Facial Emotion Recognition for Virtual Learning Environments to Reduce Dropout Rate

*Meghna Reddy[1], Priyank Kumar[2]

School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, Tamil Nadu, India

[*1]meghna.reddy2016@vitstudent.ac.in
[2]priyank.kumar2016@vitstudent.ac.in

Abstract: *Virtual learning environments (VLEs) are widening opportunities for people of all age groups to learn courses or acquire skills that appeal to them. Recent trends show that there is an increase in the use of such environments and their resources. However, we observe that there is a significant difference between the number of people registered for such courses and the number of people who complete it. We bring about assistance for the tutors monitoring these courses by helping them recognize the emotion of students over these virtual learning environments. By recognizing specific frames that failed to keep the student engaged, we can successfully increase the efficiency of virtual learning environments. We propose to achieve this by recognizing the student's emotions using a convolution neural network (CNN) model throughout the content provided, recognizing the emotion observed at each frame, and drawing detailed reports to identify the dominant emotion. By observing the dominant emotion, the tutor can take necessary actions, which would in turn help reduce the dropout rate.*

Keywords—Virtual learning environments, machine learning, facial emotion recognition, emotion detection, dropout rate

1. Introduction

Virtual Learning Environments (VLEs) are platforms that provide educational tools and informational content on a variety of topics and are available over the internet. The content provided on such platforms is in the form of text, audio, or video files. These are accessed by users of different age groups to learn and acquire new skills. Lately, a stark increase in demand for such virtual provisions was observed, as these online courses are accessible even in remote corners of the world for a minimal fee. However, studies reveal that even though there is an increase in demand and registrations for VLEs, there is a high dropout

rate among students who register for the course or content. This reduces the efficiency of VLEs.

In conventional methods (like classroom environments), the tutor can assess students' interest and engagement levels easily in person by observing their expressions, body language, actions, interactivity, and can alter their methods of teaching accordingly. However, this is quite difficult to monitor in the scenario of VLEs, where there are restrictions on the coverage area of the webcam and limitations on what can be observed.

An effective way to understand a student's engagement level in a particular course is by analyzing their emotions. An important component of understanding emotions involves recognizing the student's facial expression. Determining this would give us an understanding of the student's emotions. For instance, a happy expression indicates a student's interest in the virtual content, whereas an annoyed expression indicates disinterest. These indicators are used to determine the emotion in our proposed methodology.

The use of machine-learning models has increased greatly in the last few years for various purposes. Keeping this in mind, a machine learning model is implemented to carry out emotion detection efficiently and easily.

Machine learning models have many more real-life applications and it is truly revolutionary. Feature extraction and classification have applications in almost every field. Some applications include surveillance and security, marketing, hiring, transportation, banking, social media, medical image computing, and recommender systems.

Our proposal makes use of its applications is in online learning platforms, which are in high demand, as there is a noticeable shift and focus on moving on from conventional methods of learning.

By identifying the emotions induced on the student throughout the course, the course material can be altered to keep the student's interest throughout the course, motivating them to complete the course successfully.

In this paper, we propose a web application to assist the tutor in identifying the content that fails to keep the student engaged, motivating them to take necessary actions to prevent the student from dropping out of the course. We classify different emotions into three broad categories: interested, not interested, and neutral.

The students' emotions are recognized by our CNN model throughout the length of the provided content. The emotions are recorded against each frame and are then visualized through graphs to identify the dominant emotion of the student taking the course. The tutor can then make use of these observations.

2. Literature

Our method of reducing dropout rates from courses offered on Virtual Learning Environments involves identifying a student's expressions and classifying them into one of three categories, i.e., interested, neutral, or not interested. There exist different methods and architectures used to recognize facial expressions since, and has a wide range of applications in real-time. We implement a convolutional neural network to achieve a lower dropout rate as it is quick and efficient.

The utilization of VGG16 [1], a pre-trained Convolutional Neural Network model, was proposed to be used with the Viola-Jones face detection algorithm to detect the facial region

in images captured, and classified using the kNN classifier. The classifier classified the user's emotions into six different categories, i.e., anger, fear, joy, disgust, surprise, and sadness, successfully with an accuracy rate of 86%, with deep feature extraction boosting the accuracy by 4%.

In [2], the KLT tracker was proposed to be used for emotion recognition in video frames along with the HoG feature for feature extraction, combined with the SVM and kNN classifier. The SVM classifier was found to result in higher accuracy than the kNN classifier for the 5 classes of emotions- anger, fear, joy, sadness, and pride.

Reference [3] tested four DBN models to identify if they yielded higher accuracies than other classifiers: DemoFBVP (a simple two-layer DBN model), f+DemoFBVP (a two-layer DBN + feature selection pre-training), DemoFBVP+f (a two-layer DBN + feature selection post-training) and 3DemoFBVP (a three-layer DBN model). Multimodal features are generated with these architectures from body, voice, and psychological data. It was observed that in each case, these DBN models perform better than the SVM classifier by yielding a higher accuracy.

A hybrid model of CNN and RNN was proposed in [4] and tested against simple baseline models, including the CNN model. The integration of CNN and RNN into one model resulted in better performance, and higher accuracy compared to other baseline CNN models.

The Viola-Jones algorithm was paired with a two-module approach (eye detection module and head rotation module) in [5] to determine the concentration levels of the student in each frame. In a 10-minute sample video, the students' emotions were successfully classified into three levels of concentration- medium, low, and high.

A Feedforward Deep Convolution Neural Network (FDCNN) model is proposed in [6] to classify human emotions with body movements over a sequence of frames. The person's webcam is used to capture the body movements along with their facial expressions. The FDCNN model resulted in 95.4% accuracy, which was higher compared to other models tested. The first dataset was classified into emotions of anger, joy, fear, sadness, and pride, and the second dataset was classified into emotions coupled with actions, in different combinations (emotions-happy, sad, angry, untrustworthy, and fear, actions- sitting, walking and jumping).

Image processing techniques play a major role in feature extraction on captured images. The fields of yawn detection were explored in [7] to identify if the driver was drowsy to reduce accident rates significantly. This essentially includes detecting the face, mouth, and eyes of the person in the driver's seat.

Different combinations of feature extraction and emotion intensity recognition are compared in [21] to identify the best combination. Gabor Filters, Histogram of Oriented Gradients (HOG), and Local Binary Pattern (LBP) are the feature extraction models used, and SVM, RF, and kNN are the classifiers used for emotion intensity recognition. Various combinations from these were tested on five different datasets. The highest accuracy of the emotional intensity observed was 97.16% from the LBP+SVM classifier combination.

A three-layer architecture was proposed in [8] that gives rise to the concept of adaptive content on a well-known massive open online course (MOOC) platform named Open EdX, which enables the extended platform to update content such that it is best suited for the student based on their profile. The main aim is to promote the idea of software solutions adapting to the user, and not vice versa.

3. Proposed Work

Our proposal for reducing the dropout rate in VLEs involves three steps. The first step involves capturing the student's emotions through the webcam throughout the content provided and classifying the emotions into one of three categories: interested, neutral, and not interested. The second step involves building an API that visualizes the emotions against each frame and identifies the dominant emotion experienced with the help of graphs. The final step involves calling the API over a web application that enables the tutor to observe multiple students' reports at once.

The student's emotions are captured in real-time through their device's webcam throughout the course content. These produce 48x48 pixel frames that are then translated into their weights. These images are acquired and then preprocessed, after which feature extraction is performed on these images using image processing techniques. The preprocessed data is stored as a dataset and is used to train the convolution neural network model to classify emotions into assigned categories.

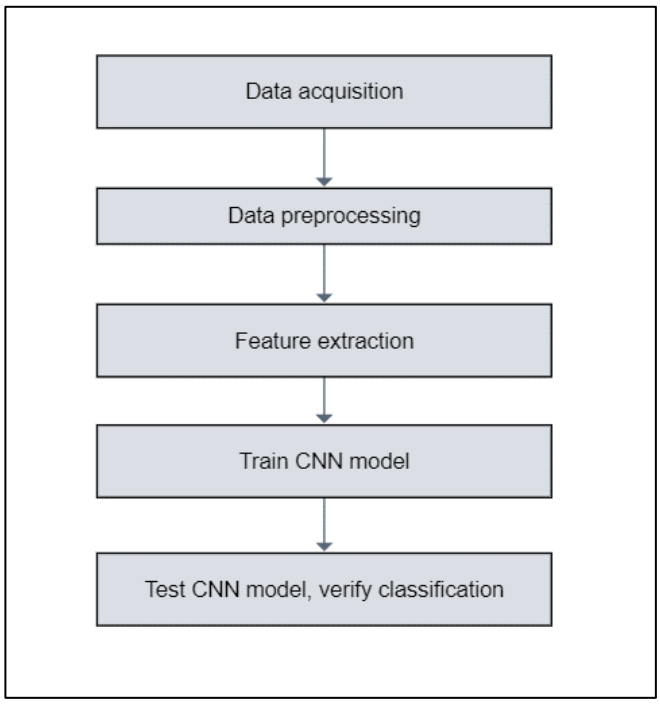

Figure 1. System design

Our second step involves visualizing the emotions onto graphs that track each emotion observed against each frame over the length of the content provided to the student, and indicate which is the dominant emotion. This is achieved using the 'plotly' library in Python.

The third step involves integrating these two modules (the emotion recognizer and the API) into a web application generated using Python's library 'dash'. At the end of the user session, the web application contains the frames vs. emotions graph. This graph is then carefully observed and the weights of each class are used to generate a bar graph simultaneously, indicating which is the dominant emotion. The tutor is allowed to view this dashboard and take necessary measures to keep the student engaged and reduce the dropout

rate. This process of identifying areas that need improvement is easily done using the frames vs. emotion graph. Students whose dominant emotion is 'not interested' are most likely to discontinue the course and dropout.

A. Dataset Description:

The dataset consists of over 35,000 image samples with different emotions. This dataset is split into a training set (80%), validation set (10%), and testing set (10%). The dataset consists of a wide range of emotions and is broadly classified into the following categories:

- **Interested**- positive emotions such as happy or surprised, which indicate students being engaged in the content
- **Neutral**- blank expressions, which don't indicate emotions that could negatively affect the result
- **Not interested-** negative emotions such as annoyed, irritated, or angry, disgusted, fearful, or sad, indicating that the student isn't interested in the content provided in the course.

B. Model Architecture:

We use a Convolution Neural Network (CNN) model to detect the student's emotions. The proposed model has four 2D convolution layers. The input layer containing the original image is connected to the first convolution layer (Conv2D layer), after which a batch normalization layer is attached, and the next convolution layer is stacked on it. These batch normalization layers boost the values of the extracted weights, which fastens the training process. Each layer is followed by a max-pooling layer that down-samples the images and reduces dimensions. This allows extracting features from the images. The combination of these convolution layers, batch normalization layers, and max-pooling layers are then followed by fully connected neural network layers, which operate on the images that have been flattened, further followed by dropout layers to prevent overfitting of the model.
The emotions are then classified into one of the three categories, namely, Interested, Neutral, and Not Interested. The model is then compiled and trained for 17 epochs using the given dataset.

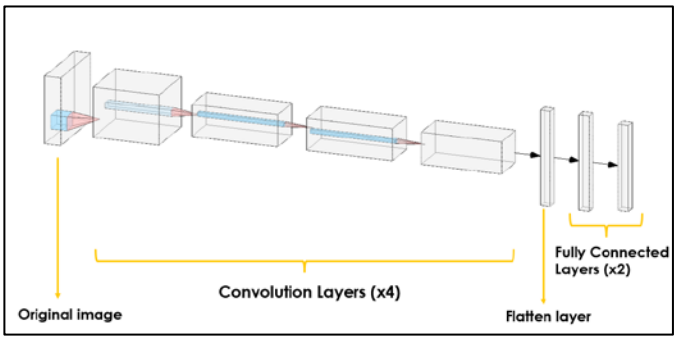

Figure 2. CNN model architecture

4. Experimental Results

The dataset is split into a training set, validation set, and a testing set. After successfully training the CNN model, we were able to classify the student's emotions into one of the three categories:

- Interested
- Neutral
- Not interested

After training the model for 17 epochs, its performance parameters are closely observed. The confusion matrix gives us the number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values.

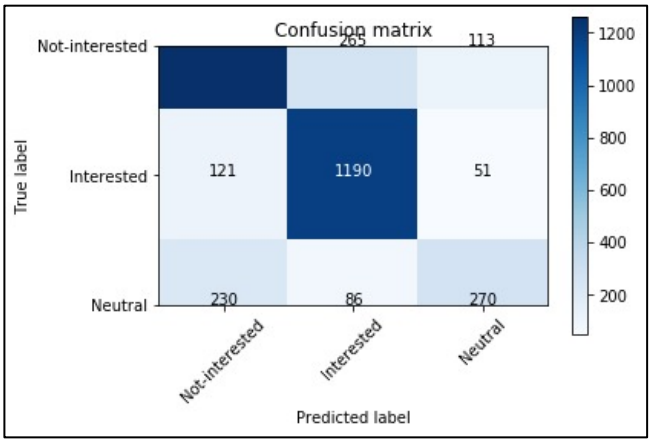

Figure 3. Confusion Matrix of the CNN model

These values are used to calculate various performance metrics such as the precision, recall, and F1 score of each category, and the overall accuracy.

The overall accuracy is the ratio between the correct number of predictions and the total number of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

accuracy			0.759	3589
macro avg	0.726	0.701	0.708	3589
weighted avg	0.752	0.759	0.752	3589

Figure 4. Accuracy of classification

As observed in Figure 4, the overall accuracy of the model is **76%**.

The precision, recall, and F1 scores for each classification are calculated (as shown in Table 1).

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Table 1. Performance matrix of the model on each category

Class	Precision	Recall	F1-Score	Support
Not-interested	0.783	0.770	0.776	1641
Interested	0.772	0.874	0.820	1362
Neutral	0.622	0.461	0.529	586

The model is then tested to determine if the emotions are accurately detected. The observations are shown in Figure 5, Figure 6, and Figure 7.

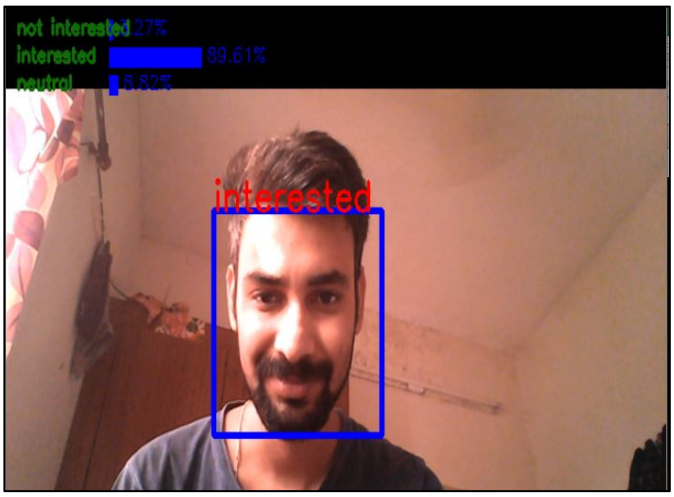

Figure 5. CNN detects student is interested

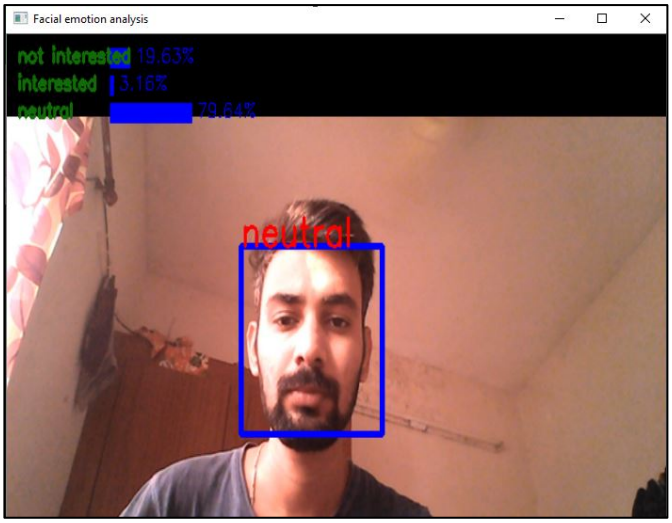

Figure 6. CNN detects student is neutral

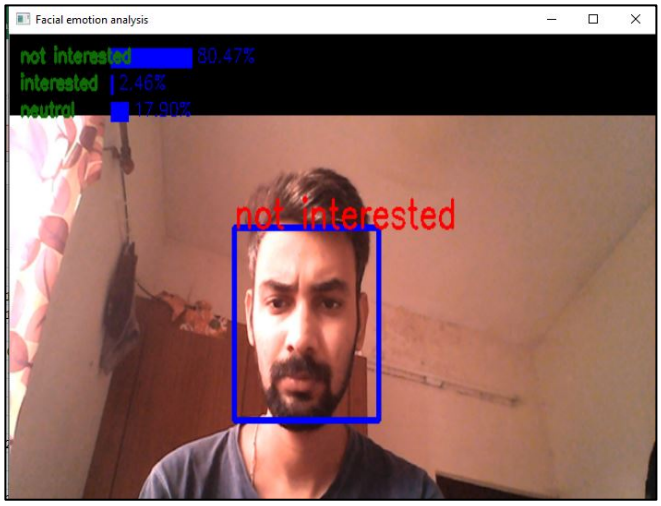

Figure 7. CNN detects student is not interested

Figure 8 compares the student's emotions against every frame on the Y-axis and X-axis respectively for the length of the content provided. The sections of content that failed to engage the student are identified through this graph.

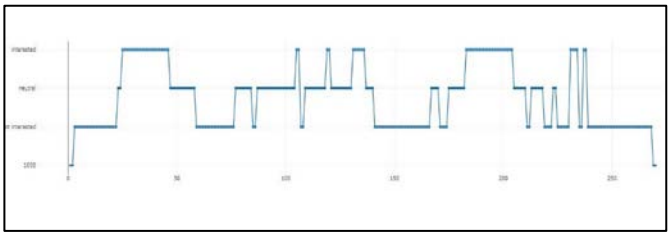

Figure 8. Emotions detected for each frame of content shown

The frequency of these emotions is then represented using a bar graph. This is to visually determine the dominant emotion a student experiences (as shown in Figure 9).

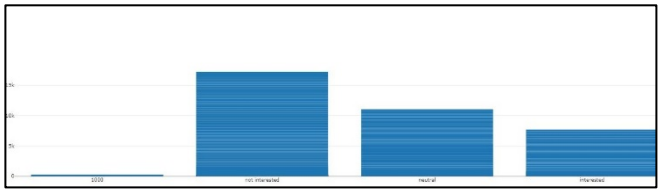

Figure 9. Class vs. Frequency bar graph

The graphs are combined into an API (refer Figure 10) for re-usability and are hosted on a web application. This is created using ‘dash’, a Python library. When the user session is live, the student's emotions are recorded against each frame (as shown in Figure 11).

Figure 10. API containing the graphs

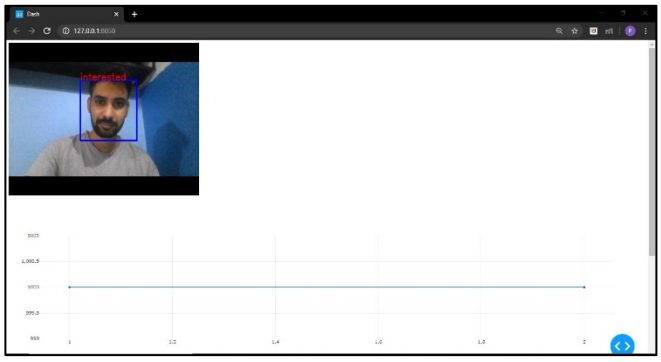

Figure 11. Web application during the user session

Once the session ends, the tutor is allowed to access every student’s report on the faculty login web application using the API imported, consisting of the emotions vs. frame graph and the class vs. frequency bar graph (Figure 12).

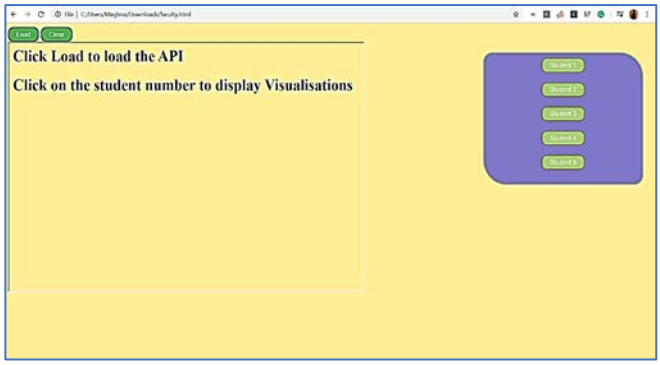

Figure 12. Web application after the user session ends

Each student's report can be accessed by the tutor through this application. If the dominant emotion is 'interested', it indicates the student is engaged, and hence a low chance of the student dropping out. But if 'not interested' is the dominant emotion observed, the tutor can take necessary measures and alter the contents to keep the student engaged for longer. This strategy enables us to reduce the dropout rate for a particular course in virtual learning environments effectively.

5. Conclusion and future work

The proposed CNN model runs with an accuracy of 76% and successfully classifies student's emotions into one of the three categories- Interested, Neutral, and Not Interested. We visualize the student's emotions throughout the content and create an API that can be re-used. A web application is built that can monitor multiple students at once. It contains the API with the frames vs. emotion graph, which portray which areas of the course material failed to keep the student engaged, and a bar graph, which denotes the dominant emotion experienced by the student.

In the future, this application can be deployed on a cloud service and reused for other virtual learning environments. This concept can even be extended to create a tool or widget compatible with various browsers where virtual learning environments are streamed that would automate this process.

6. Acknowledgements

We would like to use this opportunity to express our deepest gratitude and special thanks to Prof. Archana T (Assistant Professor (Sr), VIT), who in spite of being busy with her duties, took time out to guide and keep us on the correct path; she selflessly showed us the right approach through this project. Her belief in us motivated us into bringing our project vision to life and didn't let us get feel disheartened whenever we encountered difficulties during the implementation. We perceive this opportunity as a valuable milestone and strive to use gained skills and knowledge in the best possible way. We will continue to work on their improvement, in order to attain desired career objectives.

7. References

[1] Soltani, M., Zarzour, H., & Babahenini, M. C. (2018, March). Facial emotion detection in massive open online courses. In World Conference on Information Systems and Technologies (pp. 277-286). Springer, Cham.
[2] Mehta, D., Siddiqui, M. F. H., & Javaid, A. Y. (2019). Recognition of emotion intensities using machine learning algorithms: A comparative study. Sensors, 19(8), 1897.
[3] Ranganathan, H., Chakraborty, S., & Panchanathan, S. (2016, March). Multimodal emotion recognition using deep learning architectures. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1-9). IEEE.
[4] Santhoshkumar, R., Geetha, M. K., & Arunnehru, J. (2017). SVM–KNN based Emotion Recognition of Human in Video using HOG Feature and KLT Tracking Algorithm. International Journal of Pure and Applied Mathematics, 117(15), 621-634.
[5] Arguedas, M., Daradoumis, A., & Xhafa Xhafa, F. (2016). Analyzing how emotion awareness influences students' motivation, engagement, self-regulation, and learning outcome. Educational technology and society, 19(2), 87-103.
[6] Jyostna Devi Bodapati, N. Veeranjaneyulu (2019, May). Facial Emotion Recognition Using Deep Cnn Based Features. International Journal of Innovative Technology and Exploring Engineering (IJITEE). Volume-8 Issue-7
[7] Abtahi, S., Hariri, B., & Shirmohammadi, S. (2011, May). Driver drowsiness monitoring based on yawning detection. In 2011 IEEE International Instrumentation and Measurement Technology Conference (pp. 1-4). IEEE.
[8] Sánchez Gordón, S., & Luján-Mora, S. (2015). Adaptive content presentation extension for open edX. Enhancing MOOCs accessibility for users with disabilities.
[9] Khorrami, P., Le Paine, T., Brady, K., Dagli, C., & Huang, T. S. (2016, September). How deep neural networks can improve emotion recognition on video data. In 2016 IEEE international conference on image processing (ICIP) (pp. 619-623). IEEE.
[10] Labarthe, H., Bachelet, R., Bouchet, F., & Yacef, K. (2016). Increasing MOOC completion rates through social interactions: a recommendation system. Research Track, 471.
[11] Krithika, L. B., & GG, L. P. (2016). Student emotion recognition system (SERS) for e-learning improvement based on learner concentration metric. Procedia Computer Science, 85, 767-776.
[12] Alves, P., Miranda, L., & Morais, C. (2017). The Influence of Virtual Learning Environments in Students' Performance. Universal Journal of Educational Research, 5(3), 517-527.
[13] Boulton, C. A., Kent, C., & Williams, H. T. (2018). Virtual learning environment engagement and learning outcomes at a 'bricks-and-mortar' university. Computers & Education, 126, 129-142.
[14] Shapiro, H. B., Lee, C. H., Roth, N. E. W., Li, K., Çetinkaya-Rundel, M., & Canelas, D. A. (2017). Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers. Computers & Education, 110, 35-50.
[15] Arguedas, M., Daradoumis, A., & Xhafa Xhafa, F. (2016). Analyzing how emotion awareness influences students' motivation, engagement, self-regulation and learning outcome. Educational technology and society, 19(2), 87-103.
[16] Chao, L., Tao, J., Yang, M., Li, Y., & Wen, Z. (2015, October). Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (pp. 65-72).
[17] Chen, S., & Jin, Q. (2015, October). Multi-modal dimensional emotion recognition using recurrent neural networks. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (pp. 49-56).
[18] Salunke, V. V., & Patil, C. G. (2017, August). A New Approach for Automatic Face Emotion Recognition and Classification Based on Deep Networks. In 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA) (pp. 1-5). IEEE.
[19] Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016, March). Going deeper in facial expression recognition using deep neural networks. In 2016 IEEE Winter conference on applications of computer vision (WACV) (pp. 1-10). IEEE.
[20] Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., ... & Mirza, M. (2013, December). Combining modality specific deep neural networks for emotion recognition in video. In Proceedings of the 15th ACM on International conference on multimodal interaction (pp. 543-550).
[21] Mehta, D., Siddiqui, M. F. H., & Javaid, A. Y. (2019). Recognition of emotion intensities using machine learning algorithms: A comparative study. *Sensors*, *19*(8), 1897.