# Affective State-Based Framework for e-Learning Systems

Juan Antonio RODRÍGUEZ [a], Joaquim COMAS [a] and Xavier BINEFA [a]

[a] *Department of Information and Communication Technologies,*
*Pompeu Fabra University, Barcelona, Spain*

**Abstract.** Virtual learning and education have become crucial during the COVID-19 pandemic, which has forced a rethink by teachers and educators into designing online content and the indirect interaction with students. In an face-to-face class, some visual cues help the teacher recognize the engagement level of students, while the main weakness of the online approach is the lack of feedback that the teacher has about the learning process of the students. In this paper, we introduce a novel framework able to track the learning states, or LS, of the students while they are watching a piece of knowledge-based content. Specifically, we extract four learning states: Interested, Bored, Confused or Distracted. Finally, to demonstrate the system's capability, we collected a reduced database to analyze the affective state of the subjects. From these preliminary results, we observe abrupt changes in the LS of the audience when there are abrupt changes in the narrative of the video, indicating that well-structured and bounded information is strongly related with the learning behaviour of the students.

**Keywords.** Learning states, e-Learning, Deep Learning, Machine Learning, Facial Expression Analysis

## 1. Introduction

Online education has become very popular due to lockdown measures for the COVID-19 pandemic, and the increase of online resources in the form of MOOCS. MOOCS are very specialised courses based on any subject, that allow students to work asynchronously from any location using smart devices. However, this paradigm provokes a huge disconnection between the learner and the teacher, as the former one loses the ability of adapting the session based on the emotional reception of the students. Future E-Learning systems must integrate mechanisms to predict learning states of the students in order to drive the learning session and enhance the educational experience.

Affective Computing is the field that studies how machines will be able to assist humans or make decisions based on emotional information. The term was first introduced by Picard [1], where she also stated how E-Learning systems may integrate affective capabilities to keep the students engaged. This problem can be tackled using many sources of data in a multi-modal manner, using physiological signals like heart rate (HR), skin conductance response (SCR), blood volume pulse (BVP) and electroencephalogram (EEG) as well as facial, body pose images or audio signals. Any biological signal can correlate with the affective state of a person, however the non-invasive ones must be our

choice in order to build easy-to-use systems and put users' attention in the learning part. Following the last argument, most of the systems and models for E-Learning and Emotion Prediction use Computer Vision techniques for Facial Expression Analysis and the extraction of other visual cues. Nevertheless multi-modal fused models show superiority in accuracy [2] over just Facial Expression Analysis, at the cost of having more complex set-ups.

There has been a lot of efforts to find the best prediction/target models in order to assess emotional states. The Valance-Arousal two dimensional space introduced by Russell [3], is the choice of many works, as it encodes emotional states in two dimensions and gives a mapping to discrete classes. Other approaches use a classification model with classes of emotions focused in the task of choice. For E-Learning, classes may come from the set of Engagement, Interest, Confusion, Boredom, Distraction or Frustration. Some works may also use Emotion Levels bounded from 0 to 1 for a given class, and solve the problem by means of regression. With the explosion of Deep Learning and Neural Networks, many prediction tasks have been solved with huge accuracy and robustness. This is due to the effort on building larger models in terms of parameters, better training mechanisms and also the increment of training data available. New technologies allow for the integration of such models into smart devices in a very optimized manner, contributing for accuracy and speed. This work in progress focuses on studying many deep learning models for E-Learning Emotion Prediction and effectively use this functionality to enhance the E-Learning experience. Experiments in this paper will use a baseline model based on Convolutional Neural Networks (CNN), to focus on the system's emotion mechanics and analysis.

In this work, we focus on a specific type of resource in a learning session, which is the video. Almost every online course is supported by video based content to explain important information from a topic. These videos must be precisely designed, structured and narrated in order to achieve the goal of maintaining the student interested and transmitting the desired knowledge. We propose a system that predicts learning states from facial expressions, and links this result with a time instant of the video. With this synchronisation we can assess the evolution of emotions through the narrative of the video and also analyze if the narrative structure is properly designed.

The rest of the paper is structured as follows: Section 2 overviews literature and related work on the problem of affective state-based E-Learning. In Section 3 we describe the methodology of building the video-based E-Learning system, with details on the E-learning workflow, the prediction model and the datasets for training. In Section 4 we report preliminary experiments and results. In section 5 we reflect on ethics, fairness and privacy of works of this kind, when sensitive data is gathered from users and biased prediction models will most likely appear. Finally, Section 6 explains conclusions and future work.

## 2. Related Work

The task of Affect Recognition in E-Learning scenarios has shown challenges in both the methodology used for predicting the Learning States from raw data, and the design of plausible system models capable of using this predictions in benefit of the learning experience. The following is an overview of studies that focus on those two problems.

One of the earliest uses of physiological data for automatic emotion estimation on e-learning environments has been the work of Shen et al. [4], using signals such as HR, SCR, BVP and EEG. The authors proposed Support-Vector Machine (SVM) as prediction model in the Valence-Arousal space. In this early work, it is shown how emotion aware systems increment the interaction and engagement of the subjects.

Porta et al. [5] presented a preliminary study of a system able to adapt to the users learning states through eye tracking data. They find that variation in pupil size is highly correlated with users actions and mental effort.

Also, Ashwin et al. [6] conducted experiments in a face-to-face system focused on efficiently detect facial images from multiple users, and predict seven emotions using SVM with some well-known datasets.

Alyuz et al. [7] relied on an appearance data gathered from the camera and context information from the platform, and proposed a method with a calibration phase in order to personalize the model to the subject. Their architecture predicted Satisfaction, Boredom and Confusion using a self-collected dataset and Random Forest method.

The work presented by Chen et al. [2] introduce a hybrid model combining several modalities based on facial and physiological cues, to predict valence, arousal and attention. The authors propose a model for assisting the user based on those three indicators, using SVM and multi-modal fusion.

Luo et al. [8] discuss a new model for affective learning based in three dimensions: Attention, Emotion and Thinking. In this work, experiments where conducted using only facial images in a face-to-face e-Learning session, capturing facial and emotional data from many subjects concurrently. They use classical methods and fusion to obtain a final indicator of interest.

In the recent literature, other data-driven methods have introduced more sophisticated models and provided dedicated datasets to assess the emotional state in e-learning systems. First example is the work of Kaur et al. [9] which introduces a new dataset with collected videos and eye gaze data from students in a learning set-up, proposed for the task of Engagement Level Prediction. The authors provide many baseline models and results using modern deep learning networks.

Following the line of works in the deep learning paradigm, the use of CNN for image feature extraction has become the standard choice. Emotion Recognition studies conducted in [10], [11], [12], [13], [14] achieve State of the art accuracy on very popular datasets using CNN using both classification and regression targets.

Finally, Mukhopadhyay et al. [15] proposed a CNN-based method to assess the state of mind of the learners during an online learning session extracting Learning States based in the combinations of the eight basic emotions, namely confusion, dissatisfaction, satisfaction, and frustration.

A limitation for the CNN baseline architecture arises when images are from a video source, and vary in the time dimension. In order to reach high accuracy for spatio-temporal data patterns one must set-up a training and inference mechanic to treat data in a sequence-like fashion. In that sense, there is a lack of studies that focus on this type of data structures for Affective E-Learning systems.

The main contribution of this paper is the design and preliminary test of a novel framework for e-learning, that allows educators to obtain feedback about the LS of the students during the visualization of a video. The main aspects are the following:

1. Propose a web application-based E-Learning system focused on the visualization of video resources. The presented framework links video timestamps with LS predictions to asses the quality of the video design and the affective evolution of the student.
2. Show performance and accuracy results of a baseline data-driven model optimized for performance in edge devices. The model uses CNN feature extraction and classification.

## 3. Methodology

### 3.1. Experimental set-up and recording

The designed system can be accessed through a web application with a computer or mobile device, and the student must agree and give consent of data treatment, see part 4 Ethics, fairness and privacy. At the start of the session, the user has many video options to visualize, which are knowledge-based to introduce the student into a learning challenge. Afterwards, the system warms up, loading both face detection and learning state prediction models, as well as camera recording.

When the user feels ready, starts the visualization, and the system begins to detect faces at a chosen frame rate, which are then cropped and resized to be propagated through the neural network. The raw predictions, facial images and time records are stored in the database for further analysis.

One of the key functionalities of the system is *Ground Truth Feedback* or GTF, a mechanism that lets the teacher define questions in relevant parts of the narrative structure of the video that will be prompted to the student. This information is stored to ease the testing and validation of both the training and accuracy of the deep learning models, assess the learning process of the students, and the quality of the video.

When the visualization is finished, the recorded video of the subject is stored in the database for training purposes. The full video recording will allow further researchers to explore new ways of processing the data, having all frames and more data modalities such as head or body pose.

Algorithm 1 describes the process, and it consists of a loop over the whole video sequence, where the operations of face extraction, LS prediction and storage are performed subsequently. At every iteration, the system checks if it needs to ask a feedback question based on a list of time samples generated by the teacher. As the questions may be different, the current video time is passed as a parameter. Parallel to this process, the system is recording the camera video stream, which will be encoded in *H.264* and stored in the database as a *.webm* file. These videos will contribute to the *Learning Students Dataset*, alongside the automatic predictions and the GTF.

After the learning session, the subject moves to the *Course Analyzer*, Figure 1, an interactive dashboard that allows the learner to analyze the session in terms of the LS evoked by the video. The *Course Analyzer* generates charts that show the distribution of the learning state predictions over time, which gives an indicator of the overall reactions to the content.

We also generate a detailed table with the stored LS and facial expressions, which by clicking over, the system plays the video in that instant of time as well as shows facial

expressions of other participants at that time. This gives the teacher or content creator a tool to analyze how the narrative is being received.

```
Data: YouTube video, webcam stream
Result: LS, facial images, GTF
1  while video playing do
2      if feedbackTime(videoTime) then
3          GTF = askFedback(videoTime)
4          store(GTF)
5      end
6      faceImage = camera.detectFace()
7      if faceImage then
8          LS = nn.prediction(faceImage)
9          storeData(LS, faceImage, videoTime, timestamp)
10     end
11 end
12 storeFullVideo()
```

**Algorithm 1:** Workflow of the learning session.

Finally, we define the *Learning Timeline*, a chart that aggregates the LS over windows of time intervals. Over these intervals, a Majority Voting Rule (MVR) is computed, which returns just one LS with information also from the time domain. The *Learning Timeline* also shows the answers given by the student in the GTF. This is computed at three levels, student level, for one learning session, video level, for many visualizations of a user, and group level, for many students in the same video.
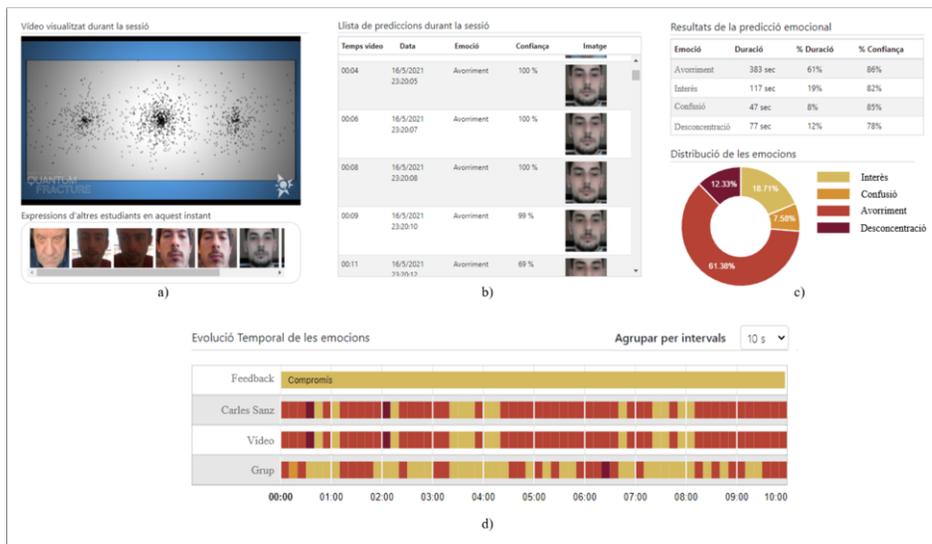


**Figure 1.** Course Analyzer. a) Video visualized by the student and facial expressions of other students in that instant. b) Learning state predictions in detail. c) Distribution of the learning states. d) Learning Timeline at three levels and feedback answers of the student.

## 3.2. Proposed method

We use data driven-based inference in order to make predictions of the LS from facial images. We use the *MobileNetV2* architecture [16] as a backbone network that is specially designed for working on edge devices efficiently. This model is based on Convolutional Neural Networks and Inverted Residual structures, which allow the network to learn powerful low-dimensional representations at a small computational cost.

This framework acts as a baseline for predicting the LS, which can effectively model the space domain of the data, i.e. it processes the images frame by frame, not taking into account the sequential nature of the data. Using the Majority Voting Rule over a time interval, the framework can give a final prediction using the time domain at low cost. However, this baseline architecture is not able to use the time domain in the training stage. For that purpose, we need more sophisticated architectures that do the prediction task using a sequence of frames as input. As a work-in-progress, we are analysing the impact of Temporal Convolutional Networks (TCN) [17]. These new models will come at a more computational cost but may exhibit more accurate and robust performance.

## 3.3. Data and Training

The training data consists of facial images paired with the annotated learning state class that the subject is evoking. The source of this data is a combination of public datasets, namely Affectnet, CK+ and DAiSEE. Affectnet dataset [18] consists of more than 1M images of facial expressions annotated with eight basic emotions: Disgust, Happy, Surprise, Fear, Angry, Contempt, Sadness and Neutral. CK+ dataset [19] is formed by 593 video sequences showing facial expressions of people, also annotated with the basis of eight emotions. CK+ gives significant amounts of data of not-in-the-wild examples, which means that the conditions of the images and facial expressions are very staged. These datasets do not contain labels for our task, but we performing a training stage for classification with the eight classes, for further transfer of learning.

Furthermore, we apply transfer learning to the model with the four LS. This process is done by freezing the weights of the first convolutional layers, and training the model with the new labels until convergence. The second training was done using the DAiSEE Dataset [20], which contains over 9000 video sequences of students in an e-Learning session. They are labelled using four labels that represent the levels of the e-Learning emotions: Boredom, Engagement, Confusion and Frustration. We performed a supervised re-annotation of the videos in order to get our four LS. This process consisted in labelling facial frames with one of the four emotions with a level 2, and supervising the fidelity of the set labels. Therefore, the training dataset has been customized for our specific task.

In Table 1, we show a summary of the datasets. As an indicator for the balance level of a given dataset, let us define Representation Balance. We propose the use of Shannon's Entropy for a Random Variable, considering the set of classes $\mathscr{U}$ and the probability distribution over the classes $P(u)$,

$$RB = -\frac{1}{\log_2(m)} \sum_{u \in \mathscr{U}} P(u) \log_2(P(u)), \tag{1}$$

where $m$ is the number of classes in the dataset. The second factor of the equation is Shannon's Entropy, and $-\frac{1}{\log_2(m)}$ is a normalization factor. When RB is closer to 1 it

means the the classes are more equiprobable. It is clear that *Affectnet* and *DAiSEE* sets are very unbalanced, a fact that will impact training and accuracy.

As two of our datasets contain videos, we must perform a preprocessing to extract the facial information from each frame and crop it to a size of 224x224. We use exhaustive frame extraction, which consists on using all frames in the video. This is done to obtain full detail over the sequences and configure the dataloader to get the samples configurable frame rate as a hyperparameter. By fine-tuning the frame rate, we expect to get a result indicating the optimum value, a trade-off between speed and accuracy.

| Dataset | # Samples | # Subjects | Labels | Representation balance |
|---------|-----------|------------|--------|------------------------|
| Affectnet | 1M images | 450000 | 8 basic emotions | 0.73 |
| CK+ | 593 videos | 123 | 7 basic emotions | 0.91 |
| DAiSEE | 9068 videos | 112 | 4 learning states | 0.61 |

**Table 1.** Description of the datasets used for train and test

The baseline model was trained using a single *NVIDIA Titan X* for 40 epochs which took around 40 hours. The chosen batch size was 64 given by the limitations on GPU memory. The learning rate was empirically set at 0.0001 and we used Adam optimizer, also based on empirical results. The loss function used for training is Categorical Cross-Entropy, the standard for a classification problem.

A final note on training and data is about the problem of class balance. As shown with the RB indicator, table 1, some classes are over-represented as well as others that have fewer samples. This will generate biases in the predictions, where the less representative classes will not be well learned. We tackled this problem by using balance weights, which penalize highly imbalanced classes.

The trained model was developed using *Keras* [21] and *Tensorflow* [22] frameworks, and we employ *Tensorflow JS* in order to optimize the model for edge devices. The process of optimization removes all the layers and operations that are only used for training, maintaining just the operations needed for inference. Furthermore, the model is stored in a JSON file containing the inference operation and the trained parameters of the model.

### 3.4. Students Learning Reaction Dataset

This work in progress will follow up with the confection of the *Students Learning Reactions (SLR) Dataset*. The dataset will facilitate video recordings of students in a learning session and the video that the user was watching. The videos will be annotated using the GTF given by the user in significant instants, and also automatic annotations using the learning states predicted by the Deep Learning model. This dataset will allow researchers to attack many tasks in Computer Vision, such as Emotions Evoked in videos or Engagement Prediction, under an End User License Agreement and the approval of the Ethics committee at UPF (CIREP). For more information refer to Section 5.

## 4. Preliminary results

This section presents the trained baseline model results using train and validation sets depicted in table 2. Moreover, we describe experimental results obtained from recording a small set of participants.

## 4.1. Training Results

The precision score (in per cent unit) and categorical Cross-Entropy are used to judge the quality of each estimated affect labels with train and validation splits (80% - 20%).

| Learning State | Val Loss | Val Acc. |
|:---:|:---:|:---:|
| **Overall** | **1.2257** | **0.8068** |
| Interested | 0.6444 | 0.8418 |
| Confused | 1.2431 | 0.7534 |
| Bored | 0.6754 | 0.8754 |
| Distracted | 1.1654 | 0.6754 |

**Table 2.** Results of validation for each LS

Results for the baseline model using *MobileNetV2* are depicted in table 2. Overall validation metrics exhibit acceptable accuracy even with just involving space information. Looking at the breakdown over LS we can see drifts in the different classes. This due to the balance of the data. Results also prove the need to involve time-domain information to make the system more robust and precise.

## 4.2. Experimental test results

In this experiment, we present the system to many participants and let them perform a learning session. The goal is to test the functionalities of the system and also validate the performance of the prediction model, comparing the predictions with the given GTF. Finally, assess the learning process at the three levels, namely student, video and group.

Table 3 presents some metrics on inference speed and video quality for different device configuration. The difference in inference speed is big when comparing high-range with low-range computers, however the values are acceptable in terms of user experience. In some cases when the computer is being stressed, video quality can get penalized due to the processing of the images.

| Configuration | Hardware | Inference speed | Video Quality |
|:---:|:---:|:---:|:---:|
| Mobile device | Android, 3GB RAM | 297ms | Subtle delays on face detection |
| Basic laptop | Intel i5, 16GB RAM | 280ms | Subtle delays on face detection |
| Mid range computer | Intel i7, 16GB RAM | 162ms | No impact |
| High range computer | Intel i9, 32GB RAM | 73ms | No impact |

**Table 3.** Performance comparison for different hardware configurations.

From the sessions performed with 20 students, we compute an MVR over intervals of 1 minute and compare the predictions with the GTF given in important parts of the video. In this case, we ask to answer which learning state are they experimenting with the video. The final precision score is 57%, which is lower than the predictions from the validation set. Empirically, we see that works very good on some participants and poorly on some others. This is due to video quality, illumination conditions and facial expressions not seen in training.

One remarkable observation is that the MVR at group level, averaging the LS overall students, denotes where are the critical instants of the narrative structure. When there are abrupt changes in the narrative, visually or conceptually, the learning states also change.

This work does not provide a comparison with other methods, as it was trained in a transfer-learning manner, using a customized version of DAiSEE dataset. Also, we cannot compare results on speed, as we have no reported values of other alternatives.

## 5. Ethics, fairness and privacy

In this work, we are very concerned with the behaviour of Artificial Intelligence algorithms regarding fairness and the occurrence of unwanted biases in the predictions. Labels that are over/under-represented in the training data yield models that disadvantage minority groups and vice versa. We attack this problem by balancing training and validation data over classes. We also adopt the approach of accounting predictions that have confidence above 70%, and ignoring the predictions below that threshold.

All works that involve user data acquiring and processing must be assessed and certified by an Ethics Committee, even more if the data is bio-metric and invasive. In that sense, this work is being evaluated by CIREP organism at UPF.

The requirement for being able to participate in the experiments is the given consent for data treatment that includes video recording and processing of facial information, as well as the publication of the resulting dataset for research purposes. Students that contribute to the *SLR Dataset* will sign the policy that will give privacy rights assured in the Spanish RGPD, and research groups that may use this dataset will get access by means of signing an EULA.

It is notorious that using this type of data (Facial and Emotional) can be intrusive, so we must assure that Face Expression-based methods only contribute to the improvement of virtual experiences, such as educational, medical or security.

## 6. Conclusions

In this paper we proposed a framework to improve the E-learning experience, with mechanics for predicting learning states and getting student feedback during video visualization. The method is based on facial expression analysis and CNN. The system is able to link video time with learning state, and in the analysis the teacher can follow the evolution at three levels of detail, student, video and group. The proposed baseline architecture presents acceptable performance with small computational resources. However, future work must be focused on studying more sophisticated models, that involve the temporal dimension of the data. Through experiments, we foresee that the narrative of the video has a direct impact on the learning states of the students. Finally, the system is ready to collect data in order to create the Students Learning Reactions Dataset, to find better solutions using Computer Vision.

## References

[1] R. W. Picard. *Affective computing*. MIT press, 2000.

[2] J. Chen, N. Luo, Y. Liu, L. Liu, K. Zhang, and J. Kolodziej. A hybrid intelligence-aided approach to affect-sensitive e-learning. *Computing*, 98(1-2):215–233, 2016.

[3] J. A. Russell. A circumplex model of affect., 1980.

[4] L. Shen, M. Wang, and R. Shen. Affective e-learning: Using emotional data to improve learning in pervasive learning environment. *Journal of Educational Technology & Society*, 12(2):176–189, 2009.

[5] M. Porta, S. Ricotti, and C. J. Perez. Emotional e-learning through eye tracking. In *Proceedings of the 2012 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–6. IEEE, 2012.

[6] TS Ashwin, J. Jose, G. Raghu, and G. R. M. Reddy. An e-learning system with multifacial emotion recognition using supervised machine learning. In *2015 IEEE seventh international conference on technology for education (T4E)*, pages 23–26. IEEE, 2015.

[7] N. Alyuz, E. Okur, E. Oktay, U. Genc, S. Aslan, S. E. Mete, Eb Arnrich, and A. A. Esme. Semi-supervised model personalization for improved detection of learner's emotional engagement. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 100–107, 2016.

[8] Z. Luo, C. Jingying, W. Guangshuai, and L. Mengyi. A three-dimensional model of student interest during learning using multimodal fusion with natural sensing technology. *Interactive Learning Environments*, 0(0):1–14, 2020.

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

[10] O. El Hammoumi, F. Benmarrakchi, N. Ouherrou, J. El Kafi, and A. El Hore. Emotion recognition in e-learning systems. In *2018 6th International Conference on Multimedia Computing and Systems (ICMCS)*, pages 1–6. IEEE, 2018.

[11] S. K. Gupta, TS Ashwin, and R. M. R. Guddeti. Students affective content analysis in smart classroom environment using deep learning techniques. *Multimedia Tools and Applications*, 78(18):25321–25348, 2019.

[12] M. Megahed and A. Mohammed. Modeling adaptive E-Learning environment using facial expressions and fuzzy logic. *Expert Systems with Applications*, 157:113460, 2020.

[13] K. P. Rao and M V P C. K. Rao. Recognition of Learners ' Cognitive States using Facial Expressions in E-learning Environments. 22(12):93–103, 2020.

[14] A. Sun, Y. J. Li, Y. Min Huang, and Q. Li. Using facial expression to detect emotion in e-learning system: A deep learning method. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10676 LNCS:446–455, 2017.

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

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[17] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks for action segmentation and detection. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:1003–1012, 2017.

[18] P. Lucey, J. F Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

[19] A. Mollahosseini, B. Hasani, and M. H. Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *arXiv*, 2017.

[20] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian. DAiSEE: Towards User Engagement Recognition in the Wild. 14(8):1–12, 2016.

[21] F. Chollet et al. Keras. https://keras.io, 2015.

[22] M. Abadi, P. Barham, K. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.