Journal on
Artificial Intelligence

Tech Science Press

DOI: 10.32604/jai.2024.048911

ARTICLE

Detection of Student Engagement in E-Learning Environments Using EfficientnetV2-L Together with RNN-Based Models

Farhad Mortezapour Shiri[1,*], Ehsan Ahmadi[2], Mohammadreza Rezaee[1] and Thinagaran Perumal[1]

[1]Faculty of Computer Science and Information Technology, University Putra Malaysia (UPM), Serdang, Malaysia

[2]Department of Electrical and Computer Engineering, University of Wisconsin, Madison, USA

*Corresponding Author: Farhad Mortezapour Shiri. Email: GS63904@student.upm.edu.my

Received: 21 December 2023 Accepted: 04 March 2024 Published: 24 April 2024

ABSTRACT

Automatic detection of student engagement levels from videos, which is a spatio-temporal classification problem is crucial for enhancing the quality of online education. This paper addresses this challenge by proposing four novel hybrid end-to-end deep learning models designed for the automatic detection of student engagement levels in e-learning videos. The evaluation of these models utilizes the DAiSEE dataset, a public repository capturing student affective states in e-learning scenarios. The initial model integrates EfficientNetV2-L with Gated Recurrent Unit (GRU) and attains an accuracy of 61.45%. Subsequently, the second model combines EfficientNetV2-L with bidirectional GRU (Bi-GRU), yielding an accuracy of 61.56%. The third and fourth models leverage a fusion of EfficientNetV2-L with Long Short-Term Memory (LSTM) and bidirectional LSTM (Bi-LSTM), achieving accuracies of 62.11% and 61.67%, respectively. Our findings demonstrate the viability of these models in effectively discerning student engagement levels, with the EfficientNetV2-L+LSTM model emerging as the most proficient, reaching an accuracy of 62.11%. This study underscores the potential of hybrid spatio-temporal networks in automating the detection of student engagement, thereby contributing to advancements in online education quality.

KEYWORDS

Student engagement detection; hybrid deep learning models; computer vision; EfficientNetV2-L; online learning environments; spatio-temporal classification

1 Introduction

The rise of online education, propelled by advancements in Internet technology has garnered widespread popularity among students [1]. In contrast to traditional teaching methods, online learning streamlines and enhances the accessibility of educational resources [2]. The global accessibility and affordability of education owe much to the transformative impact of online learning. However, amidst the benefits and growing interest in distance education, a pressing concern revolves around students' performance and active engagement in online learning environments [3].

Central to effective learning is the concept of student engagement [4], denoting active involvement in situations conducive to high-quality learning outcomes [5]. Actively engaged students generally exhibit better conceptual understanding and learning outcomes [6]. Student engagement encompasses

behavioral, cognitive, and emotional states within the learning environment [7]. Behavioral engagement requires active participation in class activities, emphasizing effort and perseverance [8], while cognition involves learning skills such as perception, storage, processing, and retrieval [9]. Emotional engagement reflects a student's active participation influenced by affective states [10], where positive emotions like happiness and interest enhance focus and engagement, while negative emotions such as boredom and frustration lead to disengagement [11].

In the realm of online learning, challenges such as lack of motivation and focus often arise, directly impacting engagement [12]. Unlike physical classrooms where teachers gauge engagement through facial expressions and social cues such as yawning, body posture, and glued eyes, assessing engagement in online environments proves significantly more intricate. Diverse electronic devices and varied backgrounds further complicate tracking students' engagement [13]. A pivotal aspect of enhancing the quality of online learning is the automated prediction of students' engagement levels [14]. This holds across various learning environments, encompassing traditional classrooms, massive open online courses (MOOCs), intelligent tutoring systems (ITS), and educational games.

Several methods exist for automating the determination of students' engagement in online education, broadly categorized into sensor-based and computer-vision-based approaches. Notably, computer-vision-based approaches, further divided into image-based and video-based methods, have garnered substantial interest. The image-based approaches rely solely on spatial information from a single image or frame which is a significant limitation. Since engagement detection is a spatio-temporal effective behavior because it is not stable over time, therefore, video-based methods emerge as more efficient and popular for detecting students' engagement [15].

Video-based methods predominantly fall into two categories: Machine learning-based and deep learning-based approaches. Machine learning-based methods extract features and employ handcrafted patterns for engagement estimation [16], while deep learning techniques dynamically learn features from training data, enabling the algorithm to discern subtle variations [17]. Deep learning methods surpass traditional machine learning in tasks requiring affective state prediction. Moreover, deep learning-based facial expression analysis in video data is non-intrusive, automated, and easily implementable [18].

This study aims to propose a new spatio-temporal hybrid deep learning model for detecting and classifying students' engagement from video data by combining the advantages of EfficientNetV2-L with four different RNN-based Models in online learning environments. The rest of the paper is organized as follows. Reviewing recent studies in student engagement detection in Section 2. Section 3 delves into the proposed deep learning approach, followed by experimental findings in Section 4 and concluding remarks in Section 5.

2 Related Works

In the realm of automatic student engagement detection, two primary methods have emerged: Sensor-based approaches and video-based methods. Sensor-based methods rely on physiological signals, encompassing heart rate variability, skin temperature, blood volume pulse, electrodermal activity (EDA), electrocardiogram (ECG), electromyogram (EMG), and electroencephalogram (EEG) [19,20].

Authors in [21] demonstrated the feasibility of distinguishing engaged and non-engaged students during lectures using wearable electrodermal activity sensors. Employing the Empatica E4 wristband

[22], which integrates blood volume pulse, acceleration, peripheral skin temperature, and electrodermal activity sensors, they recorded physiological data to achieve this distinction.

Kerdawy et al. [9] proposed a method for predicting students' cognitive states, engagement, and spontaneous attention by combining facial expression modalities and electroencephalography (EEG). They observed strong agreement between EEG and face-based models in engaged classes, with less agreement in non-engaged scenarios.

While some works explore sensor-based methods for detecting student engagement [23–25], challenges such as cost, wearability, portability, and mental privacy constraints hinder the implementation of brain-computer interface (BCI) modules in physical or online classrooms [26]. In contrast, video-based methods have gained prominence for their ease of data collection and unobtrusive evaluation processes [27].

Pise et al. [28] suggested a model that combined SqueezeNet [29] for feature extraction and temporal relational network (TRN) for connecting significant transformations between extracted spatio-temporal frames. This model achieved an accuracy of 91.30% on the DISFA+ dataset [30].

Gupta et al. [31] introduced the DAiSEE dataset, including affective states and engagement levels. They provided baseline results for four-class classification using CNN-based video classification techniques, such as InceptionNet frame level, InceptionNet video level [32], C3D training, C3D fine-tuning [33], and long-term recurrent convolutional networks (LRCN) [34], achieving accuracies of 47.1%, 46.4%, 48.6%, 56.1%, and 57.9%, respectively.

In [35], an inflated 3D convolutional network (I3D) was proposed for predicting students' engagement levels, utilizing OpenFace and AlphaPose for feature extraction, with an accuracy of 52.35% on the DAiSEE dataset.

Liao et al. [27] introduced the DFSTN model, combining long short-term memory (LSTM) with global attention (GALN) and pretrained SE-ResNet-50 (SENet) [36] for student engagement prediction. They tested the proposed method on the DAiSEE dataset and achieved an accuracy of 58.84%.

Abedi et al. [37] proposed a new end-to-end spatio-temporal hybrid method based on residual network (ResNet) [38], and temporal convolutional network (TCN) [39] for assessing student engagement in an online learning environment. While the ResNet extracts spatial features from subsequent video frames, TCN analyses the temporal changes in video frames to determine the degree of engagement. They achieved a performance increase of 63.9% on the DAiSEE dataset.

Bajaj et al. [40] utilized a hybrid neural network architecture based on ResNet and temporal convolutional network (TCN) for classifying student engagement, achieving a recognition accuracy of 53.6% on the DAiSEE dataset.

Mehta et al. [41] introduced a three-dimensional DenseNet Self-Attention neural network (3D DenseAttNet) for automatically detecting students' engagement in online learning environments. This model is designed to selectively extract relevant high-level intra-frame and inter-frame features from video data using the 3D DenseNet block. The proposed model surpassed the previous state-of-the-art, achieving a recognition accuracy of 63.59% on the DAiSEE dataset.

Gupta et al. [11] presented a deep learning approach centered on analyzing facial emotions to assess the engagement levels of students in real time during online learning. This system employs the faster region-based convolutional neural network (R-CNN) [42] for identifying faces and a modified face-points extractor (MFACXTOR) for pinpointing key facial features. The system was tested using

various deep learning architectures including Inception-V3 [32], VGG19 [43], and ResNet-50 [38] to determine the most effective model for accurately classifying real-time student engagement. The results from their experiments indicate that the system attained accuracies of 89.11% with Inception-V3, 90.14% with VGG19, and 92.32% with ResNet-50 on the dataset they developed.

Chen et al. [44] integrated gaze directions and facial expressions as separate elements in a multimodal deep neural network (MDNN) for predicting student engagement in collaborative learning settings. This multi-faceted approach was tested in an actual collaborative learning context. The findings demonstrate that the model is effective in precisely forecasting student performance within these environments.

Ahmad et al. [45] employed the lightweight MobileNetv2 model for automatic assessment of student engagement. The MobileNetv2 architecture's layers have all been fine-tuned to enhance learning efficiency and adaptability. The model's final layer was modified to classify three distinct output classes, instead of the original 1000 classes used in ImageNet. Their experimental analysis utilized an open-source dataset comprising individuals watching videos in online courses. The performance of lightweight MobileNetv2 was benchmarked against two other established pre-trained networks, ResNet-50 and Inception-V4, with MobileNetv2 achieving a superior average accuracy of 74.55%.

The authors in [46] developed a real-time system to monitor the engagement of student groups by analyzing their facial expressions and identifying affective states such as 'boredom,' 'confusion,' 'focus,' 'frustration,' 'yawning,' and 'sleepiness,' which are crucial in educational settings. This approach involves pre-processing steps like face detection, utilizing a convolutional neural network (CNN) for facial expression recognition, and post-processing for estimating group engagement frame by frame. To train the model, a dataset was compiled featuring the mentioned facial expressions from classroom lectures. This dataset was augmented with samples from three other datasets: BAUM-1 [47], DAiSEE [31], and YawDD [48], to enhance the model's predictive accuracy across various scenarios.

Sharma et al. [49] devised a method that amalgamates data on eye and head movements with facial emotional cues to create an engagement index categorized into three levels: "highly engaged," "moderately engaged," and "completely disengaged." They employed convolutional neural network (CNN) models for classification purposes and used them in the training process. Implemented in a standard e-learning context, the system demonstrated its efficacy by accurately determining the engagement level of students, and classifying them into one of the three aforementioned categories for each analyzed time segment.

Ikram et al. [50] developed a refined transfer learning approach using a modified VGG16 model, enhanced with an additional layer and meticulously calibrated hyperparameters. This model was designed to assess student engagement in a minimally controlled, real-world classroom setting with 45 students. In evaluating the level of student engagement, the model demonstrated impressive results, achieving 90% accuracy and a computation time of only 0.5 N seconds for distinguishing between engaged and non-engaged students.

3 Methodology and Proposed Model

The majority of available datasets for detecting student engagement are either privately held or small in scale, making it challenging to benchmark our research. Consequently, we opted to use the public DAiSEE dataset [31] for our evaluation and comparisons. One key limitation of current models for four-level classification on the DAiSEE dataset is their subpar accuracy. To address this issue, we leveraged EfficientNetV2-L [51] for extracting spatial features from video frames and

employed four distinct RNN-based models to capture temporal information, thereby enhancing accuracy. Notably, among the various model families, EfficientNetV2 stands out as the top performer, surpassing EfficientNet [52], ResNet [38], DenseNet [53], and Inception [32] models, which contributes to the overall improvement in accuracy. Additionally, the adoption of EfficientNetV2 substantially accelerates the training process.

Fig. 1 illustrates the block diagram of our methodology designed to predict automated student engagement in an online learning environment using video data. The proposed pipeline comprises several essential stages, including Dataset Selection: Involving the careful selection of an appropriate dataset for analysis. Pre-Processing Stage: Encompasses critical data preparation steps such as data reduction and data normalization.

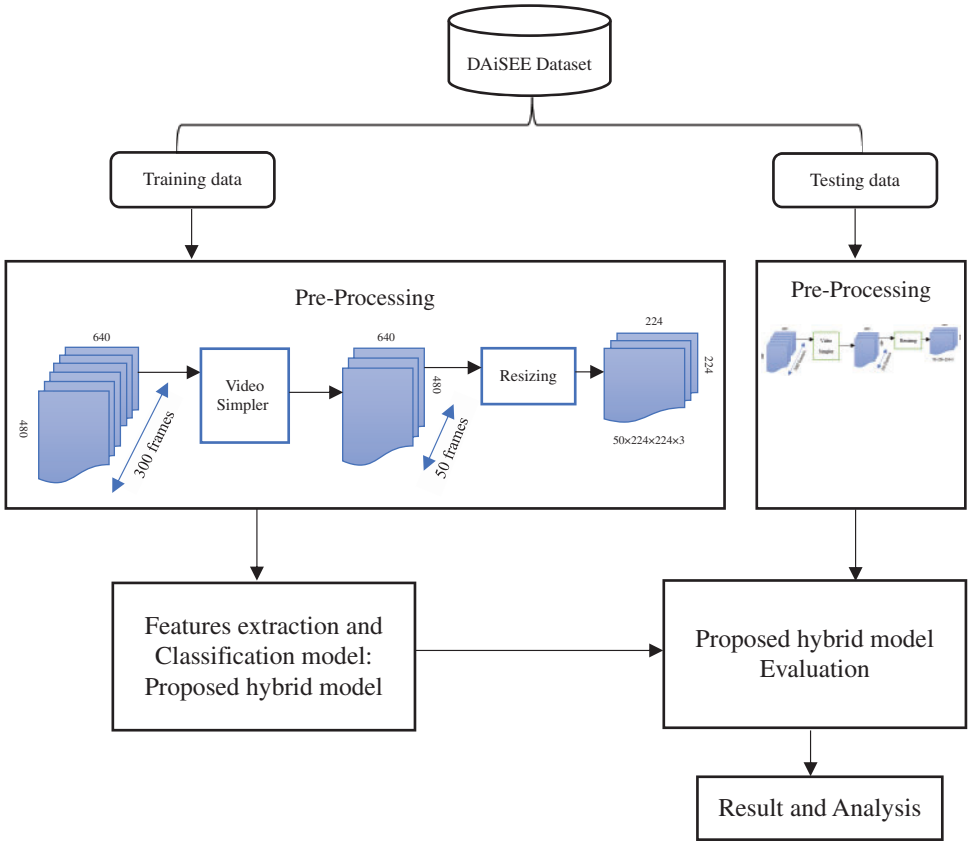


Figure 1: Block diagram of the proposed methodology for student engagement detection

Feature Extraction and Classification: Utilizes our proposed hybrid deep learning model to perform feature extraction and classification of relevant engagement levels. Model Evaluation: Includes the assessment and validation of our proposed model's performance. Experimental Result Analysis: Analyzing the outcomes of our experiments to gain insights into student engagement patterns and behavior. This comprehensive methodology is designed to enhance our understanding of student engagement in online learning by leveraging advanced deep learning techniques and rigorous data analysis procedures.

The architecture of the proposed hybrid deep learning model for detecting student engagement levels is depicted in Fig. 2. Raw video frames serve as input data to the model, generating output

across four distinct classes reflecting students' levels of engagement. Given the spatio-temporal nature of student engagement detection manifested in a sequence of video frames over time, a comprehensive analysis demands both spatial and temporal considerations. This analysis typically entails the monitoring and evaluation of students' conduct within an online learning environment by examining video footage. Regarding the spatial dimension, it involves monitoring the positions of students within the virtual classroom or e-learning platform. This encompasses identifying their screen location and observing visual cues linked to their engagement, such as eye movement and facial expressions. On the other hand, the temporal dimension concentrates on how student engagement evolves throughout an e-learning session over time. This involves tracing fluctuations in engagement levels during lectures, interactive activities, or discussions. Various features are derived from the video data to define students' behavior and involvement, encompassing aspects like facial expressions, body language, and interactions with e-learning materials. The extraction and classification of these features employ machine learning and computer vision techniques. This study employs EfficientNetV2-L to extract spatial features from video frames, while four distinct RNN-based models capture temporal information and model the sequence of frames.

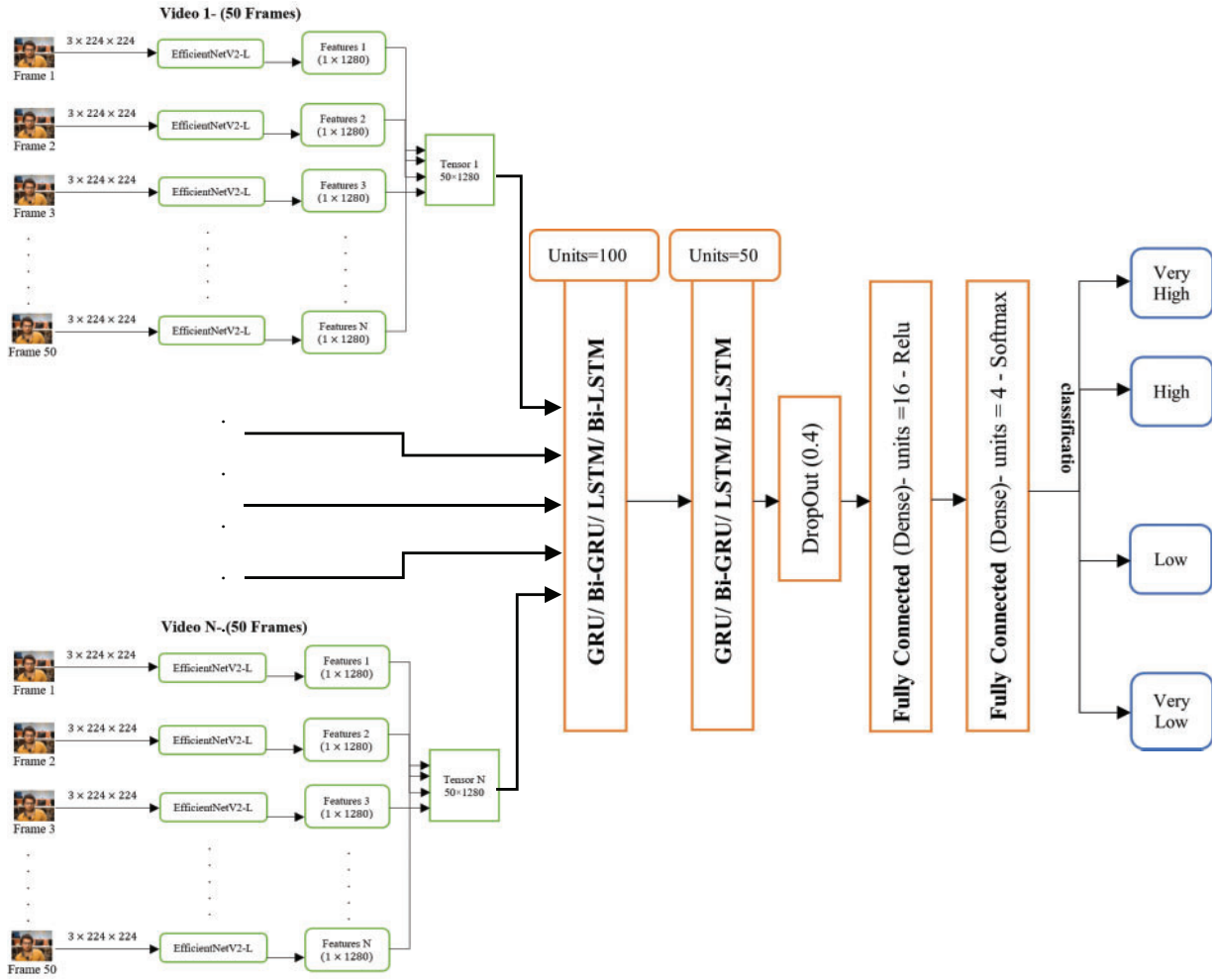


Figure 2: The structure of the proposed hybrid model for determining student engagement

The proposed hybrid models include (1) EfficientNetV2-L with gated recurrent unit (GRU), (2) EfficientNetV2-L with bidirectional GRU (Bi-GRU), (3) EfficientNetV2-L with long short-term memory (LSTM), and (4) EfficientNetV2-L with bidirectional LSTM (Bi-LSTM).

3.1 EfficientNetV2

EfficientNetV2 represents an advancement over previous models like DenseNet [53] and EfficientNet [52], demonstrating superior training speed and parameter efficiency. The architecture incorporates mobile inverted bottleneck (MBConv) [54] and fused-MBConv [55] as fundamental building blocks. Pre-training is performed on the ImageNet dataset [56]. The architecture of EfficientNetV2, illustrated in Fig. 3, distinguishes itself from the EfficientNet backbone in key aspects: 1-Increased use of both MBConv and fused-MBConv in the initial layers. 2-Preference for smaller expansion coefficients for MBConv. 3-Preference for smaller kernel size (3×3) compensated by an increased number of layers. 4-Elimination of the final stride-1 step present in the original EfficientNet, likely to address memory access overhead and large parameter size.

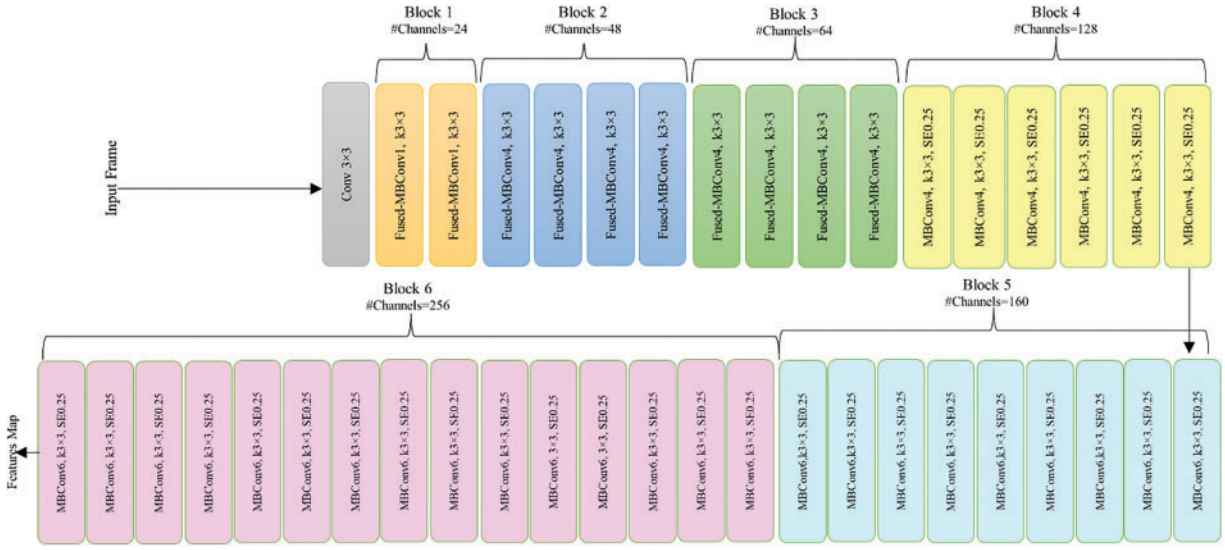


Figure 3: Architecture of EfficientNetV2

3.2 Long Short-Term Memory (LSTM)

The long short-term memory (LSTM), introduced as a seminal work in [57], epitomizes a sophisticated iteration of recurrent neural network (RNN), meticulously crafted to tackle the pervasive issue of long-term dependency [58]. Proven to excel in retaining information over extended sequences, LSTM tackles the vanishing gradient problem effectively [59]. The LSTM network processes the output from the previous time step and the current input at a given time step, producing an output sent to the subsequent time step. The last time step's final hidden layer is commonly utilized for classification [60].

The LSTM architecture includes a memory unit denoted as c, a hidden state represented by h, and three distinct gates: The input gate (i), the forget gate (f), and the output gate (o). These gates play a crucial role in controlling the flow of information in and out of the memory unit, effectively managing reading and writing operations within the LSTM framework. Specifically, the input gate determines the manner in which the internal state is updated based on the current input and the preceding internal

state. Conversely, the forget gate governs the degree to which the previous internal state is retained. Lastly, the output gate modulates the impact of the internal state on the overall system [61]. Fig. 4 demonstrates how the update process functions within the internal framework of an LSTM. More concretely, at each time step t, the LSTM initially receives an input x_t along with the previous hidden state h_{t-1}. Subsequently, it calculates activations for the gates and proceeds to update both the memory unit to c_t and the hidden state to h_t. This computational process can be outlined as follows [62]:

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right) \tag{1}$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \tag{2}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{3}$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right) \tag{4}$$

$$h_t = o_t \odot tanh\left(c_t\right) \tag{5}$$

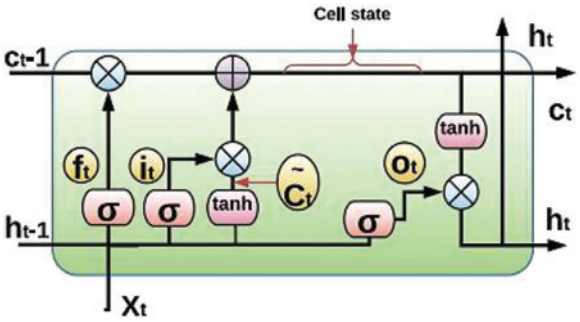


Figure 4: The inner structure of a LSTM unit

Here, the symbol $\sigma(x)$ represents the logistic sigmoid function defined as $\sigma(x) = 1/(1 + exp(-x))$. The symbol $\odot$ denotes the point-wise product operation. The parameters W and b correspond to the weights and biases associated with the three gates and the memory unit.

A version of LSTM known as bidirectional long short-term memory (Bi-LSTM) [63] addresses the drawbacks of traditional LSTM architectures by incorporating both preceding and succeeding contexts in tasks involving sequence modeling. Unlike LSTM models, which solely handle input data in a forward direction, Bi-LSTM operates in both forward and backward directions [64].

3.3 Gated Recurrent Unit (GRU)

The gated recurrent unit (GRU) serves as an alternative variant to the traditional recurrent neural network (RNN), aimed at resolving issues related to short-term memory through a design that is less complex than the long short-term memory (LSTM) [65]. By consolidating the input and forget gates found in LSTM into a singular update gate, GRU achieves an improvement in overall efficiency. Comprising update gate, reset gate, and current memory content, GRU identifies long-term dependencies in sequences. The gates allow for selective modification and utilization of data from previous time steps, aiding in the identification of long-term dependencies [66]. Fig. 5 provides a visual representation of the GRU unit's architecture.

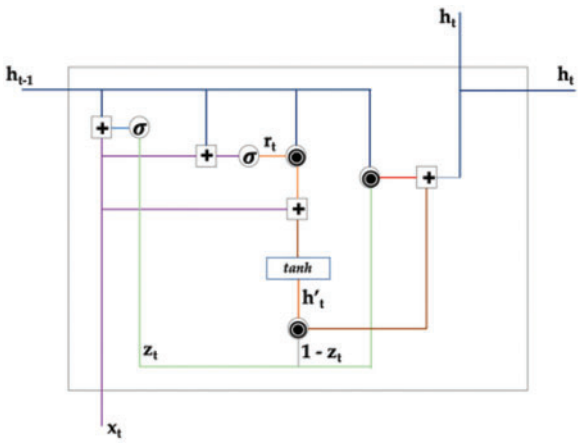


Figure 5: The inner structure of a GRU unit [66]

At time t, the GRU cell's activation, denoted as h_t^j, is determined through a weighted mix of its previous activation (h_{t-1}^j), and a candidate activation ($\tilde{h}_t^j$), as follows [67]:

$$h_t^j = \left(1 - h_{z_t}^j\right) h_{t-1} + z_t^j \tilde{h}_t^j \tag{6}$$

Here, an update gate, denoted as z_t^j, determines the extent to which the unit updates its activation or content. The formulation for this gate is given by:

$$z_t^j = \sigma \left(W_z x_t + U_z h_{t-1}\right)^j \tag{7}$$

This process involves calculating a linear combination of the current state and a newly generated state, a technique reminiscent of what is seen in LSTM units. However, unlike LSTM, GRU lacks a mechanism for regulating how much of their state is revealed, instead opting to fully disclose their entire state at each update.

The candidate activation, denoted as $\tilde{h}_t^j$, is calculated in a manner akin to the conventional recurrent unit.

$$\tilde{h}_t^j = \tanh\left(\mathrm{W} x_t + U\left(r_t \odot h_{t-1}\right)\right)^j \tag{8}$$

where r_t represents a collection of reset gates and $\odot$ indicating element-wise multiplication. When r_t approaches 0, indicating "off," the reset gate essentially causes the unit to behave as if it is processing the initial symbol of an input sequence, allowing it to forget the previously computed state. The calculation of the reset gate, denoted as r_t^j, follows a process similar to that of the update gate.

$$r_t^j = \sigma \left(W_r x_t + U_r h_{t-1}\right)^j \tag{9}$$

GRU models, which require fewer tensor operations, provide a simpler option compared to LSTM, leading to quicker training times. Nonetheless, whether to use GRU or LSTM is contingent on the particular use case and the nature of the problem being addressed [58].

A notable improvement to the GRU architecture is the Bi-GRU [68], which successfully addresses specific limitations of the standard GRU by integrating information from both past and future contexts in sequential modeling tasks. In contrast to the GRU, which handles input sequences exclusively in a forward direction, the Bi-GRU operates in both forward and backward directions.

In a Bi-GRU model, two parallel GRU layers are employed, with one processing the input data in the forward direction and the other handling it in reverse [69].

4 Experimental Results

4.1 Dataset

The prevalent datasets for student engagement detection are largely private or limited in size, posing challenges in benchmarking our research against existing work. Therefore, in this study, we conducted experiments and evaluated the proposed models using the DAiSEE dataset (Dataset for Affective States in E-Environments) [31]. The dataset comprises 112 students currently enrolled in school, aged between 18 to 30, with a predominantly Asian demographic, comprising 32 females and 80 males. A total of 9068 video clips, each lasting 10 seconds, were captured in six distinct locations such as dorm rooms, labs, and libraries, under three lighting conditions: bright, dark, and mild. Under different lighting conditions, using indoor or outdoor light sources, images or videos absorb light properties that are inextricably linked to the original image [70]. The DAiSEE dataset encompasses four affective states including confusion, boredom, engagement, and frustration, each with four levels: "very low," "low," "high," and "very high." This paper focuses predominantly on assessing student engagement levels during online learning. Table 1 presents the detailed distribution of engagement levels.

Table 1: Data distribution on the DAiSEE dataset

Level	Train	Validation	Test	Total
Very high	2494	450	814	3758
High	2617	813	882	4312
Low	213	143	84	440
Very low	34	23	4	61
Total	5358	1429	1784	8571

4.2 Result

We evaluated the four proposed deep learning models, namely EfficientNetV2-L+GRU, EfficientNetV2-L+Bi-GRU, EfficientNetV2-L+LSTM, and EfficientNetV2-L+Bi-LSTM, utilizing the DAiSEE dataset to investigate the effectiveness of each model. Before experimentation, the decision was made regarding the number of frames from each video to be fed into the model. Utilizing a vector with k-frames to represent the spatial features of a video, we aimed to balance temporal information and training time. In this study, we opted for 50 frames per video, resizing them to 224×224 to generate $50 \times 3 \times 224 \times 224$ ($L \times C \times H \times W$) tensors as inputs to the model. The EfficientNetV2-L model extracts feature vectors of dimension 1280 from successive frames, subsequently feeding them to the RNN-based module. The parameter values used are provided in Table 2.

The performance of the proposed models is summarized in Table 3. The results highlight the EfficientNetV2-L+LSTM model as the top performer among the proposed models, achieving an accuracy of 62.11%. Accuracy is measured as the ratio of correct to incorrect prediction [71].

Table 2: The parameter values used for experiments

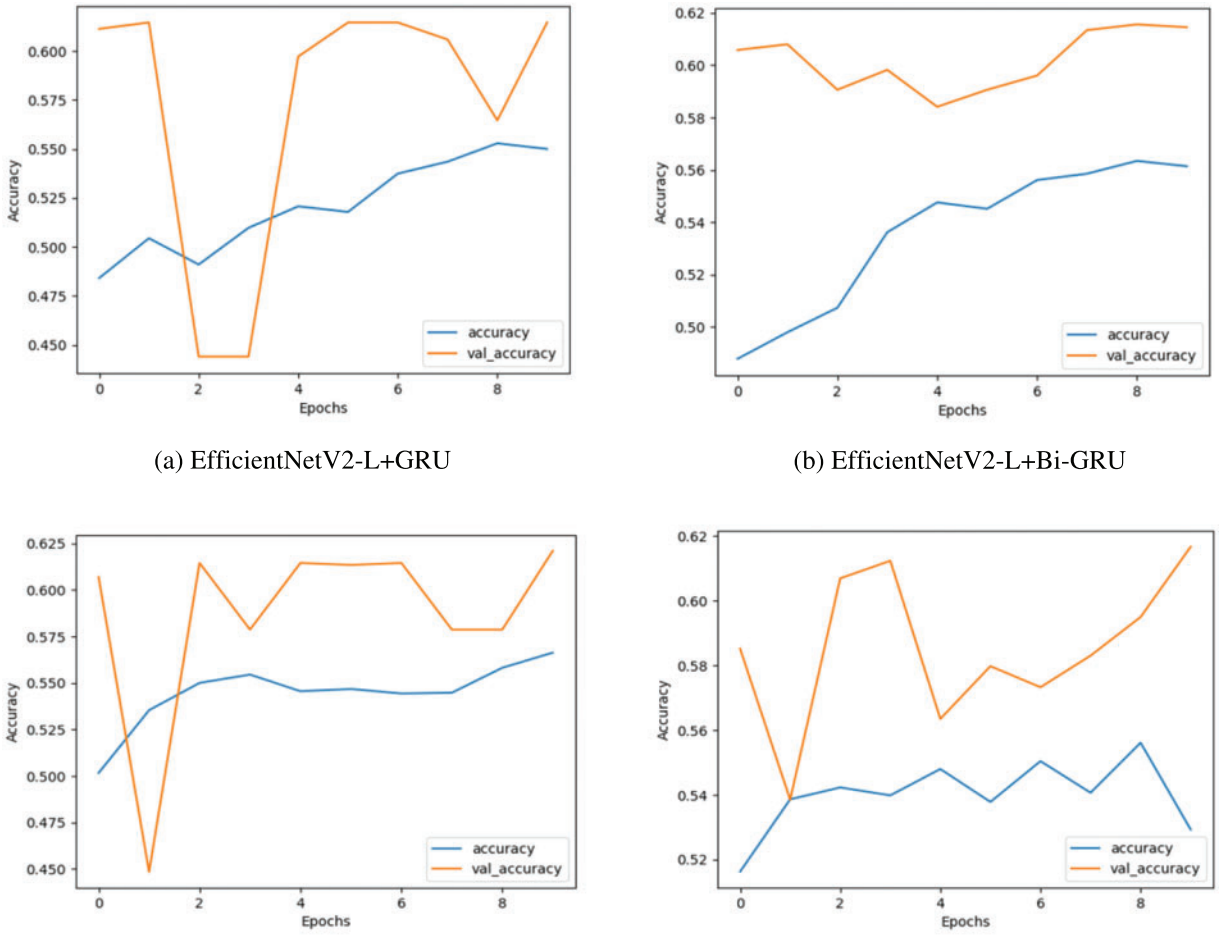
Model	EfficientNetV2L	GRU	Bi-GRU	LSTM	Bi-LSTM
Input	50 × 224 × 224 × 3	50 × 1280	50 × 1280	50 × 1280	50 × 1280
Layer 1		Unit = 100, return sequences = True	Unit = 100, return_ sequences = True	Unit = 100, return_ sequences = True	Unit = 100, return_ sequences = True
Layer 2		Dropout: 0.4	Dropout: 0.4	Dropout: 0.4	Dropout: 0.4
Layer 3		Unit = 50	Unit = 50	Unit = 50	Unit = 50
Layer 4		Dropout: 0.4	Dropout: 0.4	Dropout: 0.4	Dropout: 0.4
Layer 5		Dense: 16 units, Activation: Relu	Dense: 16 units, Activation: Relu	Dense: 16 units, Activation: Relu	Dense: 16 units, Activation: Relu
Output	50 × 1280	Dense: 4 units, Activation: softmax	Dense: 4 units, Activation: softmax	Dense: 4 units, Activation: softmax	Dense: 4 units, Activation: softmax
	weights = imagenet	Loss function: "binary_crossentropy", Optimizer function: "Adam"			
	pooling = "avg"	Batch size = 32, Epochs = 10			

Table 3: Accuracies of the four proposed models

Model	Accuracy
EfficientNetV2-L+GRU	61.45%
EfficientNetV2-L+Bi-GRU	61.56%
EfficientNetV2-L+LSTM	62.11%
EfficientNetV2-L+Bi-LSTM	61.67%

Fig. 6 illustrates accuracy and validation-accuracy diagrams, offering a visual representation of how the proposed models perform during training and testing across multiple epochs. Upon closer examination of the graphs, a noticeable trend emerges. In all the charts, the training accuracy initiates at approximately 50% in the first epoch and gradually rises to around 55% by the eighth epoch, after which it stabilizes. However, in graph (d), a decline in training accuracy is observed post the eighth epoch. As for the validation accuracy depicted in all the graphs, there are notable fluctuations. These fluctuations stem from the dataset's inherent imbalance in terms of engagement level distribution. Specifically, the number of samples with low engagement levels is considerably

lower than those with high engagement levels. In such a skewed distribution, it is plausible that a majority of the minority-level samples are misclassified as belonging to the majority engagement levels. Nonetheless, these fluctuations are less pronounced in graphs (b) and (d). Consequently, it can be inferred that bidirectional RNN models exhibit greater stability when dealing with imbalanced datasets in comparison to unidirectional RNN models.



(a) EfficientNetV2-L+GRU

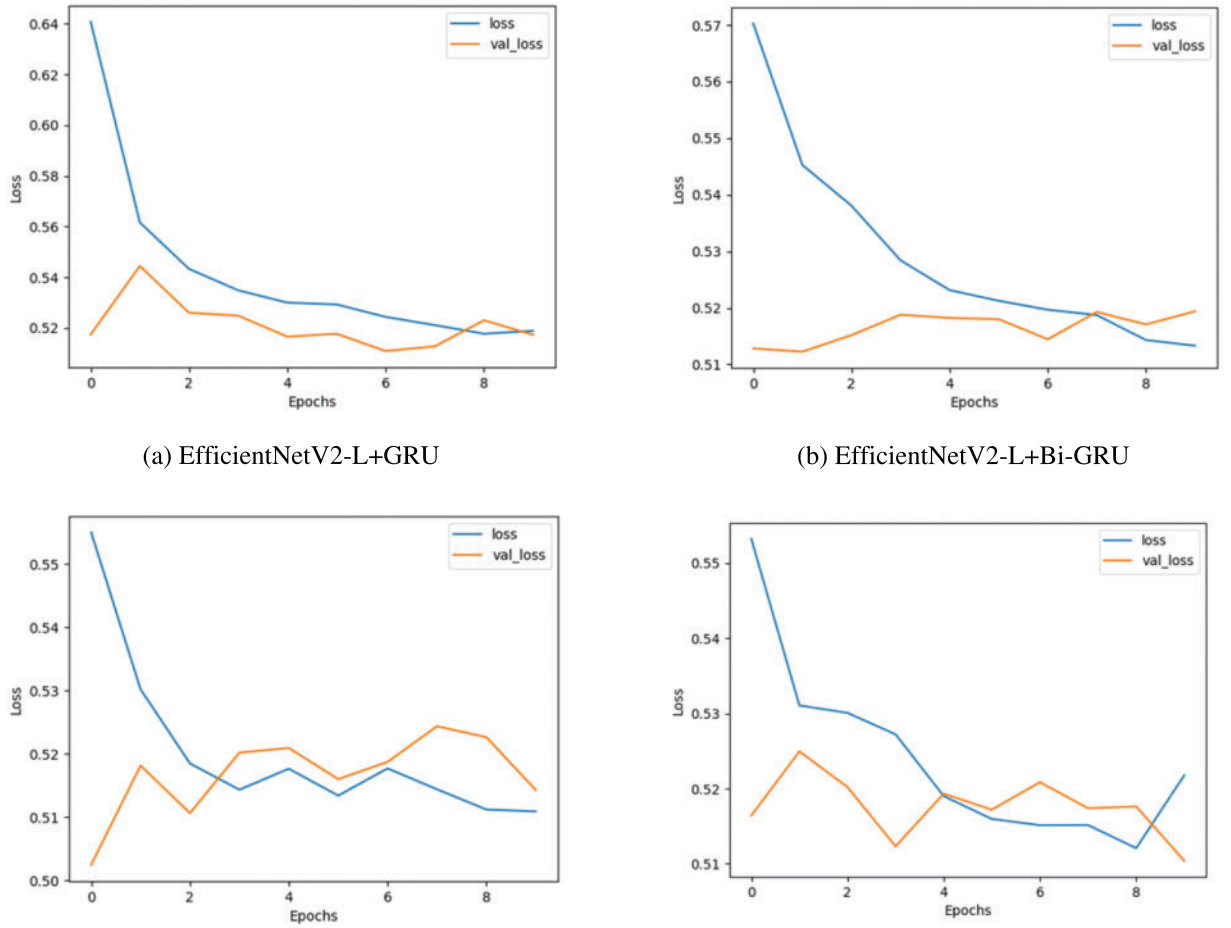
(b) EfficientNetV2-L+Bi-GRU

(c) EfficientNetV2-L+LSTM

(d) EfficientNetV2-L+ Bi-LSTM

Figure 6: The accuracy diagram of the four proposed models on training and testing

Additionally, Fig. 7 presents loss and validation-loss diagrams, visually representing the fluctuation in loss values during training and evaluation processes for the different models. The loss function quantifies the dissimilarity between predicted and actual labels. By reviewing the graphs, a clear trend emerges that the training losses have consistently diminished across all graphs. Notably, graph (c) exhibits the lowest training loss, hovering around 0.51. Moreover, the validation loss in graphs (a) and (b) demonstrates greater stability compared to graphs (c) and (d). However, it is worth noting that the final validation loss values in graphs (c) and (d), both approximately at 0.51, are lower than the values in the other two graphs, which are approximately 0.52. This observation indicates that in this specific context, the LSTM models outperform the GRU model.



(a) EfficientNetV2-L+GRU

(b) EfficientNetV2-L+Bi-GRU

(c) EfficientNetV2-L+LSTM

(d) EfficientNetV2-L+ Bi-LSTM

Figure 7: The Loss diagram of the four proposed models on training and testing

4.3 Comparison Performance

Table 4 intricately compares the outcomes of our four proposed hybrid models with previous studies utilizing the DAiSEE dataset. In a benchmark study [31], diverse deep learning models, including video-level InceptionNet, C3D fine-tuning, and long-term recurrent convolutional network (LRCN), were tested. LRCN emerged as the leader with an accuracy of 57.90%. Several other models, such as inflated 3D convolutional network (I3D) [35], convolutional 3D (C3D) neural networks with focal loss [72], ResNet+TCN with weighted loss [37], and ResNet+TCN [40], were introduced in subsequent works. Despite these efforts, the consistently superior performance of the LRCN model remained. Comparatively, DFSTN [27] surpassed LRCN with an accuracy of 58.84%, while the deep engagement recognition network (DERN) [73] which combines temporal convolution, bidirectional LSTM, and attention mechanism, achieved 60%, a 1.16% improvement over DFSTN. The Neural Turing Machine [74] exhibited an accuracy of 61.3% which is better than DERN. Notably, the proposed EfficientNetV2-L+LSTM model, achieving an accuracy of 62.11%, outperformed both LRCN and the majority of contemporary models. However, DenseAttNet [41] with 63.59%, and ResNet+TCN [37] with 63.9% outperformed previous works. This comparative analysis underscores

that our proposed models exhibit sufficient accuracy in detecting student engagement within the DAiSEE dataset compared to earlier models.

Table 4: Comparison of proposed models and previous works on DAiSEE

Model	Accuracy
InceptionNet [31]	46.40%
C3D fine-tuning [31]	56.10%
LRCN [31]	57.90%
I3D [35]	52.35%
C3D (FL) [72]	56.20%
ResNet+TCN [40]	53.60%
DFSTN [27]	58.84%
DERN [73]	60.00%
Neural turing machine [74]	61.30%
ResNet+TCN with weighted loss [37]	53.70%
C3D+TCN [37]	59.97%
ResNet+LSTM [37]	61.15%
ResNet+TCN [37]	63.90%
DenseAttNet [41]	63.59%
EfficientNetV2-L+GRU (proposed)	**61.45%**
EfficientNetV2-L+Bi-GRU (proposed)	**61.56%**
EfficientNetV2-L+LSTM (proposed)	**62.11%**
EfficientNetV2-L+Bi-LSTM (proposed)	**61.67%**

5 Conclusion

In this paper, our primary objective was to address the challenge faced by teachers in accurately and promptly detecting their students' engagement in online learning. To achieve this, we introduced four hybrid spatio-temporal models designed for detecting student engagement from video in online learning environments. These models encompassed a hybrid EfficientNetV2-L in conjunction with gated recurrent unit (GRU), a hybrid EfficientNetV2-L paired with Bidirectional GRU, a hybrid EfficientNetV2-L combined with long short-term memory (LSTM), and a hybrid EfficientNetV2-L together with Bidirectional LSTM.

The EfficientNetV2-L played a pivotal role in spatial feature extraction, while GRU, Bidirectional GRU, LSTM, and Bidirectional LSTM were employed to capture temporal information from sequential data. Our experimentation, conducted on the DAiSEE dataset featuring four levels of student engagement, demonstrated that the proposed models exhibited superior accuracy compared to the majority of previous works utilizing the same dataset. Notably, the EfficientNetV2-L+LSTM model emerged as the top performer, achieving an accuracy of 62.11%.

Despite these promising results, certain limitations exist in the current study. To address these, future research will refine the automatic recognition of learning engagement by implementing a robust face detector to crop face regions from each frame during pre-processing. Additionally, the

incorporation of attention mechanisms in the proposed models will be explored to further enhance accuracy. Furthermore, our commitment to advancing research in this domain involves testing the suggested models on diverse datasets, ensuring broader applicability and generalizability.

In essence, this study contributes valuable insights into automating the detection of student engagement in online learning environments. The demonstrated effectiveness of our hybrid models highlights their potential to provide teachers with accurate assessments of student engagement, thus contributing to the ongoing efforts to enhance the quality of online education.

Acknowledgement: The authors would like to express sincere gratitude to all the individuals who have contributed to the completion of this research paper. Their unwavering support, valuable insights, and encouragement have been instrumental in making this endeavor a success.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: F. M. Shiri, E. Ahmadi, M. Rezaee; data collection: F. M. Shiri; analysis and interpretation of results: F. M. Shiri, E. Ahmadi; draft manuscript preparation: F. M. Shiri. E. Ahmadi, M. Rezaee, T. Perumal. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The code used and/or analyzed during this research are available from the corresponding author upon reasonable request. Data used in this study can be accessed via the following link: https://people.iith.ac.in/vineethnb/resources/daisee/index.html.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Haleem, M. Javaid, M. A. Qadri, and R. Suman, "Understanding the role of digital technologies in education: A review," *Sustain. Oper. Comput.*, vol. 3, no. 4, pp. 275–285, 2022. doi: 10.1016/j.susoc.2022.05.004.

[2] I. Blagoev, G. Vassileva, and V. Monov, "A model for e-learning based on the knowledge of learners," *Cybernet. Inf. Technol.*, vol. 21, no. 2, pp. 121–135, 2021. doi: 10.2478/cait-2021-0023.

[3] A. Alhothali, M. Albsisi, H. Assalahi, and T. Aldosemani, "Predicting student outcomes in online courses using machine learning techniques: A review," *Sustain.*, vol. 14, no. 10, pp. 6199, 2022. doi: 10.3390/su14106199.

[4] P. Redmond, L. Abawi, A. Brown, R. Henderson, and A. Heffernan, "An online engagement framework for higher education," *Online Learn. J.*, vol. 22, no. 1, pp. 183–204, 2018. doi: 10.24059/olj.v22i1.1175.

[5] N. A. Johar, S. N. Kew, Z. Tasir, and E. Koh, "Learning analytics on student engagement to enhance students' learning performance: A systematic review," *Sustain.*, vol. 15, no. 10, pp. 7849, 2023. doi: 10.3390/su15107849.

[6] S. N. Kew and Z. Tasir, "Analysing students' cognitive engagement in e-learning discussion forums through content analysis," *Knowl. Manage. E-Learn.*, vol. 13, no. 1, pp. 39–57, 2021.

[7] M. Pilotti, S. Anderson, P. Hardy, P. Murphy, and P. Vincent, "Factors related to cognitive, emotional, and behavioral engagement in the online asynchronous classroom," *Int. J. Teach. Learn. High. Edu.*, vol. 29, no. 1, pp. 145–153, 2017.

[8] E. Okur, N. Alyuz, S. Aslan, U. Genc, C. Tanriover and A. A. Esme, "Behavioral engagement detection of students in the wild," in *Artif. Intell. Edu.: 18th Int. Conf.*, Wuhan, China, Jun. 28–Jul. 01, 2017, pp. 250–261.

[9] M. El Kerdawy *et al.*, "The automatic detection of cognition using eeg and facial expressions," *Sensors*, vol. 20, no. 12, pp. 3516, 2020. doi: 10.3390/s20123516.

[10] M. Mukhopadhyay, S. Pal, A. Nayyar, P. K. D. Pramanik, N. Dasgupta and P. Choudhury, "Facial emotion detection to assess learner's state of mind in an online learning system," in *Proc. 2020 5th Int. Conf. Intell. Inf. Technol.*, 2020, pp. 107–115.

[11] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimed. Tools Appl.*, vol. 82, no. 8, pp. 11365–11394, 2023. doi: 10.1007/s11042-022-13558-9.

[12] M. A. Al Mamun and G. Lawrie, "Factors affecting student behavioural engagement in an inquiry-based online learning environment," 2021. doi: 10.21203/rs.3.rs-249144/v1

[13] B. Zhu, X. Lan, X. Guo, K. E. Barner, and C. Boncelet, "Multi-rate attention based GRU model for engagement prediction," in *Proc. 2020 Int. Conf. Multi. Interact.*, 2020, pp. 841–848.

[14] J. Whitehill, Z. Serpell, Y. C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, 2014. doi: 10.1109/TAFFC.2014.2316163.

[15] 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. doi: 10.1109/ACCESS.2022.3206779.

[16] N. Bosch *et al.*, "Automatic detection of learning-centered affective states in the wild," in *Proc. 20th Int. Conf. Intell. User Interfaces*, 2015, pp. 379–388.

[17] S. Zhang, X. Pan, Y. Cui, X. Zhao, and L. Liu, "Learning affective video features for facial expression recognition via hybrid deep learning," *IEEE Access*, vol. 7, pp. 32297–32304, 2019. doi: 10.1109/ACCESS.2019.2901521.

[18] M. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: A review," *Smart Learn. Environ.*, vol. 6, no. 1, pp. 1–20, 2019. doi: 10.1186/s40561-018-0080-z.

[19] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 15–28, 2016. doi: 10.1109/TAFFC.2016.2515084.

[20] M. Bustos-López, N. Cruz-Ramírez, A. Guerra-Hernández, L. N. Sánchez-Morales, N. A. Cruz-Ramos and G. Alor-Hernández, "Wearables for engagement detection in learning environments: A review," *Biosens.*, vol. 12, no. 7, pp. 509, 2022. doi: 10.3390/bios12070509.

[21] E. Di Lascio, S. Gashi, and S. Santini, "Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors," in *Proc. ACM Interact., Mobile, Wear. Ubiquit. Technol.*, vol. 2, no. 3, pp. 1–21, 2018. doi: 10.1145/3264913.

[22] M. Garbarino, M. Lai, D. Bender, R. W. Picard, and S. Tognetti, "Empatica E3—A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition," in *4th Int. Conf. Wirel. Mobile Commun. Healthcare-Transf. Healthcare Innov. Mobile Wirel. Technol. (MOBIHEALTH),* IEEE, 2014, pp. 39–42.

[23] A. Al-Alwani, "A combined approach to improve supervised e-learning using multi-sensor student engagement analysis," *American J. Appl. Sci.*, vol. 13, no. 12, pp. 1377–1384, 2016. doi: 10.3844/ajassp.2016.1377.1384.

[24] A. Apicella, P. Arpaia, M. Frosolone, G. Improta, N. Moccaldi and A. Pollastro, "EEG-based measurement system for monitoring student engagement in learning 4.0," *Sci. Rep.*, vol. 12, no. 1, pp. 5857, 2022. doi: 10.1038/s41598-022-09578-y.

[25] D. K. Darnell and P. A. Krieg, "Student engagement, assessed using heart rate, shows no reset following active learning sessions in lectures," *PLoS One*, vol. 14, no. 12, pp. e0225709, 2019. doi: 10.1371/journal.pone.0225709.

[26] X. Hu, J. Chen, F. Wang, and D. Zhang, "Ten challenges for EEG-based affective computing," *Brain Sci. Adv.*, vol. 5, no. 1, pp. 1–20, 2019. doi: 10.1177/2096595819896200.

[27] J. Liao, Y. Liang, and J. Pan, "Deep facial spatiotemporal network for engagement prediction in online learning," *Appl. Intell.*, vol. 51, no. 10, pp. 6609–6621, 2021. doi: 10.1007/s10489-020-02139-8.
[28] A. Pise, H. Vadapalli, and I. Sanders, "Facial emotion recognition using temporal relational network: An application to e-learning," *Multimed. Tools Appl.*, vol. 81, no. 19, pp. 26633–26653, 2022.
[29] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," arXiv preprint arXiv:1602.07360, 2016.
[30] M. Mavadati, P. Sanger, and M. H. Mahoor, "Extended disfa dataset: Investigating posed and spontaneous facial expressions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. Workshops*, 2016, pp. 1–8.
[31] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," arXiv preprint arXiv:1609.01885, 2016.
[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 2818–2826.
[33] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.
[34] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 2625–2634.
[35] H. Zhang, X. Xiao, T. Huang, S. Liu, Y. Xia and J. Li, "An novel end-to-end network for automatic student engagement recognition," in *IEEE 9th Int. Conf. Electron. Inf. Emerg. Commun. (ICEIEC)*, IEEE, 2019, pp. 342–345.
[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2018, pp. 7132–7141.
[37] A. Abedi and S. S. Khan, "Improving state-of-the-art in detecting student engagement with ResNet and TCN hybrid network," in *2021 18th Conf. Robots Vis. (CRV)*, IEEE, 2021, pp. 151–157.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 770–778.
[39] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint arXiv:1803.01271, 2018.
[40] K. K. Bajaj, I. Ghergulescu, and A. N. Moldovan, "Classification of student affective states in online learning using neural networks," in *2022 17th Int. Workshop Semantic Soc. Media Adapt. Personal. (SMAP)*, IEEE, 2022, pp. 1–6.
[41] N. K. Mehta, S. S. Prasad, S. Saurav, R. Saini, and S. Singh, "Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement," *Appl. Intell.*, vol. 52, no. 12, pp. 13803–13823, 2022. doi: 10.1007/s10489-022-03200-4.
[42] L. Hai and H. Guo, "Face detection with improved face R-CNN training method," in *Proc. 3rd Int. Conf. Control Comput. Vis.*, 2020, pp. 22–25.
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[44] Y. Chen, J. Zhou, Q. Gao, J. Gao, and W. Zhang, "MDNN: Predicting student engagement via gaze direction and facial expression in collaborative learning," *Comput. Model. Eng. Sci.*, vol. 136, no. 1, pp. 381–401, 2023. doi: 10.32604/cmes.2023.023234.
[45] N. Ahmad, Z. Khan, and D. Singh, "Student engagement prediction in MOOCs using deep learning," in *2023 Int. Conf. Emerg. Smart Comput. Inf. (ESCI)*, IEEE, 2023, pp. 1–6.
[46] C. Pabba and P. Kumar, "An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition," *Expert. Syst.*, vol. 39, no. 1, pp. e12839, 2022. doi: 10.1111/exsy.12839.
[47] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 300–313, 2016. doi: 10.1109/TAFFC.2016.2553038.

[48] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, "YawDD: A yawning detection dataset," in *Proc. 5th ACM Multimed. Syst. Conf.*, 2014, pp. 24–28.
[49] P. Sharma *et al.*, "Student engagement detection using emotion analysis, eye tracking and head movement with machine learning," in *Int. Conf. Technol. Innov. Learn., Teach. Edu.*, Springer, 2022, pp. 52–68.
[50] S. Ikram *et al.*, "Recognition of student engagement state in a classroom environment using deep and efficient transfer learning algorithm," *Appl. Sci.*, vol. 13, no. 15, pp. 8637, 2023. doi: 10.3390/app13158637.
[51] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 10096–10106.
[52] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Int. Conf. Mach. Learn.*, PMLR, 2019, pp. 6105–6114.
[53] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2017, pp. 4700–4708.
[54] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2018, pp. 4510–4520.
[55] S. Gupta and B. Akin, "Accelerator-aware neural network design using automl," arXiv preprint arXiv:2003.02838, 2020.
[56] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conf. Comput. Vis. Pattern Recogn.*, IEEE, 2009, pp. 248–255.
[57] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.
[58] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU," arXiv preprint arXiv:2305.17473, 2023.
[59] V. Barot and V. Kapadia, "Long short term memory neural network-based model construction and Fne-tuning for air quality parameters prediction," *Cybernet. Inf. Technol.*, vol. 22, no. 1, pp. 171–189, 2022. doi: 10.2478/cait-2022-0011.
[60] S. Minaee, E. Azimi, and A. Abdolrashidi, "Deep-sentiment: Sentiment analysis using ensemble of CNN and Bi-LSTM models," arXiv preprint arXiv:1904.04206, 2019.
[61] W. Fang, Y. Chen, and Q. Xue, "Survey on research of RNN-based spatio-temporal sequence prediction algorithms," *J. Big Data*, vol. 3, no. 3, pp. 97–110, 2021. doi: 10.32604/jbd.2021.016993.
[62] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, "Robust LSTM-autoencoders for face de-occlusion in the wild," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 778–790, 2017. doi: 10.1109/TIP.2017.2771408.
[63] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, 2008. doi: 10.1109/TPAMI.2008.137.
[64] T. H. Aldhyani and H. Alkahtani, "A bidirectional long short-term memory model algorithm for predicting COVID-19 in gulf countries," *Life*, vol. 11, no. 11, pp. 1118, 2021. doi: 10.3390/life11111118.
[65] F. M. Shiri, T. Perumal, N. Mustapha, R. Mohamed, M. A. B. Ahmadon and S. Yamaguchi, "A survey on multi-resident activity recognition in smart environments," arXiv preprint arXiv:2304.12304, 2023.
[66] A. Dutta, S. Kumar, and M. Basu, "A gated recurrent unit approach to bitcoin price prediction," *J. Risk Finan. Manag.*, vol. 13, no. 2, pp. 23, 2020. doi: 10.3390/jrfm13020023.
[67] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.
[68] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.
[69] Y. Zhang, J. Chen, D. Wang, M. Hu, and L. Chen, "The bidirectional gate recurrent unit based attention mechanism network for state of charge estimation," *Journal of the Electrochemical Society*, vol. 169, no. 11, p. 110503, 2022.
[70] M. Kumar and S. Srivastava, "Image authentication by assessing manipulations using illumination," *Multimed. Tools Appl.*, vol. 78, no. 9, pp. 12451–12463, 2019. doi: 10.1007/s11042-018-6775-x.

[71] G. Madhu, S. Kautish, Y. Gupta, G. Nagachandrika, S. M. Biju and M. Kumar, "XCovNet: An optimized xception convolutional neural network for classification of COVID-19 from point-of-care lung ultrasound images," *Multimed. Tools Appl.*, vol. 83, no. 11, pp. 1–22, 2023. doi: 10.1007/s11042-023-16944-z.

[72] L. Geng, M. Xu, Z. Wei, and X. Zhou, "Learning deep spatiotemporal feature for engagement recognition of online courses," in *2019 IEEE Symp. Series Comput. Intell. (SSCI)*, IEEE, 2019, pp. 442–447.

[73] T. Huang, Y. Mei, H. Zhang, S. Liu, and H. Yang, "Fine-grained engagement recognition in online learning environment," in *2019 IEEE 9th Int. Conf. Electr. Inf. Emerg. Commun. (ICEIEC)*, IEEE, 2019, pp. 338–341.

[74] X. Ma, M. Xu, Y. Dong, and Z. Sun, "Automatic student engagement in online learning environment based on neural turing machine," *Int. J. Inf. Edu. Technol.*, vol. 11, no. 3, pp. 107–111, 2021. doi: 10.18178/ijiet.2021.11.3.1497.