A lightweight and reliable framework toward real-time student engagement predictions in learning analytics

Long Hoang, School of Computer Science and Information Technology, University College Cork
George Shorten, Department of Anaesthesiology and Intensive Care Medicine, University College Cork
Barry O'Sullivan, School of Computer Science and Information Technology, University College Cork
Hoang D. Nguyen, School of Computer Science and Information Technology, University College Cork

Abstract

Learning analytics can enable the provision of meaningful feedback based on the collected data, help educators to make decisions with and about learners, and improve learner performance. Student engagement predictions are a key factor in generating feedback for real-time learning analytics applications, such as dashboards. However, most previous work has been based on a heavy deep learning model, which results in challenges for deployment in real-time applications (a resource efficiency requirement in reliable AI). This paper proposes a lightweight deep-learning framework for predicting student engagement in video to address this limitation. The proposed method uses customized MobileNetV2 as the backbone, with an input size of 32 by 32 by 3, to extract features from consecutive video frames. Multi-Scale attention – Residual (MUSER) is used to capture global information and contextual representation of the extracted features. Finally, LSTM examines the temporal variations in video frames and yields the prediction result. We use the DAISEE dataset, the most popular dataset in the learning analytics community, to evaluate the proposed framework. Experimental results demonstrate that the proposed method achieves good accuracy while significantly reducing the model size compared to other approaches.

Keywords: Learning analytics, feedback, lightweight deep learning framework, student engagement predictions, reliable AI.

1. Introduction

Comprehensive data are now gathered as students employ digital learning platforms in their courses. This greatly enhances options to comprehend how students learn better [1]. The learning analytics community analyses this information to ground recommendations and clarifies learning processes [2] for enhanced learning environments [3, 4, 5]. One major issue in this work is the demand for additional, comprehensive data regarding each student's learning procedures to create timely and valuable feedback for teachers and students [6]. Student engagement is a vital priority in this case. Studies consistently indicate a strong correlation between academic success and engagement [7].

Student engagement encloses the learning environment's emotional, cognitive, and behavioural states [8]. Behavioural engagement demands in-class activities, perseverance and emphasizing effort [9], while cognition implicates education skills such as storage, perception, retrieval, and processing [10]. Active participation is controlled by affective conditions [11], whereby negative emotions such as frustration and boredom lead to disengagement, while positive emotions such as interest and happiness improve engagement and focus [12]. Various approaches have been developed to automate determination of engagement in online schooling, classified into computer vision-based and sensor-based methods. Computer vision-based techniques have garnered significant interest, and they have in turn been categorised into video-based and image-based approaches. The image-based methods depend only on spatial data from a single frame or image, which is a notable drawback. Since engagement prediction is spatiotemporal behaviour, video-based strategies have become more widely adopted and efficient for detecting student engagement [13].

Video-based approaches generally employ deep learning (DL) and/or machine learning (ML) techniques. ML techniques create features and use handcrafted patterns for student engagement computation [14], while DL methods understand features from training data, allowing the process to distinguish variations [15]. DL techniques outperform ML in tasks demanding affective condition prediction.

The contributions of this paper are two-fold. Firstly, this study introduces a new lightweight and reliable framework for detecting students' engagement from video by incorporating the strengths of Customized MobileNetV2 with Multi-Scale attention – Residual (MUSER) attention and LSTM in online learning environments. Secondly, we describe an evaluation of our method using comprehensive experiments on bench-marking datasets, including the DAISEE dataset. The experiment results indicate significant improvements in the number of model sizes with only 0.66 M parameters while keeping a good accuracy of 57.49% compared to the state-of-the-art methods.

2. Related work

Student engagement has been a critical subject in academic research. Computer vision (CV) is presented as a practical method for automated student engagement prediction [16]. At the moment, DL is broadly exploited in CV [21] because of its superior performance compared to other techniques [22], [23]. These DL approaches are used for engagement detection, providing video data to convolutional neural networks (CNNs) to predict engagement. In [19], Wu et al. presented a multi-instance learning method in the EmotiW database for engagement detection. Zhu et al. and Huang et al. [17] introduced the Deep Engagement Recognition Network (DERN), which extracts features from OpenFace [18] and combines them with temporal convolution to determine student engagement on an online education platform, reaching a precision of 60% in classification with four-level of engagement. Gupta et al. [24] identified the faces of students with the Max-Margin Face prediction method, and students' faces were categorised found on facial expression into four classes, namely, High Negative Affect, High Positive Affect, Low Negative Affect, and Low Positive Affect, employing an improved version of Inception V-3 Model. Zhu et al [20] introduced a Hybrid Attention-based Gated Recurrent Unit (GRU) network to classify features from the EmotiW database's video and predict the engagement class. Binh et al. [25] attained 80.8% using a training CNN and transfer learning with VGG19 to classify students' engagement based on their posture and gestures in the school. Nezami et al. [26] trained a simple CNN to recognize engagement in private image sets. Schulc et al [27] created a CNNLSTM method that detected nonattention and attention for commercials.

The majority of the databases employed for evaluating student engagement have either not been publicly available or are small, makingcomparison with other methods difficult and has restricted the expansion of this field. To address these challenges, Gupta et al. [28] presented the Database for Affective States in E-Environments (DAiSEE), which was employed for training numerous DL methods involving the Convolutional 3D (C3D) [30], InceptionNet [29] and long-term recurrent convolutional networks (LRCN) [31], reaching 46.4%, and 57.9% accuracy respectively in four classes of engagement prediction. Moreover, student engagement prediction has evolved as one of the demands in Emotion Prediction in the Wild (EmotiW) [33, 34] since 2018, promoting this field's evolution. Authors in [36] enhanced the Inflated 3D Convolutional Network (I3D) [37] and achieved greatest precision in the two-level engagement category, namely not engaged or engaged. Authors in [35] developed a new Gaze-AU-Pose (GAP) unit and employed a Gated Recurrent Unit (GRU) [32] layer and recurrent neural network (RNN) to predict engagement. Liao et al. [40] presented Deep Facial Spatio-Temporal Network (DFSTN) to detect the engagement class during online education, employing the SE-ResNet-50 model for spatial features retrieval and the LSTM along with global attention for temporal data retrieval. The DFSTN approach achieveda precision of 58.84% in four engagement classes on the DAiSEE database. Geng et al. [38] proposed a Convolutional 3D (C3D) model with Focal Loss [39] to predict engagement from videos.

Although the above methods perform excellently on the DAISEE dataset, none is designed for mobile applications because they rely on the heavy deep learning framework. We propose a new lightweight deep learning method to address this gap. Lightweight deep learning models demand fewer computational resources, providing efficient power consumption, allowing them to operate smoothly on mobile devices with limited hardware resources, and advancing reliable AI by

delivering consistent performance on low hardware resources (resource efficiency). Furthermore, lightweight deep learning models are available across multiple users and devices, ensuring the AI system can accommodate many users simultaneously without performance degradation, enhancing overall reliability (scalability of reliable AI).

3. Methodology

Figure 1 presents the proposed model's structure. Frames are taken from each video and transformed into spatial features by improved MobilNetV2. A unique instance of the improved MobilNetV2 model is applied to every frame. Next, MUSER is utilized to learn the extracted features' global information and contextual representation. Next, LSTM is used to understand the temporal information. After finishing the training process, the proposed method recognizes the correlation spatiotemporal features with engagement levels. The details of each component framework are discussed in the following subsection.

3.1. Customized MobileNetV2

Along with the fast development of CNN, the model structure is complicated and requires a lot of calculations. The need for hardware resources is more and more. The training process cannot be developed on resource-constrained systems, so the recent research approach is high-speed and lightweight [41]. One approach is compressing the trained network and getting a corresponding lightweight network. Another strategy is to create a small network for training, and the MobileNet V2 is a suggestive small model. MobileNet V2 compensations accuracy and model size compared to MobileNet V1 and MobileNet V3. MobileNetV2 presents the linear bottleneck and residual structure in the network.

The standard residual architecture lowers the dimension first and then expands it. The residual structure in MobilenetV2 uses the contrary order. First, a 1×1 convolution layer is employed to increase the dimension. Then, a 3×3 deep convolution layer is employed to substitute 3×3 convolution, which can significantly improve the effectiveness and decrease the computation of the model. Lastly, a 1×1 convolution layer is employed to decrease the dimension. In MobilenetV2, the ReLU is removed in the final layer and is replaced with the linear activation function. This approach can handle the issue of data loss. Moreover, extension coefficients are presented to regulate the model size [42]. When the stride is 2, no shortcut is used. In the proposed MobileNetV2 network, the shortcut is added with a 1x1 convolution layer and batch normalization layer in the bottleneck block with a stride of 2 (see Figure 2). The extra residual branch allows further improvement in the network performance.

In the original MobileNetV2, the width multiplier α is a hyper-parameter that changes the number of filters in depthwise convolutions. This multiplier affects all layers except the final convolutional layers. When $\alpha = 1.0$, the network utilizes the original number of filters. When α is less than 1.0, the number of filters in each convolutional layer is scaled down by the factor α. In TensorFlow, pre-trained networks weights are available for $\alpha = 0.35$, $\alpha = 0.5$ and $\alpha = 1.0$ [43]. In the proposed MobileNetV2 network, the value of α is 0.35.

3.2. Multi-Scale attention – Residual (MUSER)

The self-attention mechanism is extremely efficient in sequence-to-sequence learning and significantly enhances many tasks. However, it has its own drawbacks. The attention in deep layers is likely to focus in a single instance, yielding to the problem of expressing long sequences.

Multi-scale attention (MUSE) interprets parallel multi-scale expressing learning on sequence information. The original MUSE has two variants: MUSE-simple and MUSE. MUSE-simple includes the essential concept of a parallel multi-scale sequence describing data, and it encrypts the sequence in parallel, regarding various scales, supported by pointwise transformation and self-attention mechanism. MUSE is made from MUSE-simple and interprets a combination of the self-attention mechanism and convolution for learning sequences describing from various scales.

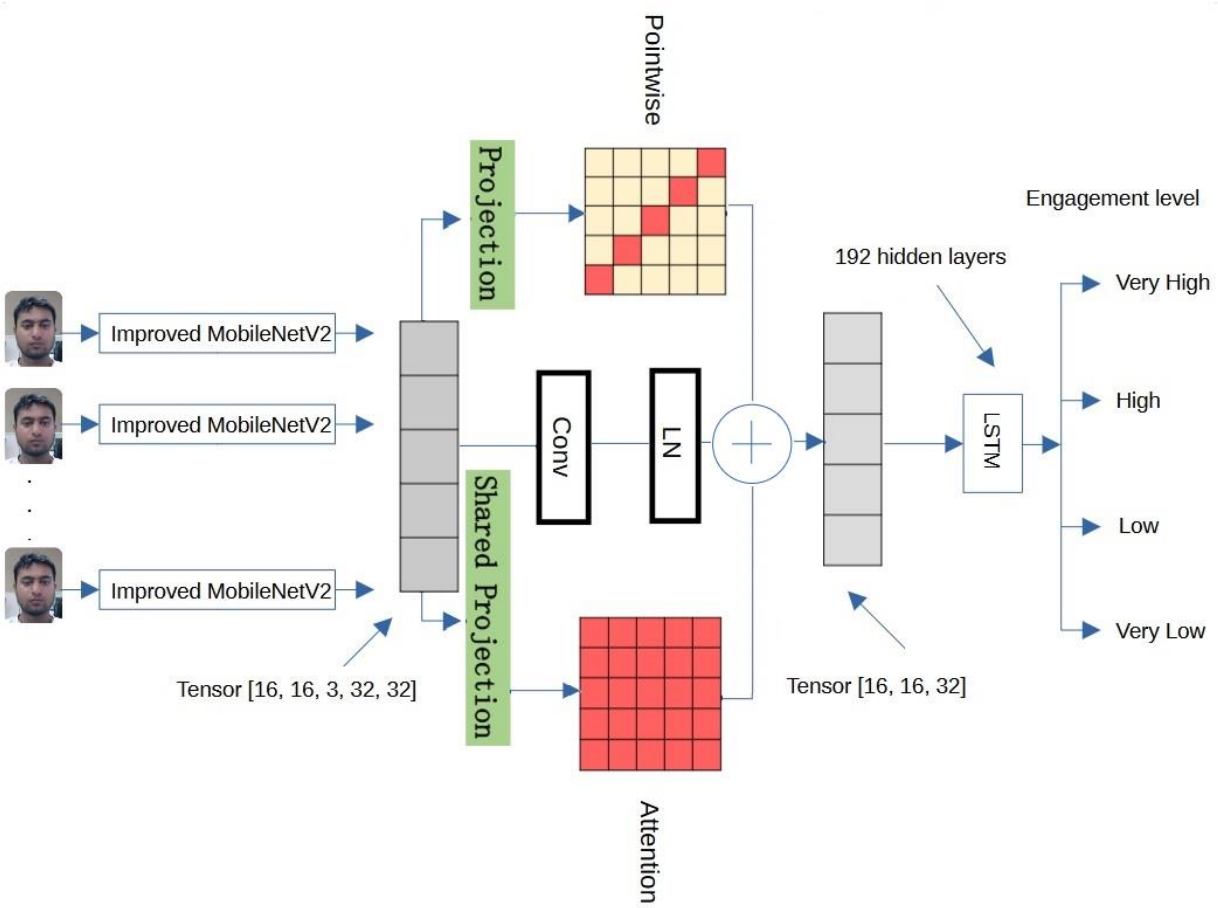

Figure 1: The PROPOSED FRAMEWORK.

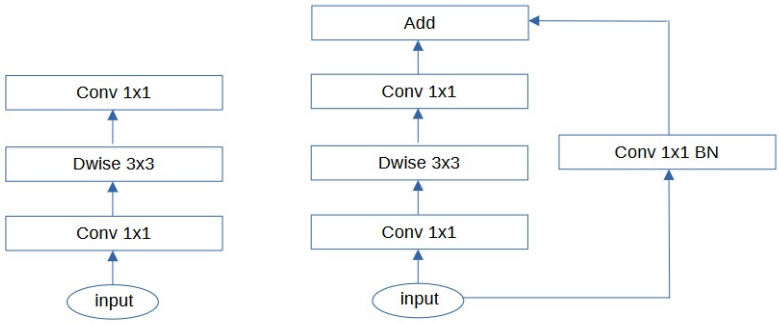

Figure 2: The original bottleneck (left) and the proposed bottleneck (right).

MUSE concentrates on machine translation and has substantially improved performance over Transformer, particularly on long sequences. More significantly, MUSE's success in the application needs multi-scale attention and complex considerations to unify semantic space.

Under a standard setting, the proposed MUSE model surpasses all earlier methods on three principal machine translation tasks. Moreover, MUSE can accelerate inference due to its parallelism [44].

In this work, we extend the MUSE simple with the layer normalization and residual mechanism to create a new attention mechanism, MUSER. The MUSER accepts the output of improved MobileNetV2 as input and creates the output data in a fusion form:

$$Y = F(X) \oplus \text{Attention}(X) \oplus \text{Pointwise}(X) \quad (1)$$

Where “Attention” indicates self-attention, “F(X)” indicates the residual network, and “Pointwise”

indicates a position-wise feed-forward network. The residual network consists of two layers: conv 1x1 and Layer Normalization. As mentioned in the previous section, conv 1x1 changes the channels' number in the feature map. Layer Normalization, a simple normalization technique to enhance the training speed for different neural network models, is used over a mini batch of the output from the conv 1x1 layer. Layer Normalization executes the operation as defined in [45,46]:

$$s = \frac{t - E[t]}{\sqrt{Var[t] + \epsilon}} * \gamma + \beta \quad (2)$$

The standard deviation and mean are computed over the final D dimensions, where D is the dimension of the normalized shape. For instance, if the standardized form is (3, 5), the standard deviation and mean are calculated over the last two dimensions of the input (i.e. input mean ((-2, -1))). γ and β are learnable affine transform parameters of the standardized form if the elementwise affine is True. The typical deviation is computed by the biased estimator.

3.3. L ong short-term memory (LSTM)

LSTM presented as a seminal in [47], is a refined version of the RNN specifically designed to address the typical problem of long-term dependency [48]. LSTMs have proven effective in maintaining data over expanded sequences, thereby resolving the vanishing gradient issue [49]. LSTM processes the result from the prior "time step" along with the current input at a provided "time step" and passes it to the next "time step". Normally, the final step of the last hidden layer is used for recognition [50].

The LSTM framework contains a hidden state described by h, a memory unit symbolised as c, and three different gates: The forget, the input, and the output gate. These gates are essential in managing the input's information flow, thereby controlling the write and read operations within the LSTM framework. The input gate decides how the internal state is updated depending on the current input and the prior internal state. In contrast, the forget gate determines how much of the previous internal state is retained [13].

Lastly, the output gate modulates the effect of the internal state on the overall system [51]. At every time step t, the LSTM obtains an input x_t along with the prior hidden state h_t−1. Later, it computes activations for the gates and updates both the hidden state $h_t$and the memory unit c_t.

The improved MobileNetV2 creates spatial features from single frames, the MUSER improves the learning feature, and the LSTM catches the temporal variations in the sequence of frames and creates the detected engagement class. For a provided video sequence, the input to the proposed framework is an L × C × H × W tensor, where C, L, W, and H correspond to the number of channels, number of frames, frame width, and frame height, respectively. The improved MobileNetV2 is the feature extractor from individual frames of the input video. Then, the MUSER enhances the extracted feature from the output of the improved MobileNetV2. Next, the enhanced feature vectors from the consecutive frames are considered as the multi-dimensional input to the consecutive time steps of an LSTM to catch temporal information in the video. The output of the final time step of the LSTM is fed to a fully connected layer to predict the engagement class from the input video.

MUSER improves the performance of engagement prediction for deep learning networks compared to other attention mechanisms such as Squeeze-and-Excitation Networks (SENet) [52] or Linformer [53]. SENet significantly improves accuracy for existing deep learning networks at a small extra computational cost. Nevertheless, it was designed for image classification and unsuitable for video prediction tasks. Linformer, on the other hand, is created to manage long sequences efficiently, making it appropriate for sequential data like video frames. However, one drawback of Linformer is that it processes sequences in nonparallel. In contrast, the MUSER attention learns sequences in parallel, so it easily captures the change in student behavior in the video data. Also, MUSER uses residual branch, helps mitigate the vanishing gradient problem, and improves features learned [54], leading to better performance than the original MUSE.

4. Experiments

The DAiSEE database, presented in [28], is employed to measure the proposed method's performance compared to the prior approaches. The database has 9,068 videos taken from 112 students to identify their states of confusion, boredom, frustration, and engagement. We just concentrate on engagement, with four engagement stages: stage 0 (very low), 1 (low), 2 (high), and 3 (very high). These states were provided by Whitehill et al. [55]. The frame rate, length, and resolution of the videos are 30 fps, 10 seconds, and 640 × 480 pixels. Table 1 displays the allocation of samples in the training, validation, and testing sets as defined in [28]. Our numerical section uses these sets to fairly compare the proposed framework with prior techniques on the DAiSEE database. We merged 1429 and 5358 videos in the validation and training sets to train the proposed framework model and make predictions on 1784 testing videos [56]. The database is highly imbalanced; only 0.63%, 1.61%, and 0.22% per cent of training, validation, and testing sets are in level 0. The extracted frames from the video were downsampled to get 3 × 32 × 32 images. The Adam is applied for parameter optimization with a batch size and learning rate of 16 and 0.001, respectively. The improved MobileNetV2 extracts feature vectors from the continuous frames and sends them to the MUSER, then LSTM. We executed the experiments using PyTorch [57] on a PC with 32 GB memory and an QUADRO M4000 8GB GPU.

Table 1. The DAISEE database allocation.

	Level 0 Very Low	Level 1 Low	Level 2 High	Level 3 Very High
Total	61	455	4419	3886
Train	46	300	2846	2468
Validate	11	74	712	641
Test	4	81	861	777

The proposed framework's performance and the comparison results with various methods on the DAiSEE database are shown in Table 2. Some of the prior methods on the DAiSEE database, not evaluated with the four-class engagement, showed accurate metrics [28], [40], [17], [19],[38] . In [28], the authors provided five models: frame-level InceptionNet, video-level InceptionNet, C3D fine-tuning, C3D feature extraction, and LRCN. The top precision in [28] was 57.9% with LRCN, outperforming other competitive networks like ResNet TCN with weighted loss and sampling [57].

Table 2. The comparison between various methods on the DAiSEE database.

Model	Accuracy	Parameters
frame-level InceptionNet [28]	47.10%	6.79M
video-level InceptionNet [28]	46.40%	6.79M
ResNet+TCN with weighted loss [56]	53.7%	11M
LRCN (CaffeNet) [28]	57.90%	61M
DFSTN [40]	58.84%	25.6M
EfficientNet B7 + Bi-LSTM [16]	66.39%	66M
EfficientNet B7 + TCN [16]	64.67%	66M
EfficientNet B7 + LSTM [16]	67.48%	66M
The proposed method (Improved MobileNetV2+MUSER+LSTM)	**57.49%**	**0.67M**

The DFSTN [40] reached a precision of 58.84% on the DAiSEE database, which surpassed LRCN by 0.94%. Abedi and Khan [56] presented ResNet TCN with weighted loss and sampling, which achieved an accuracy of 53.7%. Finally, the highest results attained by using the three models: EfficientNet B7 + LSTM [16], EfficientNet B7 + TCN [16], and EfficientNet B7 + Bidirectional LSTM [16] network on the DAiSEE database were 67.48%, 64.67%, and 66.39%, respectively. As seen in Table 2, the proposed method achievedgood accuracy while significantly reducing the model size compared to other approaches. For example, LRCN exceeded the proposed method by only 0.5% but increased the model size almost 100 times.

5. Conclusion

In this paper, we have introduced a novel architecture, Improved MobileNetV2+ MUSER attention + LSTM, for determining engagement levels in online education. We measured the proposed framework's performance and compared it to several state-of-the-art approaches. The proposed architecture significantly reduces the parameters while maintaining good accuracy compared to state-of-the-art approaches on the DAiSEE database.

Student engagement is crucial for tracking the student learning process.The proposed method is lightweight and suitable for mobile applications. Future research is required tointegrate the proposed method into the real-time learning applications.

Acknowledgments

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland, formerly Science Foundation Ireland, under Grant number 12/RC/2289-P2 at Insight Research Ireland Centre for Data Analytics at UCC, which is co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] Vytasek, J. M., Patzak, A., and Winne, P. H. 2020. Analytics for student engagement. Machine Learning Paradigms: Advances in Learning Analytics, 23-48.

[2] Roll, I., and Winne, P. H. 2015. Understanding, evaluating, and supporting self-regulated learning using learning analytics. Journal of Learning Analytics, 2(1), 7–12.

[3] Baker, R. S., and Inventado, P. S. 2014. Educational data mining and learning analytics. In Learning Analytics, Springer, Berlin, 61–75.

[4] Papamitsiou, Z., and Economides, A. A. 2014. Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence. Journal of Educational Technology & Society, 17(4), 49.

[5] Verbert, K., Duval, E., Klerkx, J., Govaerts, S., and Santos, J. L. 2013. Learning analytics dashboard applications. American Behavioral Scientist, 57(10), 1500–1509.

[6] Winne, P. H., Vytasek, J. M., Patzak, A., Rakovic, M., Marzouk, Z., Pakdaman-Savoji, A., et al. 2017. Designs for learning analytics to support information problem solving. https://doi.org/10.1007/978-3-319-64274-1_11.

[7] Ketonen, E. E., Haarala-Muhonen, A., Hirsto, L., Hänninen, J. J., Wähälä, K., and Lonka, K. 2016. Am I in the right place? Academic engagement and study success during the first years at university. Learning and Individual Differences, 51, 141–148.

[8] Pilotti, M., Anderson, S., Hardy, P., Murphy, P., and Vincent, P. 2017. Factors related to cognitive, emotional, and behavioral engagement in the online asynchronous classroom. International Journal of Teaching and Learning in Higher Education, 29(1), 145–153.

[9] Okur, E., Alyuz, N., Aslan, S., Genc, U., Tanriover, C., and Esme, A. A. 2017. Behavioral engagement detection of students in the wild. In Proceedings of the 18th International Conference on Artificial Intelligence in Education (AIED), Wuhan, China, June 28–July 01, 250–261.

[10] El Kerdawy, M., et al. 2020. The automatic detection of cognition using EEG and facial expressions. Sensors, 20(12), 3516. https://doi.org/10.3390/s20123516.

[11] Mukhopadhyay, M., Pal, S., Nayyar, A., Pramanik, P. K. D., Dasgupta, N., and Choudhury, P. 2020. Facial emotion detection to assess learner's state of mind in an online learning system. In Proceedings of the 2020 5th International Conference on Intelligent Information Technology (ICIIT), 107–115.

[12] Gupta, S., Kumar, P., and Tekchandani, R. K. 2023. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools and Applications, 82(8), 11365–11394. https://doi.org/10.1007/s11042-022-13558-9.

[13] Shiri, F. M., Ahmadi, E., Rezaee, M., and Perumal, T. 2024. Detection of student engagement in e-learning environments using EfficientNetV2-L together with RNN-based models. Journal of Artificial

Intelligence, 6. https://doi.org/10.32604/jai.2024.048911.
[14] Bosch, N., et al. 2015. Automatic detection of learning-centered affective states in the wild. In Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI), 379–388.
[15] Zhang, S., Pan, X., Cui, Y., Zhao, X., and Liu, L. 2019. Learning affective video features for facial expression recognition via hybrid deep learning. IEEE Access, 7, 32297–32304. https://doi.org/10.1109/ACCESS.2019.2901521.
[16] Selim, T., Elkabani, I., and Abdou, M. A. 2022. Students engagement level detection in online e-learning using hybrid EfficientNet-B7 together with TCN, LSTM, and Bi-LSTM. IEEE Access, 10, 99573-99583. https://doi.org/10.1109/ACCESS.2022.3206779
[17] T. Huang, Y. Mei, H. Zhang, S. Liu, and H. Yang, ''Fine-grained engagement recognition in online learning environment,'' in Proc. IEEE 9th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC), Sep. 2019, pp. 338–341.
[18] T. Baltrusaitis, P. Robinson, and L. Morency, ''OpenFace: An open source facial behavior analysis toolkit,'' in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2016, pp. 1–10.
[19] J. Wu, B. Yang, Y. Wang, and G. Hattori, ''Advanced multi-instance learning method with multi-features engineering and conservative optimization for engagement intensity prediction,'' in Proc. Int. Conf. Multimodal Interact., 2020, pp. 777–783.
[20] B. Zhu, X. Lan, X. Guo, K. Barner, and C. Boncelet, ''Multi-rate attention based GRU model for engagement prediction,'' in Proc. Int. Conf. Multimodal Interact., 2020, pp. 841–848.
[21] A. Ramamurthy, V. Sathya, M. Rochman, and M. Ghosh, ''ML-based classification of device environment using Wi-Fi and cellular signal measurements,'' IEEE Access, vol. 10, pp. 29461–29472, 2022. https://doi.org/10.1109/ACCESS.2022.3158056
[22] M. He, J. Zhang, S. Shan, M. Kan, and X. Chen, ''Deformable face net for pose invariant face recognition,'' Pattern Recognit., vol. 100, Apr. 2020, Art. no. 107113.
[23] C. Qi, J. Zhang, H. Jia, Q. Mao, L. Wang, and H. Song, ''Deep face clustering using residual graph convolutional network,'' Knowl.-Based Syst., vol. 211, Jan. 2021, Art. no. 106561.
[24] S. Gupta, T. Ashwin, and R. Guddeti, ''Students' affective content analysis in smart classroom environment using deep learning techniques,'' Multimedia Tools Appl., vol. 78, pp. 25321–25348, Oct. 2019.
[25] H. Binh, N. Trung, H. Nguyen, and B. Duy, ''Detecting student engagement in classrooms for intelligent tutoring systems,'' in Proc. 23rd Int. Comput. Sci. Eng. Conf. (ICSEC), 2019, pp. 145–149.
[26] 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,'' in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2019, pp. 273–289.
[27] A. Schulc, J. F. Cohn, J. Shen, and M. Pantic, ''Automatic measurement of visual attention to video content using deep learning,'' in Proc. 2019 16th International Conference on Machine Vision Applications (MVA), IEEE, 2019, pp. 1–6.
[28] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, ''DAiSEE: Towards user engagement recognition in the wild,'' 2016, arXiv:1609.01885.
[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ''Going deeper with convolutions,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 1–9.
[30] 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. (ICCV), Dec. 2015.
[31] J. Donahue, H. L. Anne, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. T. Saenko, and T. Darrell, ''Long-term recurrent convolutional networks for visual recognition and description,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 2625–2634.
[32] Cho, K. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
[33] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, ''EmotiW 2018: Audio-video, student engagement and group-level affect prediction,'' in Proc. 20th ACM International Conference on Multimodal Interaction (ICMI), 2018, pp. 653–656.
[34] A. Dhall, ''EmotiW 2019: Automatic emotion, engagement and cohesion prediction tasks,'' in Proc. 2019 International Conference on Multimodal Interaction (ICMI), 2019, pp. 546–550.
[35] X. Niu, H. Han, J. Zeng, X. Sun, S. Shan, Y. Huang, S. Yang, and X. Chen, ''Automatic engagement prediction with gap feature,'' in Proc. 20th ACM International Conference on Multimodal Interaction (ICMI), 2018, pp. 599–603.

[36] H. Zhang, X. Xiao, T. Huang, S. Liu, Y. Xia, and J. Li. 2019. An novel end-to-end network for automatic student engagement recognition. In Proceedings of the IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), 342–345.
[37] Carreira, J., Zisserman, A., and Vadis, Q. 2017. Action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, 6299–6308.
[38] Geng, L., Xu, M., Wei, Z., and Zhou, X. 2019. Learning deep spatiotemporal feature for engagement recognition of online courses. In Proceedings of the IEEE Symposium on Computational Intelligence (SSCI), June 2019, 442–447.
[39] Lin, T., Goyal, P., Girshick, R., He, K., and Dollár, P. 2017. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), October 2017, 2980–2988.
[40] Liao, J., Liang, Y., and Pan, J. 2021. Deep facial spatiotemporal network for engagement prediction in online learning. Applied Intelligence, 51, 6609–6621.
[41] Zhu, J., Chen, T., and Cao, J. 2019. Siamese network using adaptive background superposition initialization for real-time object tracking. IEEE Access, 7, 119454–119464. https://doi.org/10.1109/ACCESS.2019.2937166.
[42] Zhang, J., Yu, X., Lei, X., and Wu, C. 2022. A novel CapsNet neural network based on MobileNetV2 structure for robot image classification. Frontiers in Neurorobotics, 16, 1007939.
[43] Narduzzi, S., Türetken, E., Thiran, J. P., and Dunbar, L. A. 2022. Adaptation of MobileNetV2 for Face Detection on Ultra-Low Power Platform. In Proceedings of the 2022 9th Swiss Conference on Data Science (SDS), IEEE, 1-6.
[44] Zhao, G., Sun, X., Xu, J., Zhang, Z., and Luo, L. 2019. Muse: Parallel Multi-Scale Attention for Sequence to Sequence Learning. arXiv preprint arXiv:1911.09483.
[45] PyTorch LayerNorm. Available online: https://pytorch.org/docs/stable/generated/torch.nn.LayerNorm.html.
[46] Lei Ba, J., Kiros, J. R., and Hinton, G. E. 2016. Layer normalization. ArXiv e-prints, arXiv-1607.
[47] Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
[48] Shiri, F. M., Perumal, T., Mustapha, N., and Mohamed, R. 2023. A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU. arXiv preprint arXiv:2305.17473.
[49] Barot, V., and Kapadia, V. 2022. Long short term memory neural network-based model construction and Fnetuning for air quality parameters prediction. Cybernetics and Information Technologies, 22(1), 171–189. https://doi.org/10.2478/cait-2022-0011.
[50] Minaee, S., Azimi, E., and Abdolrashidi, A. 2019. Deepsentiment: Sentiment analysis using ensemble of CNN and Bi-LSTM models. arXiv preprint arXiv:1904.04206.
[51] Fang, W., Chen, Y., and Xue, Q. 2021. Survey on research of RNN-based spatio-temporal sequence prediction algorithms. Journal of Big Data, 3(3), 97–110. https://doi.org/10.32604/jbd.2021.016993.
[52] Hu, J., Shen, L., and Sun, G. 2018. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7132-7141.
[53] Wang, S., Li, B. Z., Khabsa, M., Fang, H., and Ma, H. 2020. Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768.
[54] He, K., Zhang, X., Ren, S., and Sun, J. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.
[55] Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., and Movellan, J. R. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. IEEE Transactions on Affective Computing, 5(1), 86–98. https://doi.org/10.1109/TAFFC.2014.2316163
[56] Abedi, A., and Khan, S. 2021. Improving state-of-the-art in detecting student engagement with resnet and TCN hybrid network. In Proceedings of the 18th Conference on Robots and Vision (CRV), 151–157.
[57] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. 2019. Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703.