National College of Ireland

Classification of Affective States and their Level in a Learning Environment using Neural Networks

MSc Research Project
Data Analytics

Kishan Kumar Bajaj
Student ID: x20131241

School of Computing
National College of Ireland

Supervisor: Dr. Arghir-Nicolae Moldovan

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Kishan Kumar Bajaj
Student ID:	x20131241
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Arghir-Nicolae Moldovan
Submission Due Date:	31/01/2022
Project Title:	Classification of Affective States and their Level in a Learning Environment using Neural Networks
Word Count:	6200
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Classification of Affective States and their Level in a Learning Environment using Neural Networks

Kishan Kumar Bajaj
x20131241

Abstract

Online learning has become the way of life from the last two years due to the ongoing pandemic which has completely disrupted the classroom teaching experience and the implicit feedback loop between the teacher and the students. Like technology is being used to maintain continuity in learning, it can also be used to create a feedback mechanism between the teacher and student in an online learning environment. The key to develop this is detection of different affective states and their levels exhibited by the students during learning. In this research, five different models are presented; one model to classify the correct out of four affective states (boredom, engagement, confusion and frustration) and for each affective state a separate model to classify the correct level out of four levels of each affective state (very low, low, high, very high). All these models are built upon a hybrid ResNet + TCN neural network architecture and are trained using publicly available DAiSEE dataset. The data set contains videos which are converted to sequence of frames for training the model. Another dataset taken from EmotiW2020 Challenge is also used to cross validate the engagement level classification model. The affective state classification model and the boredom level classification model outperform existing works. Confusion and Frustration level models perform almost at par with the existing models. Engagement level classification model performs at par with other models but considering the fractional amount of training data and iterations used, the performance can be considered as good as the existing baseline models. This model is assessed on, both, DAiSEE and EmotiW2020 data sets and achieves similar performance on both data sets.

1 Introduction

The last two years have been an eye opener for the whole world, disrupting industries and the way of life in ways unimaginable. One such aspect of the global community that has turned upside down is classroom teaching which has heavily impacted the student community, for better or worse is yet to be known. The positive aspect is that given the rapid advancements in the field of communications technology, the teachers and students across the globe were able to continue learning from the boundaries of their rooms which were enforced by endlessly recurring lockdown restrictions.

Nambiar (2020) conducted a survey in India in 2020, with 70 teachers and more than 400 students participants, which reported that more than 55 teachers, had the opinion that students lack involvement and interest in online classes and are not taking it seriously. More than 240 student participants of the survey support this and have

reported difficulty in concentration in online classes and frequent distractions. This is a serious problem for the global learning community since online learning is the only way to go until the pandemic subsides which nobody can say for certain. Dewan et al. (2019) in their have thoroughly detailed the utilisation of computer vision technologies to use the facial expression of the students in real time and predict their level of engagement in a cost-effective, real time and non-intrusive manner. This can further be expanded to predict the level of other important affective states that are critical in a learning environment.

The question this research project addresses is – Can a pretrained neural network classify affective states and their level on different data sets with consistent accuracy?To address this research question, following research objectives are prescribed:

- Identify two or more data sets which can be cross examined for affective state classification or affective state level classification.
- Investigate previous research conducted using the identified data sets or similar data sets.
- Train the existing state-of-the-art model on one of the appropriate data sets.
- Evaluate the performance of the trained model for all the identified data sets.
- Benchmark the performance of the model using suitable parameters from previous research and present a detailed comparison with the previous works using those parameters.

To successfully achieve the above objectives and answer the research question, this research proposes to use two publicly available data sets, Dataset for the Affective States in E-Environments (DAiSEE), Gupta et al. (2016), which is used to train a neural network for classification of Affective States and the Level of each Affective State, and Engagement prediction in the wild data set from the Eighth Emotion Recognition in the Wild Challenge (EmotiW2020), Kaur et al. (2018), which will be used along with DAiSEE to evaluate the model which is trained on DAiSEE. DAiSEE consists of more than 9000 videos of 112 subjects in a learning environment and each video being 10 second long. The videos are recorded while performing learning activity in multiple locations to train the model efficiently for real life use cases. EmotiW2020 consists of 195 videos of 78 subjects and each video is 3 to 7 minutes long and is recorded over Skype.

This research work contributes multiple novelties to the domain of affective states and their level. For the first time in the domain of engagement level classification, the model, built for classifying engagement level into four levels (0 - very low, 1 - low, 2 - high and 3 - very high), is trained on one data set (DAiSEE) has been cross evaluated on a completely different data set (EmotiW2020) and has successfully produced similar results (50% accuracy and 28% unweighted average recall) for both data sets. Furthermore, this model uses 85% less training data (53580 frames vs. 354350 frames) and 90% less model training iterations (10 vs. 100) and still achieves 28% unweighted average recall score compared to 33.6% in the current state-of-the-art model. The affective state classification model, which classifies four different affective states (0 – Boredom, 1 – Engagement, 2 – Confusion, 3 - Frustration), achieves the highest accuracy of 80.9% significantly outperforming 53.4% accuracy state-of-the-art model. The Boredom level classification model is also the best performing model with an accuracy of 48% when compared to the model of DAiSEE dataset creators with an accuracy of 36.5%.

The rest of the paper is structured as follows: Section 2 reviews and summarizes the relevant research works conducted previously in the domain of affective states and their

level along with the public data sets available. Section 3 elaborates in detail each step of the proposed research methodology for successful completion of this research. Section 4 describes the design of the hybrid model used in this research. Section 5 lists out various parameters used in the implementation of this research along with the hardware and software configurations. Section 6 evaluates each experiment conducted in this research using relevant metrics and observations and compares the outcomes with previous works. Section 7 concludes this research work and lays out possible future enhancements.

2 Related Work

This section critically reviews the previous works carried out in the field of affective states and their levels and is further divided into sub sections; Affective States 2.1, Level of Affective States 2.2, Public Data Sets 2.3 and Summary 2.4.

2.1 Affective States

After analyzing 24 studies in multiple fields of technology enhanced learning such as simple computer interfaces, intelligent tutoring systems, simulation environments and serious games, D'Mello (2013) identifies the affective states experienced by the students. Engagement/flow, boredom, and confusion affective states were the most frequent, while the frequency of affective states such as frustration, curiosity and happiness varied significantly. Baker et al. (2010) emphasizes on detection of boredom and confusion as boredom and poor learning were found to be closely associated and confusion was one of most frequent affective states within learning environment such as Intelligent Tutoring systems, problem solving games and dialogue tutors. As identified by D'Mello (2013), Baker et al. (2010) also did not find the frustration affective state to be as persistent as confusion and engagement and also weakly associated with poor learning.

Leong (2020) presents a FaceNet Embedding and Long short-term memory (LSTM) based approach to detect boredom and frustration on DAiSEE dataset, using separate models for each of the two, with an accuracy of 52.15% and 70.67%, respectively. Classification of affective states on multiple databases, such as CK+, JAFFE and DAiSEE, is presented by Rao and Rao (2020) using a hybrid CNN model with manually engineered features and has achieved an accuracy of 99.95%, 71.4% and 53.4%, respectively.

2.2 Level of Affective States

Binomial classification of level of Engagement affective state on DAiSEE dataset, whether engaged or not engaged, is presented by Dash et al. (2019) where a custom CMConv model achieved an accuracy of 91.1%. The same problem is attempted using a 3-dimensional CNN model by Werlang and Jaques (2021) and an accuracy of 82.1% is achieved. Murshed et al. (2019) proposed a custom CNN based model to classify the level of engagement affective state into 3 categories, namely, no engagement, normal engagemment and high engagement, and still managed to achieve higher accuracy, 92.3%, than Dash et al. (2019) and Werlang and Jaques (2021) who used only 2 classes for the same problem in their research. El Kerdawy et al. (2020) also presents a CNN based model to classify engagement levels into engaged and not engaged classes, but on a custom private data set and not on DAiSEE data set, with a F2 score of 0.82.

Multiple classification studies, Liao et al. (2021), Abedi and Khan (2021a), Dresvyanskiy et al. (2021), Gupta et al. (2016) and Abedi and Khan (2021b), have been carried out on DAiSEE data set to classify the level on engagement affective across four different levels, namely, very low (0), low (1), high (2) and very high (3), as summarized in Table 1. The creators of DAiSEE data set, Gupta et al. (2016), used a long-term recurrent convolutional network (LRCN) based model to perform this classification and achieved an accuracy of 57.9%. This was followed by Liao et al. (2021) improving it by reaching an accuracy of 58.8% by implementing a Deep Facial Spatiotemporal Network (DFSTN) based classification model. Abedi and Khan (2021a) used latent affective, behavioral and affect features with a temporal convolutional network (TCN) model to reach an accuracy of 63.3% and then further improved the accuracy to 63.9% by implementing a hybrid model consisting of Residual neural network (ResNet) and TCN in Abedi and Khan (2021b). Dresvyanskiy et al. (2021) uses an ensemble architecture based model consisting of OpenFace, EmoVGGFace2, VGGFace2-SA and Recurrent neural network (RNN) and achieve a very high unweighted average recall (UAR) score of 44.3% even though the accuracy is only 39.02%.

Regression studies are also carried out to predict the intensity of engagement, on a scale of 0 (low) to 1 (high), on EmotiW2020 data set by Liao et al. (2021) and Abedi and Khan (2021a), as seen in Table 1. Liao et al. (2021) achieves a mean square error (MSE) score of 0.0736 by using a DFSTN based regression model while Abedi and Khan (2021a) uses a LSTM based model with behavioral and affect features and achieves a better MSE score of 0.0673.

Even though, engagement is the most popular affective state when it comes to predicting the level of affective state, Gupta et al. (2016), who are indeed the original creators of the DAiSEE data set, have also presented the classification of level of other affective states such as boredom, confusion and frustration on it across the same four levels, namely, very low (0), low (1), high (2) and very high (3), using five different frame and video based models. The accuracy of the best performing model, i.e. LRCN, for classifying the level of each of the three affective states in listed in Table 1.

2.3 Public Data sets

Dataset for affective States in E-Environments (DAiSEE), created by Gupta et al. (2016) and Kamath et al. (2016), is used in a lot of recent studies, as reviewed in previous sub sections, related to student engagement and affective state detection. This data set provides more than 9000 video clips which are meticulously labelled using the wisdom of the crowd for four affective states, namely, boredom, engagement, confusion and frustration and each affective state is further labelled for its level again categorized into a scale of four levels; very low (0), low (1), high (2) and very high (3).

Another similar data set is "Engagement prediction in the Wild" data set from the Eighth Emotion Recognition in the Wild Challenge (EmotiW2020) which is created by Kaur et al. (2018). The publicly available version of this data set provides around 200 video clips which are labelled for four levels of intensity of engagement; completely disengaged(0), barely engaged(1), engaged(2) and highly engaged(3).

Table 1: Research studies realted to classification and prediction of Affective States and Level of Affective State on DAiSEE and EmotiW2020 data set.

Research Work	Data set	Classes	Best Model	Accuracy
Liao et al. (2021)	DAiSEE	Engagement Level 0/1/2/3	DFSTN	58.8%
	EmotiW2020	Regression		MSE - 0.0736
El Kerdawy et al. (2020)	Custom	Engaged/Not Engaged	CNN	F2 - 82.0%
Abedi and Khan (2021a)	DAiSEE	Engagement Level 0/1/2/3	(Latent Affective + Behavioral + Affect) features + TCN	63.3%
	EmotiW2020	Regression	Clip-level (Behavioral + Affect) features + LSTM	MSE - 0.0673
Dresvyanskiy et al. (2021)	DAiSEE	Engagement Level 0/1/2/3	OpenFace + EmoVGGFace2 + VGGFace2-SA + RNN	UAR - 44.3%
Murshed et al. (2019)	DAiSEE	Engagement Level No/Normal/High	Custom CNN based	92.3%
Gupta et al. (2016)	DAiSEE	Boredom Level 0/1/2/3	LRCN	53.7%
		Engagement Level 0/1/2/3		57.9%
		Confusion Level 0/1/2/3		72.3%
		Frustration Level 0/1/2/3		73.5%
Dash et al. (2019)	DAiSEE	Engaged/Not Engaged	CMConv	91.1%
Abedi and Khan (2021b)	DAiSEE	Engagement Level 0/1/2/3	ResNet + TCN	63.9%
Werlang and Jaques (2021)	DAiSEE	Engaged/Not Engaged	3D CNN	82.1%
Rao and Rao (2020)	DAiSEE	Boredom/Engagement/ Confusion/Frustration	CNN (uptoFC) + Handcraft_features with SVM	53.4%

2.4 Summary

Studies related to classifying or predicting level of affective states are mostly restricted to the engagement affective state despite the emphasis laid by D'Mello (2013) and Baker et al. (2010) on the significance and impact of other affective states like boredom, confusion and frustration. Possible reason for this could be, as also pointed out by Nezami et al. (2020), the unavailability of sufficient amount of good quality training data and the debatable annotations. Gupta et al. (2016) and Kaur et al. (2018) have made great contribution in an effort to address this problem by making systematically annotated large amount of data publicly available for frutiful research in the field of affective state detection and measuring the level of affective state.

3 Methodology

This research follows the Knowledge Discovery in Databases (KDD) methodology. Each phase of this research, according to the KDD methodology, is illustrated in detail in Figure 1 and is further elaborated in the following sub sections; 3.1 Data Selection, 3.2 Data Preprocessing and 3.3 Data Transformation, 3.4 Data Modelling and 3.5.

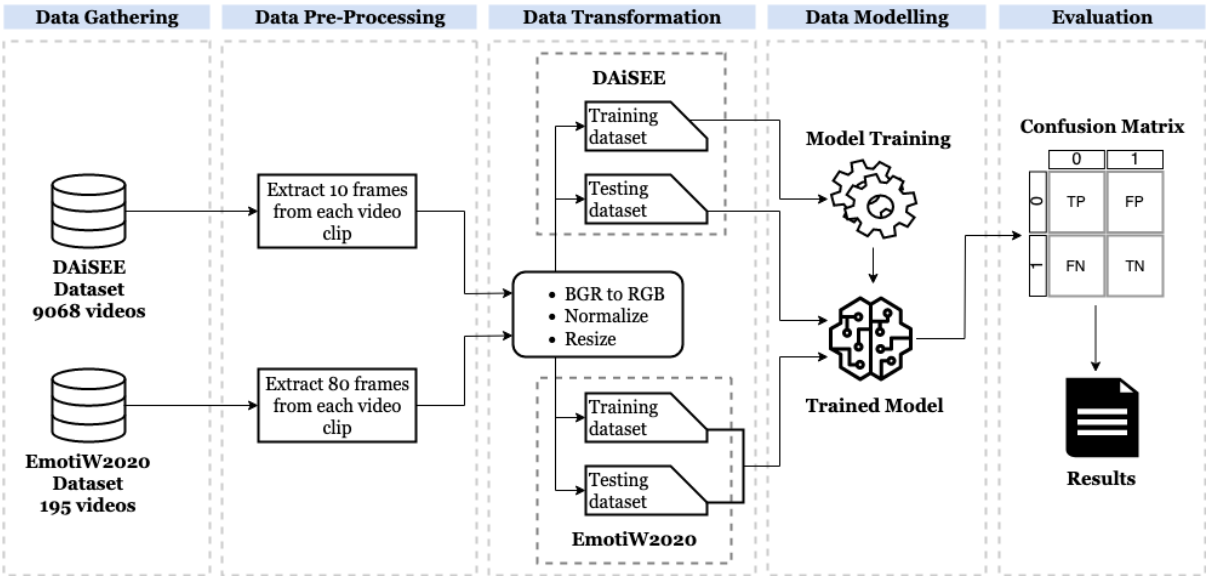


Figure 1: High Level illustration of KDD based methodology adopted to conduct this research

3.1 Data Selection

3.1.1 DAiSEE

Dataset for Affective States in E-Environments (DAiSEE), Gupta et al. (2016) and Kamath et al. (2016), is selected for this research which is available to download from Indian Institute of Technology Hyderabad's website[1]. DAiSEE contains video clips of subjects recorded in an e-learning setting and closely imitates "in the wild" environment settings. The dataset is annotated using crowd sourcing and each video clip is labelled for one of the four levels, namely, very low(0), low(1), high(2) and very high(3) for four engagement related affective states, namely, boredom, confusion, engagement and frustration.

It consists of 9068 video clips and each clip is of 10 seconds duration. These video clips are of 112 subjects who are between 18 to 30 years old of Asian race with a female to male ratio of 2:5. The resolution of the video clips is 640x480 pixels and the frame rate is 30 frames per second (fps). The video clips are recorded in 6 varied locations like laboratories, dormitory rooms, library, etc. under light, neutral and dark illumination conditions. All these features provide huge amount of data with a lot of variety for training state of the art neural network based deep learning models.

[1] `https://iith.ac.in/~daisee-dataset/`

3.1.2 Engagement Prediction in the Wild from EmotiW2020

Engagement prediction in the wild data set from the Eighth Emotion Recognition in the Wild Challenge (EmotiW2020), Kaur et al. (2018), is also selected for one of the experiments of this research. The data set can be obtained from EmotiW2020's website[2]. This data set also contains video clips of subjects recorded in various environments like hostel rooms, open grounds and computer laboratories to simulate "in the wild" environment. The data set is annotated by a team of 5 annotators and each video clip is labelled on the basis of four levels of intensity of engagement, namely, completely disengaged(0), barely engaged(1), engaged(2) and highly engaged(3).

This data set consists of 195 video clips and each clip varies in duration from 3 to 7 minutes. These video clips are of 78 subjects who are between 19 to 27 years old with a female to male ratio of 1:2. The resolution of the video clips is 640x480 pixels and the frame rate is 30 frames per second (fps) but since the video is recorded over Skype the frame rate is not constant and suffers from frame drops. This data set is highly similar to the DAiSEE data set and will be very insightful to evaluate the model which is trained on DAiSEE data set.

3.2 Data Preprocessing

3.2.1 DAiSEE

All 9068 video clips, of 10 seconds each, are programmatically parsed using OpenCV library in Python and 10 frames are extracted, with an extraction frequency of 1 frame per second, from each video clip as seen in Figure 2. This process is carried out for all three splits of the data, training, test and validation.

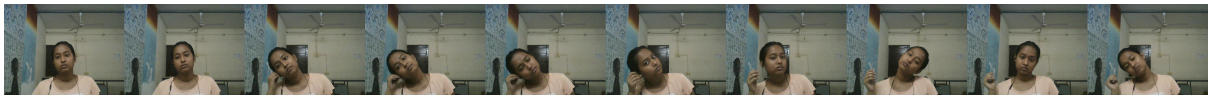


Figure 2: Ten different consecutive frames obtained from one of the video clips

The annotation labels for each video clip are originally provided in comma separated values format files for all 3 splits, training, test and validation. For each affective state, a separate set of training, test and validation csv files is created containing the absolute path of the folder containing the extracted frames from each video and respective label for the level of that affective state. After this process a total of 4 sets of training, test and validation csv files is available, one for each affective state, namely, boredom, confusion, engagement and frustration. The distribution of samples for each class (level) in training and test split of each set is tabulated in Table 2.

One more set of training, test and validation csv files is created for classification of affective state instead of the level of affective state. This is done by selecting the most dominant affective state in a video clip which is determined by identifying the affective state with highest level among the four. For example, if a video clip has low (1) level of boredom, high (2) level of engagement, very low (0) level of confusion and very low (0) level of frustration, then in this case the affective state of the video clip is labelled as engagement as engagement affective state has the highest level among all the four affective

[2] `https://sites.google.com/view/emotiw2020/challenge-details`

Table 2: Sample distribution for each level of the four affective states in training and test split of DAiSEE data set

Level	Affective State							
	Boredom		Engagement		Confusion		Frustration	
	Samples	Class Percentage	Samples	Class Percentage	Samples	Class Percentage	Samples	Class Percentage
Training Split								
0	2433	45.4%	34	0.6%	3616	67.5%	4183	78.1%
1	1696	31.7%	213	4.0%	1245	23.2%	941	17.6%
2	1073	20.0%	2617	48.8%	431	8.0%	191	3.6%
3	156	2.9%	2494	46.6%	66	1.2%	43	1.0%
Test Split								
0	823	46.1%	4	0.2%	1200	67.2%	1388	77.8%
1	584	32.7%	84	4.7%	427	23.9%	316	17.7%
2	338	19.0%	882	49.4%	136	7.6%	57	3.2%
3	39	2.2%	814	45.6%	21	1.2%	23	1.3%

states. The affective states are encoded as; boredom (0), engagement (1), confusion (2) and frustration (3). The distribution of samples for each class (affective state) in training and test split is tabulated in Table 3.

Table 3: Sample distribution for each affective state in training and test split of DAiSEE data set

Affective State	Training Split		Test Split	
	Samples	Class Percentage	Samples	Class Percentage
0 - Boredom	1023	19.1%	317	17.8%
1 - Engagement	4245	79.2%	1432	80.3%
2 - Confusion	56	1.1%	18	1.0%
3 - Frustration	34	0.7%	17	1.0%

3.2.2 EmotiW2020

Similar to DAiSEE data set, video clips in this data set are also programmatically parsed using OpenCV library in Python and frames are extracted from each video clip. Since, the video clips in this data set are not of equal duration and vary from 3 minutes to 7 minutes in length, the frames are extracted using a step size of 30 given the videos exhibit a frame rate of 30 frames per second. Using this method, the minimum number of frames obtained from each video clip is 80 and hence we use 80 frames, across the length of the video, from each video clip for our experiment. The original data is already split in the ratio of 75:25 for training:test and the frames are extracted for both the splits respectively. The annotation labels for each video clip are provided in comma separated values format files for, both, training and test split. The distribution of samples for each level of engagement in training and test split is tabulated in Table 4.

3.3 Data Transformation

For both data sets, post preprocessing, the color space of extracted frames is transformed from blue-green-red (BGR) to red-green-blue (RGB) using OpenCV library in Python

Table 4: Sample distribution for each level of engagement in training and test split of Engagement Prediction in the Wild data set

Engagement Level	Training Split		Test Split	
	Samples	Class Percentage	Samples	Class Percentage
0	4	2.7%	4	8.5%
1	35	24.0%	9	19.2%
2	89	54.1%	19	40.4%
3	28	19.2%	15	31.9%

because OpenCV uses the BGR color space and the input images need to be of RGB color space. After the color space transformation frames are normalized using the PyTorch Torchvision library in Python to get the data within a range and reduce the skewness for more efficient and faster learning. Since the images have to be loaded in the range of [0,1] the normalization transformation is carried out using a mean of 0.485, 0.456 and 0.406 and a standard deviation of 0.229, 0.224 and 0.225 for red, green and blue channels, respectively. After normalization, frames are resized from width×height of 640×480 pixels to width×height of 224×224 pixels as this is the minimum height and width required for the input images.

3.4 Modelling

Post transformation, the training split of DAiSEE dataset along with each of the five sets of input csv files is used for training the model, as seen in Figure 1, for respective experiments which are, Active State Classification, Boredom Level Classification, Engagement Level Classification, Confusion Level Classification and Frustration Level Classification. For each experiment there is a separate train and test csv file. The train csv file contains the path of the folders with frames that are to be used for training and the class label against each path for the respective experiment. Similarly the test csv file contains the path of the folders with frames that will be used to evaluate the model trained with the train csv files and check if the class label against each path for that experiment matches the output of the model. The csv files serve as a pointer to the location of the frames and contain the class label for each set of frames.

The model consists of ResNet + TCN hybrid architecture classifier which is explained in detail in Section 4. The test split of DAiSEE dataset is used to evaluate the model after training. For Engagement Level Classification, the train and test split of Emotiw2020 data set is also used along with DAiSEE test split for evaluation.

3.5 Evaluation

This research is evaluated on the basis of multiple standard measures such as model accuracy percentage, training loss, class wise precision and recall score and unweighted average recall (UAR) score. Accuracy or classification rate is the number of correctly classified samples divided by the total number of samples. Correctly classified samples is the sum of both True Positives and True Negatives while total number of samples is the sum of True Positives, True Negatives, False Positives and False Negatives. UAR measure is useful in imbalanced data set classification to detect if one or more classifiers are not upto the mark, as also stated in Dresvyanskiy et al. (2021), but it does not consider false

positives. For a multi-class classification problem, as seen in Figure 3, UAR is the mean of Recall measures of each class, the formula for which is as below.

Recall is the number of correctly classified positive samples divided by the total number of positive samples. Correctly classified positive samples is True Positives and total number of positive samples is the sum of True Positives and False Negatives. Precision is the number of correctly classified positive samples divided by the total number of predicted positive samples. Correctly classified positive samples is True Positives and total number of predicted positive samples is the sum of True Positives and False Positives. High recall and low precision indicate that most of the positive examples are correctly recognised but there are a many false positives. On the other hand, low recall and high precision indicate that a lot of positive samples are missed to be classified but the predicted positive samples are indeed positives.

		Predicted Labels			
		Class 0	Class 1	Class 2	Class 3
True Labels	Class 0	TRUE POSITIVES	FALSE NEGATIVES	FALSE NEGATIVES	FALSE NEGATIVES
	Class 1	FALSE POSITIVES	TRUE NEGATIVES		
	Class 2	FALSE POSITIVES		TRUE NEGATIVES	
	Class 3	FALSE POSITIVES			TRUE NEGATIVES

Figure 3: Confusion Matrix for multiple classes depicting the True Positives, True Negatives, False Positives and False Negatives for Class 0

$$Precision(R) = \frac{TP}{TP + FP}$$

$$Recall(R) = \frac{TP}{TP + FN}$$

$$UAR = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{R_1 + R_2 + \cdots + R_n}{n}$$

Other measures such as confusion matrix and model training history graphs showing accuracy vs. loss for each of the 10 epochs are also used to further deep dive into the class wise performance of the model and assess various aspects of it.

Confusion matrix is used for visualizing the performance of the algorithm and helps in identification of mislabelling between the classes.

4 Design Specification

Each classification problem in this research is solved using a hybrid neural network architecture, Abedi and Khan (2021b), which is made up of a Residual Network Szegedy et al. (2015) and a Temporal Convolutional Network (TCN) Bai et al. (2018). The ResNet architecture uses the frames generated from the video clips to extract spatial features from them. On top of the spatial features a TCN is built for modelling the variation in temporal features since the affective states and their levels can vary from one frame to another in the same video clip.

This unique architecture does not require handcrafting and features or manually extracting them from images or video clips because when the network is trained in a joint manner, it automatically learns the features. The ResNet is trained to extract features post removal of the final layer of ResNet18 which is a fully connected layer. It takes the input from sequence of video frames in the form of tensor [N (number of frames), C (number of color space channels), H (height of each frame), W (width of each frame)].

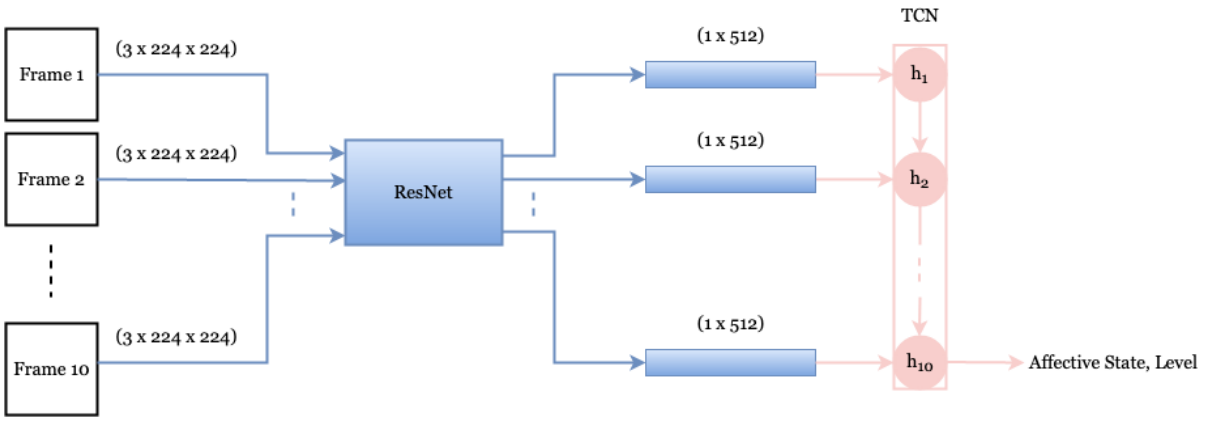


Figure 4: Architecture of the ResNet + (TCN) Hybrid Neural Network which takes a sequence of frames from a video as an input and predicts the class of the video as per training.

The temporal features in the sequence of video frames are modelled using a dilated temporal convolutional network which uses the feature vectors from the sequence of frames as input. The fully connected layer takes input from TCN's final time step and predicts the class using a softmax function.

While training the model, in each epoch random samples are used from the training data due to which samples from classes with low number of samples might not be used. This is likely to happen in most of the iterations which can affect the classification results of the model. To address this issue, custom sampling is used during training so that samples from each class are included in all batches of Stochastic Gradient Descent optimizer (SGD) during the model training process.

5 Implementation

The inputs to the ResNet and TCN network were tensors of order [[10 (number of frames), 3 (number of color space channels), 224 (height of each frame), 224 (width of each frame)]] which means a sequence of 10 frames of dimensions 3×224×224, which is the standard input expected by the ResNet18 architecture Paszke et al. (2019). Along with this a Stochastic Gradient Descent (SGD) optimizer is used with a learning rate of 0.001 and batch size of 4 which helps in optimizing the parameters. The sequence of frames are used for extraction of feature vectors with a dimension of 512 by ResNet which is further used by TCN. For best results Bai et al. (2018), 8 levels, 128 hidden units, dropout of 0.25 and kernel size of 7 is used for TCN parameters. Each model is trained for 10 epochs and the model developed in each iteration is saved. The evaluation of the model is done by analyzing the training loss and training accuracy using the model training history graphs.

Hardware configuration used for this implementation is; CPU: Intel Core 8^{th} Generation i5-8250 CPU @ 1.60GHz, RAM: 16 GB, GPU: NVIDIA GeForce MX150 2 GB.

Software and libraries used for this implementation are; Operating System: Windows 10, Programming Language: Python 3.8.12, IDE: Spyder 5.1.5, Libraries: pytorch 1.10.0, torchvision 0.11.1, sklearn, pandas 1.3.4, numpy 1.21.2

6 Evaluation

6.1 Affective State Classification

In this experiment, the model was evaluated on the basis of standard classification measures such as accuracy, precision and recall along with a confusion matrix. The model was trained for 10 epochs and 4th epoch, as seen in Figure 5, was selected on the basis of highest accuracy (80.9%) and minimum loss (67.5%). The unweighted average recall score for the model is 80%. The overall accuracy of the model is highest when compared to the other four CNN and SVM based models, in Rao and Rao (2020), with the highest accuracy of 53.4%. The model has a high precision and recall for 1-Engagement class, as seen in Table 5, but fails to classify even a single sample of class 2-Confusion and 3-Frustration, as is seen in the confusion matrix in Figure 6.

Table 5: Affective State Classification Model performance metrics

Class	Precision	Recall	F-beta
0 - Boredom	64.7%	6.9%	2.5%
1 - Engagement	81.2%	99.2%	89.2%
2 - Confusion	0%	0%	0%
3 - Frustration	0%	0%	0%

Table 6: Affective State Classification Model comparison

Reference	Model	Accuracy
Rao and Rao (2020)	Handcrafted features with SVM	33.2%
Rao and Rao (2020)	CNN (softmax)	52.6%
Rao and Rao (2020)	CNN (upto FC) with SVM	52.6%
Rao and Rao (2020)	CNN (upto FC) + Handcraft_features with SVM	53.4%
This research	ResNet + TCN	**80.9%**

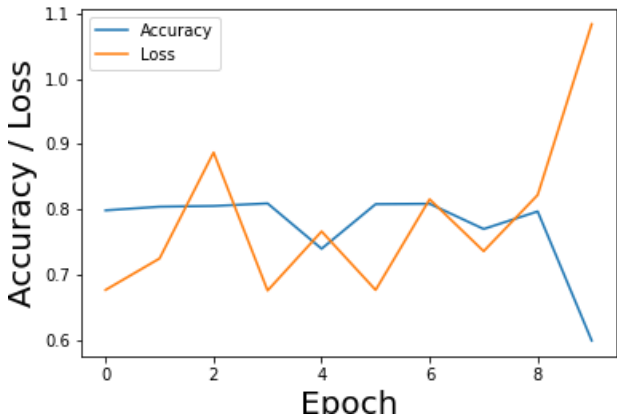


Figure 5: Model Training Graph for Affective State Classification Model

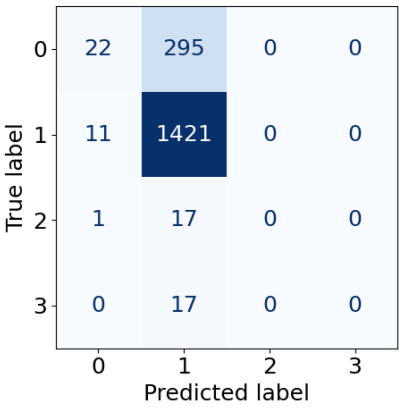


Figure 6: Confusion Matrix for Affective State Classification Model

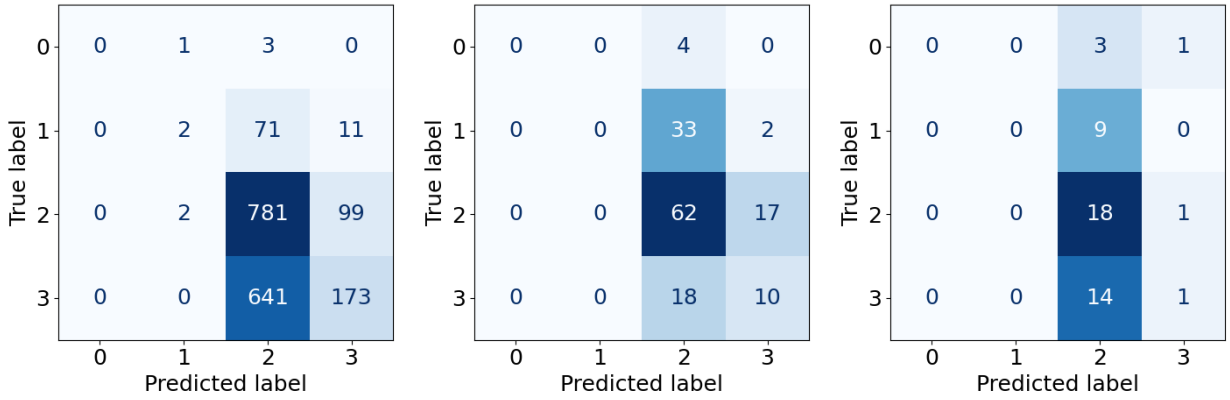
6.2 Engagement Level Classification

In this experiment, the model was trained on DAiSEE training split and evaluated on DAiSEE test split as well as Engagement in the Wild, both, training and test splits. This model was also trained for 10 epochs and the best model was selected on the basis of accuracy and unweighted average recall score for all three evaluations of this experiment. The evaluation of this model is done on the basis of multiple measures such as accuracy, unweighted average recall, number of training samples used, total number of frames used in training and the number of iterations the model was trained for. These metrics are important for a fair comparison with other similar studies conducted in the past such as Dresvyanskiy et al. (2021) and Abedi and Khan (2021b). Factors like, the amount of training data, the time taken to complete the training (which depends on the number of iterations) and the infrastructure used for training, play a critical role in determining the feasibility for further research and also indicate whether a model with similar performance can be obtained with much less infrastructure resources and time.

As seen in Table 7, the model achieved 53.6% accuracy with an unweighted average recall score of 28% on the DAiSEE test split. A similar score is also obtained on EmotiW2020's Engagement in the Wild data set train split with an accuracy of 49.3% and unweighted average recall of 28.6% while on the test split of the same data set the scores are 44.4% and 25.3% respectively. If we look at the model classification for each class in Figure 7, there is a common pattern in all three confusion matrices, the model is classifying most of the level 1 and 3 samples as level 2 and has not classified even a single sample as level 0.

Table 7: Engagement Level Classification Model performance comparison including performance on Emotiw Data set and comparison of amount of training data used and training iterations

Reference	**Data Set**	**Model**	**Accuracy**	**UAR**	**Training Samples**	**Frames per sample**	**Frames used in Training**	**Epochs**
Dresvyanskiy et al. (2021)	DAiSEE	RNN based ER system	29.0%	44.3%	5358	50	267,900	50
Abedi and Khan (2021b)	DAiSEE	ResNet + TCN	63.9%	33.6%	7087	50	354,350	100
This Research	DAiSEE	ResNet + TCN	53.6%	28.0%	5358	10	**53,580**	**10**
This Research	Emotiw (Train set)	ResNet + TCN	49.3%	28.6%	5358	10	**53,580**	**10**
This Research	Emotiw (Test set)	ResNet + TCN	44.4%	25.3%	5358	10	**53,580**	**10**
Tran et al. (2015)	DAiSEE	Fine-Tuned C3D	57.8%	30.7%	-	-	-	-
Abedi and Khan (2021b)	DAiSEE	C3D + TCN	59.9%	31.5%	-	-	-	-
Abedi and Khan (2021b)	DAiSEE	ResNet + TCN + weighted sampling and loss	53.7%	37.1%	-	-	-	-
Liao et al. (2021)	DAiSEE	DFSTN	58.8%	35.5%	-	-	-	-



(a) DAiSEE **(b)** EmotiW2020(Train) **(c)** EmotiW2020(Test)

Figure 7: Confusion Matrices for Engagement Level Classification of DAiSEE (a), Engagement Prediction in the Wild Train set (b) and Test set (c)

6.3 Other Affective States' Level Classification

There were three separate experiments for the classification of the level of each of the three remaining affective states, namely, Boredom, Confusion and Frustration. All three models were trained for 10 epochs and the best model for each of the three experiments was selected on the basis of highest accuracy.

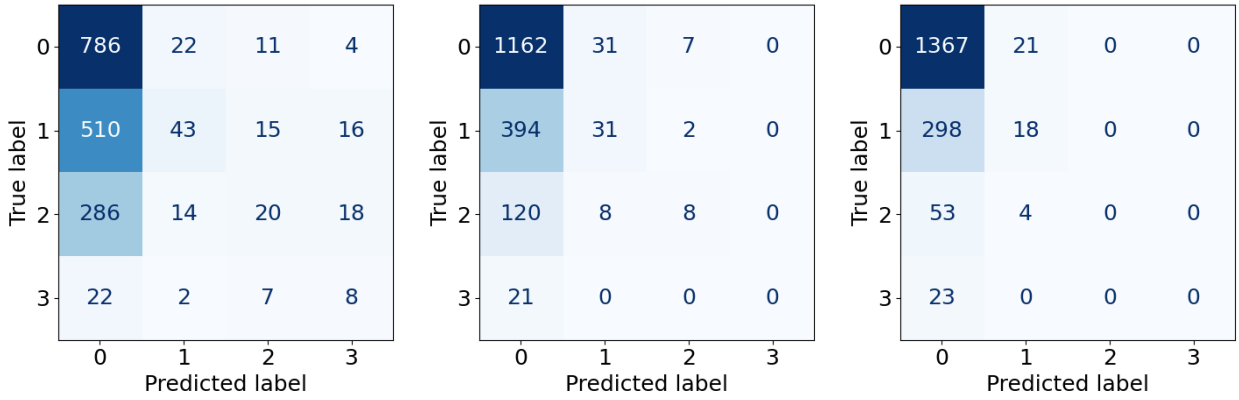
In the experiment for classification of level of Boredom, the model achieves an accuracy of 48.0%. As seen in Table 8, this accuracy is much higher than the accuracy achieved by, both frame based models, Inception Net and Emotion Net in Gupta et al. (2016). The accuracy achieved is 36.5% and 35.9% for Inception Net and Emotion Net, respectively. The confusion matrix for this model in Figure 8a shows that although the model has classified samples in all the four levels, majority of the samples of levels 1, 2 and 3 are still classified as level 0.

The model for classification of level of Confusion affective state has achieved an accuracy of 67.3% which is much higher than the 57.5% accurate Emotion Net model in Gupta et al. (2016), as seen in Table 8, but at the same time falls a bit short of Inception Net model in the same study whose accuracy is 70.3%. The model has classified samples in levels 0, 1 and 2 but all samples of level 3 have been classified as level 0 which can be seen in the confusion matrix in Figure 8b.

Similar to the performance of the above model, the model for classification of level of Frustration also performs better than the Emotion Net model but still fails to outperform the Inception Net model in Gupta et al. (2016). The model achieved an accuracy of 77.6% in classifying the level of Frustration which is better than the Emotion Net model with an accuracy of 73.1% and is very close to the Inception Net model whose accuracy is 78.3%. This model has also classified majority of the samples as level 0 and has not classified any sample as level 2 or level 3 as seen in the confusion matrix of the model in Figure 8c

Table 8: Boredom, Confusion and Frustration Level Classification Model performance comparison

Reference	Model	Level of Affective State		
		Boredom	**Confusion**	**Frustration**
Gupta et al. (2016)	Inception Net	36.5%	**70.3%**	**78.3%**
Gupta et al. (2016)	Emotion Net	35.9%	57.5%	73.1%
This research	ResNet + TCN	**48.0%**	67.3%	77.6%



(a) Boredom **(b)** Confusion **(c)** Frustration

Figure 8: Confusion Matrices for Affective State's Level Classification of Boredom (a), Confusion (b) and Frustration (c) Affective states

6.4 Discussion

Even though the model to classify affective states resulted in a much better accuracy in comparison to existing work of Rao and Rao (2020), there were no classifications in two classes, confusion (2) and frustration (3), out of the four available classes. The reason for the bias in classifications can be atrributed to the heavily imbalanced class distribution, also pointed out by Dresvyanskiy et al. (2021), as seen in Table 3 where two out of four classes have only 2% of samples.

The results of Engagement Level classification model were satisfactory given that in this research only 15-20% data is used for training and the model training iterations also were only 10-20%, when compared to the works of Dresvyanskiy et al. (2021) and Abedi and Khan (2021b), and still the model was able to achieve results close enough to them. Again, if we refer Table 2, it is clear that Level 0 and 1 for Engagement affective state have less than 5% samples because of which there were negligible classifications in these two classes for all 3 evaluated data sets, namely, DAiSEE(Test split), EmotiW2020(Train split) and Emotiw2020(Test split), as seen in Figure 7a, 7b and 7c, respectively.

Regarding the models for classification of level of boredom, confusion and frustration affective states, the results are impressive when compared to existing models of Gupta et al. (2016). The performance of model trained for classification of level of boredom is more than 30% better when compared to Inception Net and Emotion Net. Results for classification of level of confusion are almost 20% better than the Emotion Net but fall a bit short when compared to Inception Net. Similarly the frustration level classification model also has a higher accuracy when compared to Emotion Net but just falls a hair short of Inception Net.

The results of the experiment conducted in this research clearly prove that a pre-trained neural network can classify the level of affective state on different data sets with a consistent accuracy. Two data sets, DAiSEE and EmotiW2020, are identified for engagement affective state level classification and model is trained on the DAiSEE data set. The performance of the model trained on DAiSEE data set is evaluated on both, DAiSEE and EmotiW2020, data sets and is consistent for both data sets, as seen in Table 7, where accuracy is 53.6% and 49.3% and unweighted average recall score is 28.0% and 28.6% for DAiSEE and EmotiW2020 data set, respectively.

7 Conclusion and Future Work

This research presents an approach to classify the affective states and their levels in a learning environment with the help of hybrid ResNet + TCN bases neural network model. The performance of each model is compared to existing works in the domain using the same data set but different approaches. The engagement level model is also tested on another dataset with similar results as the original data set which proves the robust training of the model even with a fraction of training data and iterations. The most challenging part is correctly classifying the samples of classes with very low number of training samples.

In future there is significant scope to extend and enhance this research for even better results. Although this research has achieved the highest accuracy of 80.9% in classification of the four affective states (0 – Boredom, 1 – Engagement, 2 – Confusion, 3 - Frustration) in the DAiSEE data set compared to previous works, there is not even a single classification in class 2 – Confusion and 3 – Frustration. This can be attributed to the huge imbalance in the data set where these two classes without a single classification have only 1% training samples each. Considering the option to enhance the data set by adding more samples of confusion and frustration classes is not possible, this experiment can be made more meaningful by classifying only 0 – Boredom and 1 – Engagement classes. However, the sample distribution of even these two classes is highly imbalanced where the approximate ratio of samples of Boredom to Engagement class is 1:4. This can be handled by applying the random under-sampling technique on the majority class 1 – Engagement, until both the classes are balanced out. This should help reduce the misclassification of 0 – Boredom class as 1 – Engagement class and significantly improve the recall score of 0 – Boredom class which is currently only 6.9%.

Another enhancement which should result in improved results is the fragmentation of EmotiW2020 data set. The EmotiW2020 data set had almost equivalent duration of videos for training as the DAiSEE data set but the ratio of training samples 3:100 as Emotiw2020 data set had only 150 training samples and DAiSEE data set had more than 5000 training samples. This huge difference was because of the significant difference in the duration of each video clip of both data sets. The video clips in EmotiW2020 data set were 5 minutes long on an average whereas the ones in DAiSEE data set were 10 seconds long each. This prevented the research from training the model on EmotiW2020 data set and bench-marking it on DAiSEE data set and comparing the performance of both the models on their own data set as well as external data sets. Preprocessing the EmotiW2020 data set to divide each video clip into multiple fragments of 10 seconds each will increase the number of samples and enable the cross comparison of both models on both data sets.

References

Abedi, A. and Khan, S. (2021a). Affect-driven engagement measurement from videos, *arXiv preprint arXiv:2106.10882* .

Abedi, A. and Khan, S. S. (2021b). Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network, *arXiv preprint arXiv:2104.10122* .

Bai, S., Kolter, J. Z. and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, *Universal Language Model Fine-tuning for Text Classification* .

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T. and Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments, *International Journal of Human-Computer Studies* **68**(4): 223–241.

Dash, S., Dewan, M. A. A., Murshed, M., Lin, F., Abdullah-Al-Wadud, M. and Das, A. (2019). A two-stage algorithm for engagement detection in online learning, *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*, IEEE, pp. 1–4.

Dewan, M. A. A., Murshed, M. and Lin, F. (2019). Engagement detection in online learning: a review, *Smart Learning Environments* **6**(1): 1–20.

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology., *Journal of Educational Psychology* **105**(4): 1082.

Dresvyanskiy, D., Minker, W. and Karpov, A. (2021). Deep learning based engagement recognition in highly imbalanced data, *International Conference on Speech and Computer*, Springer, pp. 166–178.

El Kerdawy, M., El Halaby, M., Hassan, A., Maher, M., Fayed, H., Shawky, D. and Badawi, A. (2020). The automatic detection of cognition using eeg and facial expressions, *Sensors* **20**(12): 3516.

Gupta, A., D'Cunha, A., Awasthi, K. and Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild, *arXiv preprint arXiv:1609.01885* .

Kamath, A., Biswas, A. and Balasubramanian, V. (2016). A crowdsourced approach to student engagement recognition in e-learning environments, *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9.

Kaur, A., Mustafa, A., Mehta, L. and Dhall, A. (2018). Prediction and localization of student engagement in the wild, *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8.

Leong, F. H. (2020). Deep learning of facial embeddings and facial landmark points for the detection of academic emotions, *Proceedings of the 5th International Conference on Information and Education Innovations*, pp. 111–116.

Liao, J., Liang, Y. and Pan, J. (2021). Deep facial spatiotemporal network for engagement prediction in online learning, *Applied Intelligence* pp. 1–13.

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

Nambiar, D. (2020). The impact of online learning during covid-19: students' and teachers' perspective, *The International Journal of Indian Psychology* **8**(2): 783–793.

Nezami, O. M., Dras, M., Hamey, L., Richards, D., Wan, S. and Paris, C. (2020). Automatic recognition of student engagement using deep learning and facial expression, *19th Joint European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML PKDD 2019*, Springer, Springer Nature, pp. 273–289.

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, *Advances in neural information processing systems* **32**: 8026–8037.

Rao, K. P. and Rao, M. C. S. (2020). Recognition of learners' cognitive states using facial expressions in e-learning environments, *Journal of University of Shanghai for Science and technology* **22**(12): 93–103.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015). Going deeper with convolutions, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks, *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497.

Werlang, P. S. and Jaques, P. A. (2021). Student engagement recognition from videos: A comparison between deep learning neural network architectures, *Bulletin of the Technical Committee on Learning Technology (ISSN: 2306-0212)* **21**(3): 7–12.