# A Deep Learning Approach for Automatic Detection of Learner Engagement in Educational Context

Asma Ayari[1,2][a], Mariem Chaabouni[1][b] and Henda Ben Ghezala[1][c]
[1]*Univ. Manouba, ENSI, RIADI LR99ES26, Campus Manouba, 2010, Tunisia*
[2]*Esprit School of Engineering, Tunis, Tunisia*

Keywords: Technology-Enhanced Learning, Education, e-Learning, Learner's Engagement, Engagement Detection, Artificial Intelligence, Deep Learning, CNN, VGG-16.

Abstract: The rise of online learning modalities, including fully online, hybrid, hy-flex, blended, synchronous, and asynchronous formats, has transformed educational landscapes. However, assessing learners' engagement in these environments, where direct teacher-student interaction is limited, poses a significant challenge for educators. In this context, Artificial Intelligence (AI) emerges as a transformative force within education, leveraging advanced algorithms and data analysis to personalize learning experiences and enhance teaching methodologies. The detection of engagement in educational settings is critical for evaluating the effectiveness of instructional strategies and fostering student participation. This study presents an approach to assess the learner's engagement through detecting the facial emotions and classifying the level of engagement. This approach proposes a deep learning model specifically designed for automatic engagement detection in educational environments, employing a Convolutional Neural Network (CNN) approach. We introduce an optimized CNN model tailored for recognizing learner engagement through facial expressions in distance learning contexts. By integrating the foundational elements of traditional CNN architectures with the widely acclaimed VGG-16 model, our approach harnesses their strengths to achieve exceptional performance. Rigorous training, testing, and evaluation on an augmented dataset demonstrate the efficacy of our model, which significantly surpasses existing methodologies in engagement recognition tasks. Notably, our approach achieves an accuracy of 94.10% with a loss rate of 10.39%, underscoring its potential to enhance the assessment of learner engagement in online education.



## 1 INTRODUCTION

In the teaching and learning process, learner engagement plays a central role, as it has a major influence on the effective progress of activities and the acquisition of learning outcomes. Classified as a multifaceted construct, including behavioral, emotional, and cognitive engagement (Fredricks et al. 2004), it needs to be closely analyzed, particularly for e-learning environments. Therefore, the evaluation of learner engagement is multidimensional, involving behavioral dimension (class participation or time spent on an online platform), emotional dimension (enthusiasm, interest, frustration, reactivity) and cognitive dimension (initiative-taking, critical thinking).

Several definitions of engagement have been proposed in scientific works: learner engagement was defined as: "the interaction between the time, effort and other relevant resources invested by both learners and their institutions intended to optimize the learner experience and enhance the learning outcomes and development of learners and the performance, and reputation of the institution" (Trowler, 2010). According to Kuh (2009), "Engagement premise is straightforward and easily understood: the more learners study a subject, the more they know about it, and the more learners practice and get feedback from faculty and staff members on their writing and collaborative problem solving, the deeper they come to understand what they are learning".

Often associated with higher academic performance and increased knowledge retention,

[a] https://orcid.org/0000-0003-4722-328X
[b] https://orcid.org/0009-0000-1345-4042
[c] https://orcid.org/0000-0002-6874-1388



Ayari, A., Chaabouni, M. and Ben Ghezala, H.
A Deep Learning Approach for Automatic Detection of Learner Engagement in Educational Context.
DOI: 10.5220/0013283200003932

In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 1*, pages 372-379
ISBN: 978-989-758-746-7; ISSN: 2184-5026

engagement is known to be one of the key qualitative indicators. According to D'Errico et al. (2016), engagement can influence attitudes and motivation, strengthening perseverance and resilience in the face of learning difficulties. This information about the learner engagement will help the teachers to make the e-Learning environment more interactive and adaptable according to the needs of the learners. In online learning, direct teacher-learner interaction is limited. This poses a significant challenge for educators. Engagement in online settings can be experienced differently from traditional in-class environments. Studies indicate that how teachers perceive student engagement not only influences their interactions with those students but can also impact student performance and grades (Bergdahl, 2022). Although, it remains difficult to quantify and to observe this dimension in a direct and objective manner.

In this context, institutes, educators, and researchers are constantly seeking innovative methods to enhance learner engagement and improve learning outcomes. Engagement detection methods can be broadly categorized into three main types: manual, semi-automatic, and automatic. Manual engagement detection methods rely on human observation, interaction, or subjective reporting to assess learner engagement, involving direct human involvement in the data collection and analysis process. Examples include Direct Observation and Self-Report Surveys (Larson & Richards, 1991) (Shernoff et al., 2000) (O'Brien & Toms, 2010), (Grafsgaard et al., 2012). Semi-automatic methods combine elements of manual observation with automated data collection or analysis techniques, aiming to leverage both human judgment and technological capabilities (D'Mello et al., 2014) (Nezami et al. 2017) (Qi et al, 2023).

In contrast, automatic engagement detection methods primarily rely on technology and computational algorithms to assess learner engagement without direct human intervention. These methods analyse data streams, signals, or inputs to automatically infer engagement levels. Automatic methods are further categorized into computer vision-based methods, sensor data analysis, and log-file analysis, depending on the type of information processed for engagement detection.

Computer vision-based methods are categorized into three main groups: facial expression analysis, gesture and posture recognition, and eye movement tracking (Murshed et al, 2019), (Nguyen et al., 2019), (Zhang et al., 2020), (Toti et al. 2021), (Sharma et al. 2022), (Gupta et al. 2023). Certain research endeavors integrate multiple modalities to enhance accuracy. Artificial Intelligence has emerged as a powerful tool in this pursuit, offering the capability to analyze learner engagement and tailor educational experiences to individual needs. Machine learning algorithms, such as deep learning, have gained significant attention in the field of education due to their potential in enhancing learner engagement and improving learning outcomes.

This paper aims to contribute to this context of research by presenting an engagement detection approach based on an optimized deep learning model, specifically a Convolutional Neural Network-based approach, for the automatic detection of engagement in educational settings. In this paper, we focus on the emotional dimension and more precisely on the detection of emotions through the analysis of learners' faces during a synchronous learning session and their classification according to the degree of engagement.

The proposed deep learning approach aims to address the need for more innovative methods in enhancing learner engagement and improving learning outcomes in educational contexts.

The remainder of the paper is structured as follows. Section 2 reviews prior studies on engagement detection in distance education. Section 3 outlines the CNN models examined in this research, including the proposed model for detecting learner engagement. Section 4 details the methodology and experimental setup, while Section 5 analyses the results. Finally, Section 6 concludes the paper and suggests potential directions for future related research.

## 2 LITERATURE REVIEWS

The primary objectives of e-learning are to enhance learners' skills, address their individual needs, and streamline the learning process. To achieve these goals, employing adaptive learning systems is essential, as learners vary significantly in cognitive and physical characteristics. In a previous paper (Ayari et al. 2023), we have introduced a new learner model comprising two main categories of dimensions: (1) static dimensions, which include demographic information, interests, background, experience, learning style preferences, and learning context; and (2) dynamic dimensions, encompassing learning situation, materials, emotions, motivation, engagement, and interaction.

Building on this previous work, the present paper specifically examines the engagement dimension, a crucial factor impacting the quality of online education. Learner engagement reflects the effort a

student invests to maintain psychological commitment to the learning process. Various methods are employed to measure this parameter. In our research, we concentrate on the application of deep learning models to detect engagement within educational settings.

The use of deep learning has significantly enhanced the ability to predict various parameters, including learner engagement. Notably, some models developed as early as 1943 continue to effectively address contemporary challenges. This enduring success has attracted considerable interest from researchers, leading to the adoption of a range of machine learning techniques tailored to meet diverse educational needs, with a primary focus on improving learning outcomes. Numerous studies have demonstrated the precision with which these methods can identify and interpret patterns of engagement, further establishing their relevance in educational settings.

Recent research on engagement detection has largely concentrated on identifying the best-performing deep learning architectures, such as CNNs, LSTMs, and hybrid models. The effectiveness of these models largely hinges on their ability to detect key features, especially in facial expression recognition, which plays a crucial role in assessing engagement. Facial expressions such as eye gaze, blinking, and changes in mouth and eyebrow positioning are strong indicators of attentiveness, confusion, or boredom. By capturing these features in real-time, these models can more precisely gauge learner engagement.

For instance, Dewan, et al. (2019) compared different CNN architectures, including All-CNN, Very Deep Convolutional Neural Networks (VD-CNN), and Network-in-Network (NiN-CNN), with a proposed CNN model for predicting learners' engagement. The analysis involved utilizing a dataset of learner face images to discern facial expressions indicative of engagement. Through this comparative study, the authors sought to identify the most effective CNN architecture for engagement prediction tasks. In comparison to the three types of CNNs, the proposed model achieves an accuracy of 92.33%, demonstrating its effectiveness.

In another study, Chen, Zhang, Wang, and Li (2021) introduced the Multimodal Deep Neural Network (MDNN) algorithm to predict learners' engagement. This algorithm combines the capabilities of CNNs and Long Short-Term Memory (LSTM) networks to process multimodal data, such as images of learners. By leveraging both CNN and LSTM architectures, the proposed MDNN algorithm demonstrates promising results in engagement prediction. The proposed model achieves an accuracy of 74% in the evaluation results.

Similarly, Liao, Wang, and Wu (2021) utilized the LSTM algorithm in conjunction with a variant of CNN known as the Squeeze and Excitation Network (SENet) for predicting learner engagement. Their approach involves analyzing sequences of photos extracted from videos to infer engagement levels. By incorporating SENet alongside LSTM, the authors aimed to enhance the model's ability to capture nuanced features relevant to engagement prediction. This model achieves a classification accuracy of 58.84%.

Bajaj et al. (2022) implemented a hybrid neural network architecture combining ResNet and a temporal convolutional network (TCN) to classify learner engagement, achieving a recognition accuracy of 53.6%.

In another approach, Mehta et al. developed a 3D DenseNet Self-Attention neural network (3D DenseAttNet) to automatically detect learner engagement in online learning environments. The model leverages the 3D DenseNet block to selectively extract high-level intra-frame and inter-frame features from video data. It achieved a recognition accuracy of 63.59% on the DAiSEE dataset.

Gupta et al. (2023) introduced a deep learning method that analyzes facial emotions to assess learner engagement in real time during online learning. The system uses a Faster R-CNN (region-based convolutional neural network) (Hai & Guo, 2020) for face detection and a modified face-points extractor (MFACXTOR) to identify key facial features. To determine the most effective model for accurate classification of real-time learner engagement, they tested various deep learning architectures, including Inception-V3 (Szegedy et al., 2016), VGG19 (Simonyan & Zisserman, 2014), and ResNet-50 (He et al., 2016). Their results showed that the system achieved accuracies of 89.11% with Inception-V3, 90.14% with VGG19, and 92.32% with ResNet-50 on their custom dataset.

Building on the advancements in facial emotion analysis for engagement detection, Ahmad et al. (2023) further explored the potential of deep learning architectures by utilizing the lightweight MobileNetv2 model to automatically assess learner engagement. The architecture was fine-tuned to improve learning efficiency and adaptability, with the final layer modified to classify three distinct output classes instead of the original 1000 classes from ImageNet. Their experiments, conducted using an open-source dataset of individuals watching online course videos,

compared the performance of MobileNetv2 against two other pre-trained networks, ResNet-50 and Inception-V4. MobileNetv2 outperformed both models, achieving an average accuracy of 74.55%.

Lastly, Ikram et al. (2023) introduced an improved transfer learning method that employs a modified VGG16 model, which includes an additional layer and finely tuned hyperparameters. This model was specifically designed to evaluate learner engagement in a minimally controlled, real-world classroom environment with 45 learners. In assessing learner engagement levels, the model achieved remarkable results, attaining an accuracy of 90%.

Overall, these studies underscore the growing interest in leveraging deep learning methodologies, particularly CNNs, for detecting and analyzing engagement dynamics in educational contexts. By harnessing the power of deep learning algorithms and multimodal data processing, researchers aim to develop robust and accurate models capable of understanding and predicting learner engagement levels effectively.

# 3 ENGAGEMENT DETECTION USING CNN MODEL

The following sections will introduce the fundamental CNN model, the VGG16 model and our proposed model.

## 3.1 Fundamental CNN Model

Convolutional neural networks (CNNs) represent a subset of artificial neural networks renowned for their ability to discern patterns within grid-like data structures, such as images. Constituting a generic CNN architecture are fundamental components including an input layer, multiple hidden layers, and an output layer. Within the hidden layers, a diverse array of elements can be found, including convolutional layers, activation layers, pooling layers, normalization layers, and fully connected layers.

## 3.2 VGG16 Model

The VGG16 model, a deep convolutional neural network, has garnered considerable attention in the field of computer vision. Originating from the Visual Graphics Group at the University of Oxford, it stands out for its straightforward yet powerful approach to image recognition tasks. Its strength lies in its capacity to extract intricate features directly from raw input images. By using many convolutional layers, the model can capture hierarchical representations of the input data, which enables it to recognize and classify a wide range of objects and patterns.

The VGG-16 architecture comprises 16 weight layers, encompassing 13 convolutional layers and 3 fully connected layers. While the initial layers capture general features, the subsequent layers focus on specifics, necessitating a transition from the general to the specific somewhere within the network (Yosinski et al., 2014). To mitigate overfitting, a pre-trained strategy is employed, leaving some initial layers untouched while training the final layers (Yamashita et al., 2018). The initial layers primarily handle convolution or feature extraction, while the ultimate layer is dedicated to classification. Notably, the last three layers consist of fully connected layers with Rectified Linear Unit (ReLU) activation functions. Customizing these final layers is essential to enhance performance. Below is the pseudocode delineating the modifications required for the last three layers.

## 3.3 Proposed Model

Our proposed model is a hybrid architecture that combines a custom convolutional neural network (CNN) and a pre-trained VGG16 model, aimed at improving image classification accuracy through effective feature extraction.

The custom CNN processes input images through three convolutional blocks, employing increasing filter sizes (32, 64, 128) and max-pooling layers with a pool size of 2x2. This structure facilitates the extraction of hierarchical features, with each block designed to capture different levels of abstraction. The VGG16 model, pre-trained on the ImageNet dataset, serves as a robust feature extractor, leveraging its extensive learned representations from diverse natural images. The choice of VGG16 is justified by its proven efficacy in image classification tasks, allowing the model to benefit from established feature representations.

Both sets of features are concatenated and processed through a dense layer with 512 units and ReLU activation, enabling the model to learn complex feature interactions. The final output layer consists of a dense layer with a SoftMax activation function for accurate multi-class classification.

The model was trained using the Adam optimizer, starting with a learning rate of 1e-4 to ensure stable convergence. We utilized categorical cross-entropy loss over 50 epochs with a batch size of 32 to effectively balance training efficiency and resource management. Early stopping was implemented to

mitigate overfitting by continuously monitoring validation loss, thereby enhancing the model's generalization capability.

The relevance of coupling a custom CNN with a pre-trained VGG16 model lies in the complementary strengths each offers for analyzing learner engagement in educational settings. The pre-trained VGG16 model brings the advantage of leveraging features learned from a vast and diverse dataset (ImageNet), which includes general image patterns that can be useful for identifying facial expressions. These features help recognize nuanced facial expressions that may reflect learner engagement or lack thereof.

On the other hand, the custom CNN is designed to learn specific features from the educational dataset at hand, which may contain unique patterns related to the specific educational context. For example, it can fine-tune its filters to better recognize patterns in learner facial expressions that are directly tied to engagement during learning activities, beyond the more general image recognition capabilities of VGG16.

By combining these two approaches, the architecture benefits from both general image feature extraction (via VGG16) and task-specific learning (via the custom CNN). This allows the model to more effectively analyze and classify levels of learner engagement, particularly when dealing with diverse or subtle variations in engagement across learners. This combination improves generalizability while still being sensitive to the nuances of learner interaction in educational environments.

# 4 METHODOLOGY AND EXPERIMENTS

The following sections will detail the architecture of the proposed model, describe the dataset used for engagement recognition, outline the model training process, and present the model evaluation.

## 4.1 Proposed Model Architecture

This architecture allows the model to leverage both the learned features from the pre-trained VGG16 model and the features learned directly from the input images by the custom CNN branch, potentially improving its performance on your classification task.

The custom CNN consists of three convolutional blocks. The first block applies to 32 filters, the second 64 filters, and the third 128 filters. Each convolutional layer uses a 3x3 kernel size, followed by a ReLU activation function to introduce non-linearity. Max-pooling layers with a 2x2 pool size are applied after each convolution to reduce spatial dimensions, controlling overfitting and computational complexity while preserving important features. The flattened outputs of the custom CNN and the pre-trained VGG16 model are concatenated for further processing. This basic CNN architecture efficiently extracts hierarchical, low-to-high-level features, enabling it to capture patterns crucial for image classification.

The model comprises several convolutional layers (Conv2D) and max-pooling layers (MaxPooling2D), organized into blocks denoted as "block1," "block2," and so forth. Each convolutional layer extracts features from the input data, while max-pooling layers reduce spatial dimensions to control overfitting and computational complexity. The "flatten" layers transform the multi-dimensional feature maps into one-dimensional arrays for further processing. Additionally, a concatenation layer merges the flattened outputs of two distinct pathways within the network. The fully connected layers (Dense) at the end of the architecture perform classification based on the learned features.

The presented architecture showcases a deep learning methodology for both feature extraction and classification, making it particularly well-suited for image recognition tasks. The model was trained using the Adam optimizer with an initial learning rate of 1e-4, and the categorical cross-entropy loss function was used to handle the multi-class classification. Training was performed over 50 epochs with a batch size of 32, and early stopping was employed to avoid overfitting by monitoring the validation loss.

Figure 1 represents our proposed deep learning architecture.

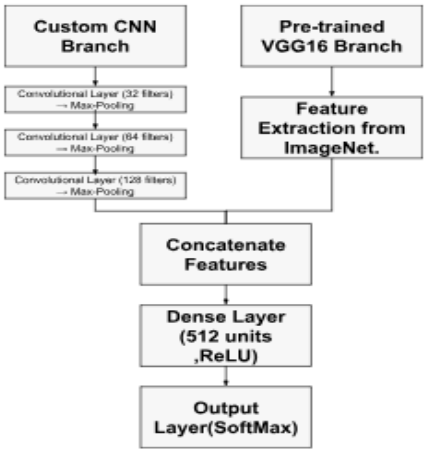


Figure 1. The structure of the proposed model for engagement detection

### 4.2 Description of the Dataset Engagement Recognition

The recognition of the necessity for extensive, well-categorized, openly accessible datasets for the purposes of training, assessing, and setting benchmarks has been widespread. Consequently, numerous initiatives have been undertaken in recent years to fulfil this requirement.

The authors of this paper use two types of datasets: original dataset and augmented dataset. The dataset is an open-source collection titled "Student Engagement." It includes data related to learner participation in academic activities, focusing on various features of learner engagement. Designed for researchers and educators to study engagement patterns and improve educational outcomes. The dataset consists of 2,120 photos of learners (both girls and boys) captured during online classes, with the aim of predicting engagement. The photos are categorized into six classes: looking away, bored, confused, drowsy, engaged, and frustrated. All images are resized to 200x200 pixels to optimize model processing time. Table 1 displays the dataset's organization.

Table 1: The organization of our dataset.

| Engaged | Not Engaged | Total |
|---|---|---|
| Class 1: Engaged 347<br>Class 2: Confused 369<br>Class 3: Frustrated 360 | Class 4: Bored 358<br>Class 5: Drowsy 263<br>Class6: Looking away 423 | 2120 |

The authors of this paper apply the technique of data augmentation. This method helps increase the diversity and variability of the dataset by applying transformations like rotation, flipping, scaling, cropping, or adding noise. This variety can assist models in better generalizing to new, unseen data. The table 2 outlines the transformation undergone by a dataset through augmentation techniques for training a model. Initially comprising 2,120 images across 6 classes, the dataset is expanded exponentially during training epochs, reaching approximately 67,840 images per epoch through augmentation methods like rotation, shift, shear, zoom, and flip. This augmentation process contributes to a more comprehensive understanding of the data's variability and enhances the model's ability to generalize. Despite the increase in dataset size, the image dimensions remain unchanged at 200 x 200 pixels, maintaining consistency. With a batch size doubled to 64 and an extended training duration of 500 epochs, the augmented dataset offers a rich pool of over 6 million samples, all while ensuring class balance and maintaining data normalization. This augmentation strategy enriches the dataset's diversity, ultimately fostering a robust and adaptable model.

## 5 RESULTS AND DISCUSSION

The following sections present the results of image data augmentation, as well as the accuracy and loss of our model.

### 5.1 Results of Image Data Augmentation

In an efficient deep learning model, the validation error should consistently decrease alongside the training error, and data augmentation stands as a potent technique to mitigate overfitting (Shorten & Khoshgoftaar, 2019).

### 5.2 Accuracy and Loss

The accuracy of our proposed model on the validation dataset is approximately 94.10%. This means that our model correctly classified 94.10% of the images in the validation set.

The loss of our model on the validation dataset is approximately 0.1039. It represents how well our proposed model's predictions match the true labels in the validation set.

In Table 2, the authors of this paper present a comparison between the engagement prediction models used in the above-mentioned research and our proposed model which is the combination of CNN model and VGG16.

Table 2: Comparison between accuracy of the engagement prediction models and our proposed model.

| Model used | Input data | Accuracy (%) |
|---|---|---|
| CNN | Learners' photos | 92.33 |
| All-CNN | Learners' photos | 75.97 |
| NiN-CNN | Learners' photos | 83.22 |
| VD-CNN | Learners' photos | 86.45 |
| VGG-16 | Learners' photos | 85 |
| SENet (CNN + LSTM) | Learners' photos | 59 |
| Xception | Learners' photos | 88 |
| Multimodal Deep Neural Network (MDNN) (CNN + LSTM) | Learners' photos | 74 |

Table 2: Comparison between accuracy of the engagement prediction models and our proposed model (cont.).

| Model used | Input data | Accuracy (%) |
|---|---|---|
| LSTM + Squeeze and Excitation Network (SENet) | Learners' photos | 58.84 |
| ResNet + Temporal Convolutional Network (TCN) | Learners' photos | 53.6 |
| 3D DenseNet Self-Attention Neural Network (3D DenseAttNet) | Learners'videos | 63.59 |
| MobileNetv2 | Learners'videos | 74.55 |
| Modified VGG16(Transfer Learning) | Learners'videos | 90 |
| Proposed Model (CNN+VGG-16) | Learners's photos | 94,10 |

# 6 CONCLUSION AND FUTURE WORK

The objective of this paper is to address the challenge that teachers face in accurately and efficiently detecting learner engagement in online learning environments. To address this, we introduce a new CNN model that combines two popular architectures the basic CNN model and the VGG-16 model specifically designed for detecting learner engagement.

Our experiments, conducted on a student engagement dataset featuring six levels of learner engagement (engaged, confused, frustrated, bored, drowsy and looking), demonstrated a high accuracy of 94.10% in engagement classification, outperforming most previous works using the same input. We can conclude that the current research results are satisfactory. This study provides valuable perspectives on automating the detection of student engagement in online learning environments. Researchers may explore avenues for optimizing our proposed model in future experiments conducted within this domain. This could include changing the architecture of our models and improving other metrics. Furthermore, conducting a comparative analysis against alternative models may provide valuable insights into the efficacy of our approach.

The authors of this paper are integrating the proposed model within an e-learning application to enable to exploit the real-time detection of learner engagement and adapt the learning experience. This practical application will further validate the effectiveness of our model in educational contexts.

# ACKNOWLEDGEMENTS

We extend our gratitude to all individuals involved in the preparation and review of earlier iterations of this document.

# REFERENCES

Ahmad, N., Khan, Z., Singh, D. (2023). Student engagement prediction in MOOCs using deep learning. In 2023 International Conference on Emerging Smart Computing and Information (ESCI), IEEE, pp. 1–6.

Ayari, A., Chaabouni, M., & Ben Ghezala, H. (2023). New learner model for intelligent and adaptive e-learning system. *2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA)*, 1-7. https://doi.org/10.1109/INISTA59065.2023.10310323

Bajaj, K. K., Ghergulescu, I., Moldovan, A. N. (2022). Classification of student affective states in online learning using neural networks. In 2022 17th International Workshop on Semantic Social Media Adaptation and Personalization (SMAP), IEEE, pp. 1–6.

Bergdahl, N. (2022). Engagement and disengagement in online learning. *Computers & Education*, *188*, 104561.

Chen, X., Zhang, Y., Wang, L., Li, T. (2021). Multimodal Deep Neural Network (MDNN) algorithm to predict students' engagement. In *Journal of Educational Data Science*, vol. 5, no. 2, pp. 45–58.

Dewan, A., Murshed, M., Lin, F. (2019). Engagement detection in online learning: A review. In *Smart Learn. Environ.*, vol. 6, pp. 1–17.

D'Errico, F., Paciello, M., Cerniglia, L. (2016). When emotions enhance students' engagement in e-learning processes. In Journal of e-Learning and Knowledge Society, vol. 12, no. 4.

D'Mello, S., Lehman, B., Pekrun, R., Graesser, A. (2014). Confusion can be beneficial for learning. In *Learn. Instr.*, vol. 29, pp. 153–170.

D'Mello, S. K., Craig, S. D., Sullins, J., Graesser, A. C. (2006). Predicting affective states expressed through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue. In *Int. J. Artif. Intell. Educ.*, vol. 16, no. 1, pp. 3–28.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research*, *74*(1), 59-109.

Gupta, S., Kumar, P., Tekchandani, R. K. (2023). Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. In Multimedia Tools and Applications, vol. 82, no. 8, pp. 11365–11394. DOI: 10.1007/s11042-022-13558-9.

Grafsgaard, J. F., Fulton, R. M., Boyer, K. E., Wiebe, E. N., Lester, J. C. (2012). Multimodal analysis of the implicit affective channel in computer-mediated textual

communication. In ACM Int. Conf. on Multimodal Interaction, California, USA, pp. xx–xx.

Hai, L., Guo, H. (2020). Face detection with improved face R-CNN training method. In Proceedings of the 3rd International Conference on Control, Computer Vision, pp. 22–25.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Ikram, S., et al. (2023). Recognition of student engagement state in a classroom environment using deep and efficient transfer learning algorithm. In Applied Sciences, vol. 13, no. 15, pp. 8637. DOI: 10.3390/app13158637.

Kuh, G. D. (2009). The national survey of student engagement: Conceptual and empirical foundations. In New Directions for Institutional Research, vol. 2009, no. 141, pp. 5–20.

Larson, R. W., Richards, M. H. (1991). Boredom in the middle school years: blaming schools versus blaming students. In *Am. J. Educ.*, vol. 99, no. 4, pp. 418–443.

Liao, Y., Wang, J., Wu, Y. (2021). Deep Facial Spatio-temporal Network (DFSTN) for predicting student engagement based on facial expressions and spatio-temporal features. In *Applied Intelligence*, 2021, pp. 287–303. DOI: 10.1007/s10489-020-01995-1.

Mehta, N. K., Prasad, S. S., Saurav, S., Saini, R., Singh, S. (2022). Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement. In Applied Intelligence, vol. 52, no. 12, pp. 13803–13823. DOI: 10.1007/s10489-022-03200-4.

Murshed, M., Dewan, M. A. A., Lin, F., & Wen, D. (2019, August). Engagement detection in e-learning environments using convolutional neural networks. In 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (pp. 80-86). IEEE.

Nezami, O. M., Richards, D., & Hamey, L. (2017). Semi-supervised detection of student engagement.

Nguyen, Q. T., Binh, H. T., Bui, T. D., & NT, P. D. (2019, December). Student postures and gestures recognition system for adaptive learning improvement. In 2019 6th NAFOSTED Conference on Information and Computer Science (NICS) (pp. 494-499). IEEE.

O'Brien, H. L., Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. In *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 1, pp. 50–69.

Qi, Y., Zhuang, L., Chen, H., Han, X., & Liang, A. (2023). Evaluation of Students' Learning Engagement in Online Classes Based on Multimodal Vision Perspective. Electronics, 13(1), 149.

Shorten, C., Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. In *Journal of Big Data*, vol. 6, pp. 1–48.

Sugden, N., Brunton, R., MacDonald, J., Yeo, M., Hicks, B. (2021). Evaluating student engagement and deep learning in interactive online psychology learning activities. In Australas. J. Educ. Technol., vol. 37, pp. 45–65.

Sharma, P., Joshi, S., Gautam, S., Maharjan, S., Khanal, S. R., Reis, M. C., ... & de Jesus Filipe, V. M. (2022, August). Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In International Conference on Technology and Innovation in Learning, Teaching and Education (pp. 52-68). Cham: Springer Nature Switzerland.

Toti, D., Capuano, N., Campos, F., Dantas, M., Neves, F., & Caballé, S. (2021). Detection of student engagement in e-learning systems based on semantic analysis and machine learning. In Advances on P2P, Parallel, Grid, Cloud and Internet Computing: Proceedings of the 15th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2020) 15 (pp. 211-223). Springer International Publishing.

Trowler, V. (2010). Student engagement literature review. In The Higher Education Academy, vol. 11, no. 1, pp. 1–15.

Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. In arXiv preprint arXiv:1409.1556.

Shernoff, D. J., Csikszentmihalyi, M., Schneider, B., Shernoff, E. S. (2000). Student engagement in high school classrooms from the perspective of flow theory. In *Sociol. Educ.*, vol. 73, pp. 247–269.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818–2826.

Xu, R., Li, C., Paterson, A. H., Jiang, Y., Sun, S., Robertson, J. S. (2018). Aerial images and convolutional neural network for cotton bloom detection. In *Frontiers in Plant Science*, vol. 8, pp. 1–17

Yosinski, J., Clune, J., Bengio, Y., Lipson, H. (2014). How transferable are features in deep neural networks? In *arXiv preprint arXiv:1411.1792*.

Yamashita, R., Nishio, M., Do, R. K. G., Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. In *Insights Into Imaging*, vol. 9, pp. 611–629. [CrossRef] [PubMed]

Zhang, Z., Li, Z., Liu, H., Cao, T., & Liu, S. (2020). Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology. Journal of Educational Computing Research, 58(1), 63-86.