator subjectivity, data imbalance	Accuracy: 72.38%
SMOTE Extensions (Distance ExtSMOTE, Dirichlet ExtSMOTE, FCRP SMOTE, BGMM SMOTE) [11]	Simulated and real-world datasets (medical, fraud)	Improves performance, handles outliers	Computational complexity, sparse datasets	PR-AUC, MCC with Dirichlet ExtSMOTE

III. EXPECTED RESULTS

The proposed system is expected to achieve reliable and accurate classification of student engagement levels into three categories are Engaged high, Engaged low, and Engaged not listening. By leveraging the DAiSEE dataset and employing a pre-trained EfficientNetB0 model for feature extraction, followed by a custom neural network for classification, the system is anticipated to demonstrate strong performance in predicting engagement. Initial experiments indicate that the model achieves a training accuracy of 80%, showcasing its ability to learn patterns from the data, while maintaining a robust testing accuracy, ensuring generalization to unseen data. The integration of SMOTE (Synthetic Minority Over-sampling Technique) is expected to address class imbalance, further enhancing the model's ability to classify underrepresented engagement categories. Ultimately, the system aims to provide real-time, actionable insights to educators during online classes, enabling them to identify disengaged students and take timely corrective actions, thereby improving the overall learning experience.

IV. CONCLUSION

In conclusion, this study presents an innovative AI-driven system for monitoring and predicting student engagement in online learning environments. By leveraging the DAiSEE dataset and employing a combination of deep learning techniques such as feature extraction using EfficientNetB0 and classification with a custom neural network the system effectively categorizes engagement into three levels are Engaged high, Engaged low, and Engaged not listening. The integration of SMOTE ensures balanced representation of all engagement categories, enhancing the model's performance. Initial results, including a training accuracy of 80%, demonstrate the system's potential to provide reliable and actionable insights for educators.

The real-time integration of this system with Jitsi Meet highlights its practical applicability in online education. By enabling educators to identify disengaged students and intervene promptly, the system fosters a more interactive and productive learning environment. This work underscores the transformative potential of AI-driven tools in addressing the challenges of virtual learning and improving educational outcomes. Future research could explore advanced architectures, larger datasets, and real-world deployment to further enhance the system's accuracy and scalability. Overall, this study contributes to the growing field of educational technology, offering a scalable solution to enhance student engagement in online classrooms.

REFERENCES

[1] M. M. Santoni, T. Basaruddin, K. Junus, and O. Lawanto, "Automatic detection of students' engagement during online learning: A bagging ensemble deep learning approach," IEEE Access, vol. 12, pp. 9606396073, 2024.

[2] P. Buono, B. De Carolis, F. D'Errico, et al., "Assessing student engagement from facial behavior in on-line learning," Multimed Tools Appl,vol. 82, pp. 12859–12877, 2023.

[3] I. Dubovi, "Cognitive and emotional engagement while learning with vr:The perspective of multimodal methodology," Computers Education,vol. 183, p. 104495, 2022.

[4] T. Selim, I. Elkabani, and M. A. Abdou, "Students engagement level detection in online e-learning using hybrid efficientnetb7 together with tcn, lstm, and bi-lstm," IEEE Access, vol. 10, pp. 99573–99583, 2022.

[5] Y. Kiarashinejad, S. Abdollahramezani, and A. Adibi, "Deep learning approach based on dimensionality reduction for designing electromagnetic nanostructures," npj Comput Mater, vol. 6, p. 12, 2020.

[6] Y. Chen, J. Zhou, Q. Gao, and W. Zhang, "Mdnn: Predicting student engagement via gaze direction and facial expression in collaborative learning," Computer Modeling in Engineering Sciences, vol. 136, no. 1, pp. 382–401, 2023.

[7] P. Sharma, S. Joshi, S. Gautam, S. Maharjan, S. R. Khanal, M. C. Reis, J. Barroso, and V. M. d. J. Filipe, "Student engagement detection using emotion analysis, eye tracking, and head movement with machine learning," in Proceedings, University of Massachusetts Boston, USA,2022.
[8] J. A. Alhijaj and R. S. Khudeyer, "Integration of efficientnetb0 and machine learning for fingerprint classification," Informatica (Slovenia), vol. 47, no. 5, 2023.
[9] M. U. Abdullah and A. Alkan, "A comparative approach for facial expression recognition in higher education using hybrid-deep learning from students' facial images," Traitement du Signal, vol. 39, no. 6, pp. 1929–1941, 2022.
[10] O. M. Nezami, M. Dras, L. Hamey, D. Richards, S. Wan, and C. Paris, "Automatic recognition of student engagement using deep learning and facial expression," arXiv, 2019
[11] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Enhancing smote for imbalanced data with abnormal minority instances," Machine Learning with Applications, vol. 18, p. 100597, 2024.
[12] A. Gupta, A. D'Cunha, K. N. Awasthi, and V. N. Balasubramanian, "Daisee: Towards user engagement recognition in the wild.," arXiv: Computer Vision and Pattern Recognition, 2016.
[13] B. Bekta¸s and E. Dandıl, "Development of android based mobile video conference application using jitsi meet platform," 06 2021.