Applied Intelligence (2022) 52:13803–13823
https://doi.org/10.1007/s10489-022-03200-4

Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement

Naval Kishore Mehta[1,2] · **Shyam Sunder Prasad**[1,2] · **Sumeet Saurav**[1,2] · **Ravi Saini**[1,2] · **Sanjay Singh**[1,2]

Accepted: 4 January 2022 / Published online: 18 March 2022

Abstract

Today, due to the widespread outbreak of the deadly coronavirus, popularly known as COVID-19, the traditional classroom education has been shifted to computer-based learning. Students of various cognitive and psychological abilities participate in the learning process. However, most students are hesitant to provide regular and honest feedback on the comprehensiveness of the course, making it difficult for the instructor to ensure that all students are grasping the information at the same rate. The students' understanding of the course and their emotional engagement, as indicated via facial expressions, are intertwined. This paper attempts to present a three-dimensional DenseNet self-attention neural network (DenseAttNet) used to identify and evaluate student participation in modern and traditional educational programs. With the Dataset for Affective States in E-Environments (DAiSEE), the proposed DenseAttNet model outperformed all other existing methods, achieving baseline accuracy of 63.59% for engagement classification and 54.27% for boredom classification, respectively. Besides, DenseAttNet trained on all four multi-labels, namely boredom, engagement, confusion, and frustration has registered an accuracy of 81.17%, 94.85%, 90.96%, and 95.85%, respectively. In addition, we performed a regression experiment on DAiSEE and obtained the lowest Mean Square Error (MSE) value of 0.0347. Finally, the proposed approach achieves a competitive MSE of 0.0877 when validated on the Emotion Recognition in the Wild Engagement Prediction (EmotiW-EP) dataset.

Keywords Engagement recognition · Online learning · Attention network · Spatio-temporal features

This article belongs to the Topical Collection: *Artificial Intelligence Applications for COVID-19, Detection, Control, Prediction, and Diagnosis*

✉ Shyam Sunder Prasad
shyam.ece56@gmail.com

Naval Kishore Mehta
naval.mehta95@gmail.com

Sumeet Saurav
sumeet@ceeri.res.in

Ravi Saini
ravi@ceeri.res.in

Sanjay Singh
sanjay@ceeri.res.in

[1] Academy of Scientific and Innovative Research(AcSIR), Ghaziabad, India

[2] CSIR-Central Electronics Engineering Research Institute(CSIR-CEERI), Pilani, India

1 Introduction

The global outbreak of COVID-19 has led to closure of educational institutes worldwide. According to a recent survey by UNESCO, due to the COVID-19 pandemic, over 190 countries' schools and 61 countries' higher education institutions have been forced to shut down [1]. The absence of in-person education in schools, colleges, and universities has adversely affected approximately 156 million students throughout the world.[1] According to experts, normal classroom instruction is unlikely to return anytime soon. In such circumstances, students are heavily dependent on online learning environment to complete their courses through virtual mode platforms like Zoom, WebEx, and Google meet [2]. Due to the availability of the internet worldwide, learning through virtual mode has become the new normal and has finally reached remote corners of the country. Students can learn and communicate with teachers, and their peers

[1] https://en.unesco.org/covid19/educationresponse, accessed on 22/06/2021

from far-off locations [3]. In terms of transportation, lodging, and the overall cost of institution-based learning, it is considered an alternative cost-effective solution for physical learning [4–6]. The main problem in synchronous learning environments, such as e-learning platforms, Massive Open Online Courses (MOOCs), traditional classrooms, seminars, and so on, is to maintain students' comfort with course content. Moreover, lack of interaction with instructors and the absence of traditional classroom socialization has led to the increased withdrawal rate among students' using online learning environments [4, 7]. Hence, detecting learners' involvement has become critical in virtual education to give individual pedagogical assistance. Analyzing the students' engagement can be one of the possible phenomenon that can help measure the above parameters effectively. Encouraging student engagement is beneficial not just in online learning but also in other learning environments such as conventional classrooms, problem-solving and creative educational games, and online tutoring systems [8, 9].

One can consider student engagement in any learning environment a blend of behavioral, cognitive, and emotional states. Behavioral engagement requires the active participation and involvement of the students in classroom activities. On the other hand, cognitive engagement incorporates students' zeal to master learning and self-dedication towards the learning process. Whereas emotional engagement is one of the affective states of the student, and it indicates the students' active participation in the class [10–12]. Students who are emotionally and cognitively engaged in learning tend to put more effort into their studies and are more persistent and efficient in meeting the courses' demands than students with relatively less emotional and cognitive engagement. It is, therefore, one of the fundamental determinants of the welfare and overall growth of students. Emotional engagement in the teacher-student and peer group relationships has an indirect effect on the perceived influence of students' cognitive engagement [10, 13–15]. Engagement intensity is strongly related to facial expressions, upper-body posture, and overall environmental factors, but facial expressions are the most natural nonverbal way of expressing engagement in the online learning environment [1, 8]. For the longest time, datasets largely covered the seven basic expressions of happy, neutral, sorrow, anger, surprise, disgust, and contempt [16–19]. The emphasis is currently on monitoring a students' emotional state in a learning environment to obtain concrete results of students' involvement. A recent study identified face expressions connected to self-reported and evaluated learning-centered emotional states such as boredom, engagement, confusion, and frustration [20–22]. The relationship between the strength and timing of facial emotions is important in a synchronous learning environment [23].

There are primarily two approaches to estimate the affective states, one based on machine learning and the other on deep learning. The machine learning-based method acquire facial features and estimate engagement levels by using hand-crafted patterns [23–26]. While deep learning techniques learn on-the-fly features from training data, allowing the algorithm to identify the fine-grained facial variations, and outperforms the traditional machine learning-based method in affective state prediction tasks [21, 27–31]. Moreover, the deep learning-based approaches are non-intrusive, and the resources used by these methods to capture and analyze facial expressions video data are low-cost, automated, and easy to implement [8]. Existing deep learning methods for emotion recognition are divided into two categories: static image-based approaches and video sequence-based methods [18]. It is more natural to categorize facial expressions from consecutive frames in a video, as the video sequence provides significantly more information for facial expression recognition (FER) than static facial images. The video sequence-based methods are subdivided into a combination of spatial and temporal CNN models, a three-dimensional convolutional neural network (3D CNN), and a hybrid of CNN and long-short term memory (LSTM) network [21, 27–29, 32–34]. In this paper, we have used an end-to-end 3D CNN-based method for video analysis and prediction.

Since the development of C3D neural networks [35], several deep 3D CNNs have been proposed for numerous computer vision applications [36]. For instance, Hara et al. [37] proposed several variants of the ResNet3D and ResNeXt3D CNN for human action recognition in video sequences. In another work, Ruiz et al. [38] proposed a 3D DenseNet for the classification of Alzheimer's disease. Most of the existing 3D CNNs are compute-intensive and thus cannot be utilized for real-time applications. On the other hand, our proposed 3D DenseNet, a modified 3D variant of the DenseNet-121 CNN [39] is tailored to attain the right balance between accuracy and computational cost. The designed model with 19.30 million parameters requires 76.70 megabytes (MB) memory footprint and thus can be deployed on a low-cost embedded platform for real-time detection of student engagement in a video stream. Besides, the self-attention blocks (spatial, temporal, and spatial-temporal) in the proposed model helps it to extract enhanced features for efficient detection of students' affective states in a video sequence. For the multi-class and multi-label classification of the affective states, we explored three loss functions, namely the Cross-Entropy (CE) [40], Class-Balanced Cross-Entropy (CB-CE), and Class-Balanced Focal loss (CB-FL) [41]. The MSE and Class-Balanced Mean Square Error (CB-MSE) were utilised in the regression experiment. We assessed the feasibility

and robustness of the proposed technique using publicly available DAiSEE [21], and EmotiW-EP [42, 43] datasets. The proposed systems' automatic and timely input to the instructor can improve the students' learning experience. The system can assess and identify a students' absent-mindedness during a lecture session, improve content delivery effectiveness, and boost productivity and learning gains. The major contributions of our work are summarized as follows:

- Design and implementation of a robust and efficient three-dimensional DenseNet self-attention neural network (DenseAttNet). The designed neural network is trained and tested for multi-class classification of affective states, with a baseline accuracy of 63.59% for engagement and 54.27% for boredom.
- Training and testing of the multi-label DenseAttNet on all four affective states: boredom, engagement, confusion, and frustration with accuracies of 81.17%, 94.85%, 90.96%, and 95.85%, respectively.
- The model has achieved a baseline MSE of 0.0347 on the DAiSEE engagement validation set and a competitive MSE of 0.0877 on the EmotiW-EP.
- The manuscript also includes a real-time implementation in a classroom setting, as well as Gradient-based Localization Class Activation Mapping (Grad-CAM)[44] is used for visual interpretation and understanding of model predictions.

The rest of the paper is organized as follows. Background and related work are discussed in Section 2. The details of our proposed method DenseAttNet and variation of loss function is described in Section 3. Section 4 discusses the qualitative and quantitative analysis and results of the engagement prediction as a classification and regression problem on DAiSEE and EmotiW-EP datasets, along with a thorough comparison with the benchmark results. Section 5 concludes the paper.

2 Related work

Momentary expressions that transmit emotions include muscle movements such as raising the brows, wrinkling the forehead, rolling the eyes, and curling the lip [45]. Students who are anxious may have a depressed brow, a drawn-together brow, horizontal or vertical forehead creases, and trouble maintaining eye contact. In order to be a good receiver of student communication, a lecturer must be aware of many of the subtle nonverbal cues that their students express [46]. Automated computer vision and deep learning-based approaches are the most popular methods for assessing learners' engagement based on their facial expression [16–19, 47–49]. There are several studies in the literature on detecting learners' engagement. Minyu et al. [50] suggested a theoretical basis for explaining student's facial expressions and auto-assessment for teaching and learning in the classroom. The student's attention is based on the head posture and five learners' expressions of focused, surprised, confused, joyful, and distracted. They characterize student's learning affects (SLA) and build an SLA transfer model using learning facial expressions and attentiveness. However, before it can be used to assess the efficacy of classroom instruction and learning, their theoretical SLA analysis must be tested "in the wild". Thomas et al. [51] has extracted multimodal features from both the speaker and listener audio and video data. The feature set includes audio features (extracted from Praat, OpenSMILE, pyAudioAnalysis toolbox), facial features (extracted from OpenFace), and posture features (extracted from OpenPose). Although their system aims to improve learning by analyzing instructor-learner interactions in the classroom, it does not reflect the level of engagement of learners. Kerdawy et al. [52] proposed an approach to use electroencephalography (EEG) and facial expression modalities to anticipate students' cognitive states, engagement, and spontaneous attention. They observed that, while the EEG and face-based models demonstrated significant agreement in the engaged classes, there was less consensus in the non-engaged case. The combination of the two modalities, EEG and facial expression, has the potential to improve performance, but implementing a brain-computer interface (BCI) module in a practical classroom or online learning situation would be difficult because of mental privacy, wearability, portability, and cost constraint [26, 53, 54]. Bhardwaj et al. [49], have introduced a deep learning approach to compute the students' Mean Engagement Score (MES) in real-time through emotion detection and emotion weights picked up from a survey carried out on students in an hour-long classroom. Tongu et al. [48] used Microsoft's emotion identification API to identify emotions like sadness, joy, fear, anger, surprise, and contempt throughout the lecture. Students' emotions were examined in relation to department, lecture hours, gender, session information, and other factors. Their research has revealed patterns of different emotional states over time, but it would be ideal to focus only on the emotions that are significant in a classroom setting, and all of these methods are unreliable in the real world. Gupta et al. [21] has created a DAiSEE dataset of 112 students that includes four affective states of boredom, engagement, confusion, and frustration with degrees of engagement (low, very low, high, and very high) in e-learning environments. The aim is to determine how engaged students are in online classrooms. They reported baseline results using several CNN-based video classification approaches, such as InceptionNet [55], C3D

[35], and Long-term Recurrent Convolutional Networks (LRCN) [56] models with 46.4%,56.1%, and 57.9% for four-class classification of engagement. Using LRCN, they have attained the Top-1 accuracy of 94.6% for the engagement label in binary classification, labelling from (low, very low) categories to "not engaged (0)" and (high, very high) categories to "engaged (1)". They only presented the performance metric as an accuracy in their studies and didn't address class imbalance problem. Wang et al. [29] also proposed a 2D CNN-based architecture that could detect the amount of engagement on still images with 57% accuracy. Huang et al. [27] has proposed a Deep Engagement Recognition Network (DERN) that comprises temporal convolution, bidirectional LSTM, and an attention mechanism. They have achieved a Top-1 accuracy of 60.0% of engagement recognition on DAiSEE and 94.2% for binary classification. However, all the above strategies did not account for class imbalances and were not validated on other engagement datasets. Zhang et al. [30] introduced Weighted Single RGB-stream inflated 3D convolutional network (WSRGB -I3D) with weighted cross-entropy loss and obtained 52.30% accuracy to classify four labels of engagement and achieved the highest accuracy of 98.82% in binary not-engaged/engaged, classification problem, but performed poor in four class classification. Geng et al. [57] offer a Convolutional 3D (C3D) approach for recognizing student interest in videos by modeling both visual and motion information. In order to tackle the class-imbalanced data distribution problem in engagement recognition, they employed focal loss [41]. The focal loss collects additional feature information from the various samples and increases accuracy to 56.2% over the C3D baseline by adaptive reducing the weight of high engagement samples while increasing the weight of low engagement samples. For engagement prediction, Liao et al. [28] suggested the Deep Facial Spatio-temporal Network (DFSTN). The model includes a pre-trained Squeeze-and-Excitation-ResNet-50 (SENet) as a facial spatial feature extractor, as well as a hidden state LSTM Network with a global attention that captures temporal information and increases result efficiency. They evaluated their method on the DAiSEE and reported 58.84% accuracy in the four-class engagement classification. To address the problem of imbalanced data distribution in the DAiSEE, a combination of cross-entropy loss and center loss is used to learn more discriminatory features. Although the methods presented above addressed the class imbalance problem for four class engagement but the highest accuracy obtained was less than 60%, and the confusion matrix presented in [28] did not perform well for minority classes. EmotiW-EP is another challenging dataset for predicting student engagement. [42, 43]. Niu et al. [58] introduced the Gaze-AU-Position (GAP) feature, which takes into consideration the test subject's gaze, action units, and head pose for engagement prediction in EmotiW-EP. EmotiW 2018 challenge winners Yang et al. [33] approach build a multi-modal regression model, including local binary pattern (LBP), convolutional 3D (C3D), and statistical temporal features such as gaze, head, and body posture, followed by LSTM-FC layers. The accuracy of each affective state for the model trained on authentic DAiSEE dataset is presented, and no approach other than [28] has been tested on both DAiSEE and EmotiW-EP. Therefore, a deep learning-based solution that can address class imbalances with improved accuracy and be evaluated on multiple datasets is required. In addition to the multi-class classification, we introduced DenseAttNet for multi-label classification, which predicts all four affective states with a single model. In another set of experiments the DenseAttNet is modified to perform the regression task, and the DenseAttNet pre-trained with DAiSEE is evaluated on the EmotiW-EP training and validation set to ensure the models' robustness.

3 Proposed method

Figure 1 shows the block diagram of the proposed pipeline for automated student engagement prediction in a video sequence. The proposed pipeline consists of multiple stages, including temporal down-sampling of video clip frames, face detection and alignment with Dlib [59], and prediction with the proposed neural network. The aligned face image is resized and concatenated to get $30\times224\times224\times3$-dimension image cube that is fed as input to the DenseAttNet (described in Section 3.2) model to predict affective states.

3.1 Pre-processing

In DAiSEE, a 10-second video clip with a resolution of 640 × 480 captured at 30 frames per second results in 300 frames. A significant proportion of the frames are redundant. Therefore, we chose 30 frames at a frame interval of 10 from the 300 frames in the video clip. As a next step, we use a robust real-time Dlib face detector to crop face regions from each selected frame. Afterward, the cropped faces are passed to a face alignment function that aligns the faces and rejects the frames which lack frontal face alignment or have facial occlusions, as shown in Fig. 12. The clean frontal faces are then resized to $224 \times 224 \times 3$. Thus, for each 10-second video, $30 \times 224 \times 224 \times 3$ ($D \times C \times H \times W$)-dimension image cubes are given to the DenseAttNet, where D is the number of frames extracted from the video, C is the number of channels, and H and W are the height and width of a video frame.

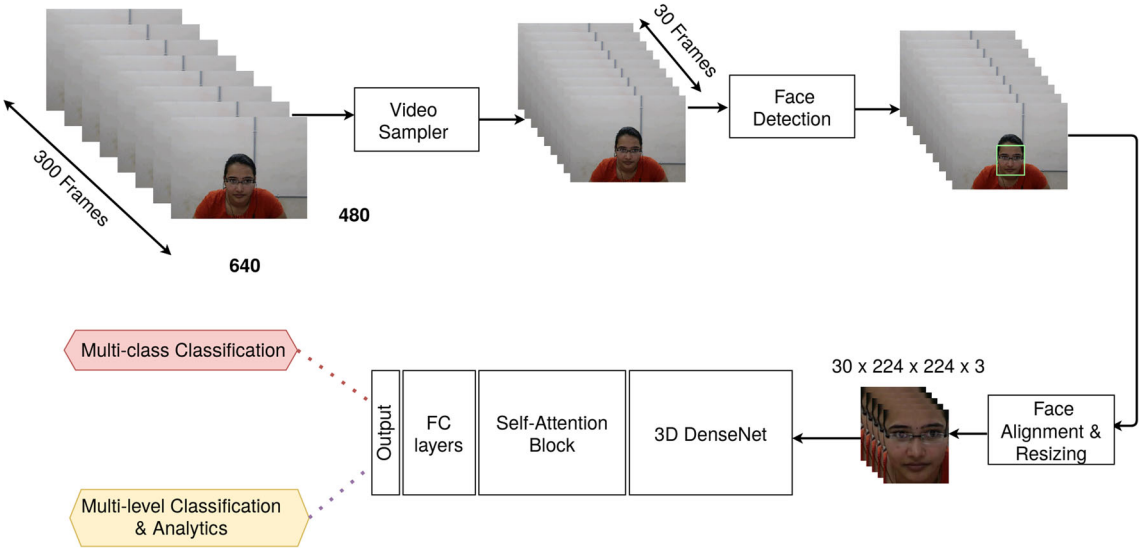


Fig. 1 Proposed pipeline for automatic students' engagement detection

3.2 DenseNet self-attention network (DenseAttNet)

One can express the task of students' engagement in a video sequence as spatial-temporal variations in the facial expressions. Therefore, it requires deep learning architectures that can extract the spatial-temporal relationship among the video sequence frames. It appears intuitive to use 3D CNN that is efficient in learning spatial-temporal relationships in video sequences. Furthermore, as a non-local operation, we used a self-attention module to characterize the global dependencies of the 3D CNN feature maps [60]. The self-attention module can help the model to focus on more relevant regions of the image, resulting in better performance. Such an attention approach enhances the discovery of new patterns in data by allowing models to learn deeper correlations between spatial or temporal dependencies between any two points in the input feature maps. In general, regardless of their proximity on the input image or feature maps, any two spots with similar features or strong dependencies will be represented in the correlation matrix and contribute to the final response by learning and focusing on critical areas of the input feature maps while suppressing irrelevant information [60–64].

Figure 2 presents our proposed DenseAttNet, which combines 3D DenseNet-121 [39] and a 3D self-attention module to capture global relationship between the features. It comprises multiple important building blocks, including 3D Convolutional dense and transitional blocks followed by a self-attention block, and fully connected classifier layers.

3.2.1 3D DenseNet

An image cube of students' facial expressions is supplied as input to the proposed 3D DenseNet-121 neural network. The DenseNet's dense block concatenates additional input feature maps from previous layers and feeds their feature maps to successive layers at a growth rate of $k = 32$ [39]. Each dense block in Fig. 2 comprises a batch normalization (BN), rectified linear unit (ReLU), and $1 \times 1 \times 1$ Convolution (Conv) layers followed by another set of BN, ReLU, and $3 \times 3 \times 3$ Conv layers. After each of the dense blocks, namely 3D Dense Block 1, 3D Dense Block 2, and 3D Dense Block 3 the transition layers are employed as bottleneck layers that comprises of BN, ReLU, and $1 \times 1 \times 1$ Conv followed by a $2 \times 2 \times 2$ average pooling layer with stride $1 \times 2 \times 2$. The bottleneck layers cut the number of input feature maps to half, increasing computational efficiency. Eventually, the fourth dense block output of $1024 \times 5 \times 7 \times 7$ ($C_o \times D_o \times H_o \times W_o$)-dimension image cube is given to the self-attention module.

3.2.2 Self-attention layer

We used three strategies to link global features across spatial, temporal, and both spatial and temporal dimensions. The attention module represented in Fig. 4a, b, and c reflect intra-frame, inter-frame, and both intra-frame and inter-frame dependencies in a video.

The DenseNet layers' output $x \in \mathbb{R}^{C_o \times D_o \times H_o \times W_o}$ dimensional image cube is transformed to the key $k(x_i)$, query $q(x_j)$ and value $v(x_i)$ by using $1 \times 1 \times 1$ convolution

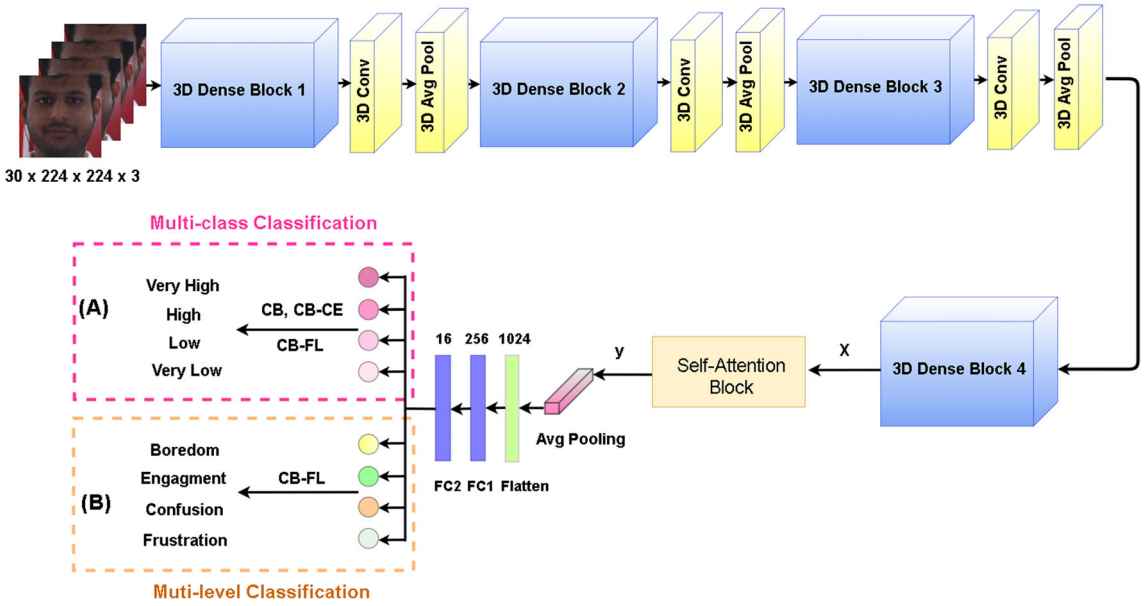


Fig. 2 DenseNet, self-attention layer, and FC classification layers are the three components of the proposed DenseAttNet

filters, where $k(x_i) = W_k x$, $q(x_j) = W_q x$, $v(x_i) = W_v x$ and W is the weight of the learned convolution filter. We have reduced the channel number of $\hat{C_0}$ to be $C_o/4$ for memory efficiency. As shown in Fig. 4a the spatial self-attention module is used to compute intra-frame dependencies of facial images. It computes each pixels' relationships to all other pixels in the frame and passes only those features with dominant intra-frame dependencies, as illustrated in Fig. 3a and mathematically expressed by (1). In (1), $S_{j,i} \in \mathbb{R}^{\hat{C_o} \times D_o \times H_o \times W_o \times H_o \times W_o}$ is the spatial correlation matrix obtained by softmax normalization of the inner product of $q(x_j)$ and $k(x_i)$. The relationships between pixels are represented by the spatial dimension $(H_o \times W_o) \times (H_o \times W_o)$ correspond to a 49×49 matrix.

$$S_{j,i} = \frac{exp(k(x_i)^T q(x_j))}{\sum_{i=1}^{H_0 \times W_0} exp(k(x_i)^T q(x_j))}. \tag{1}$$

The output attention features across spatial $o_{S\hat{att}} = \sum_{i=1}^{H_0 \times W_0} v(x_i) S_{j,i} \in \mathbb{R}^{\hat{C_o} \times H_o \times W_o \times D_o}$ is fed through the $1 \times 1 \times 1$ convolution filter, which results in o_{Satt} attention feature maps with increase in channel number to the original C_o channels. The final result attained by adopting the spatial attention module y_s is given by:

$$y_s = \gamma o_{Satt} + x. \tag{2}$$

Temporal self-attention in Fig. 4b is used to compute inter-frame dependencies of facial images and relates the global features from other facial images in temporal domain. As illustrated in Fig. 3b and mathematically expressed by (3), $T_{j,i}$ is $\mathbb{R}^{\hat{C_o} \times H_o \times W_o \times D_o \times D_o}$ dimensional temporal correlation matrix. The relationships between pixels in depth dimension $D_o \times D_o$ correspond to a 5×5 matrix.

$$T_{j,i} = \frac{exp(k(x_i)^T q(x_j))}{\sum_{i=1}^{D_0} exp(k(x_i)^T q(x_j))}. \tag{3}$$

The output attention features across temporal $o_{T\hat{att}} = \sum_{i=1}^{D_o} v(x_i) T_{j,i} \in \mathbb{R}^{\hat{C_o} \times D_o \times H_o \times W_o}$ is fed through the $1 \times 1 \times 1$ convolution filter, which results in o_{Tatt} attention feature maps with increase in channel number to the original C_o channels. The final result attained by adopting the temporal attention module y_t is given below:

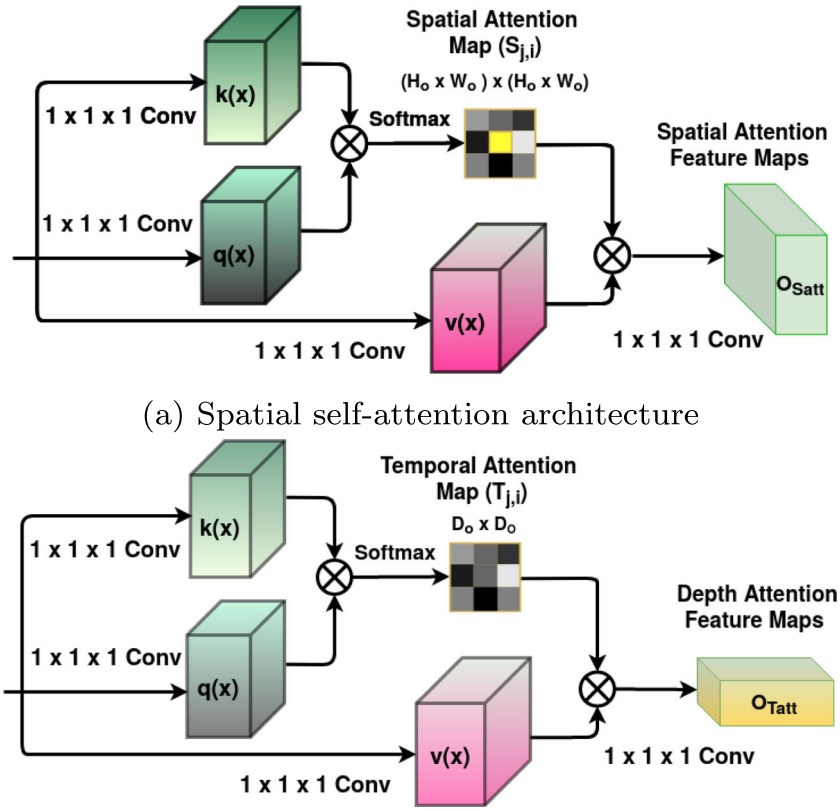
$$y_t = \gamma o_{Tatt} + x. \tag{4}$$

The hybrid module, which combines both inter-frame and intra-frame dependencies, is shown in Fig. 4c, and the final result obtained by combining the spatial and temporal attention module is given in the following equation:

$$y_h = \gamma (o_{Tatt} + o_{Satt}) + x. \tag{5}$$

where γ is a scalar parameter that is initially set to 0 and can be learned as training advances further. Allow the network to rely on local cues at first, then gradually increase the γ value to give non-local evidence more weight.

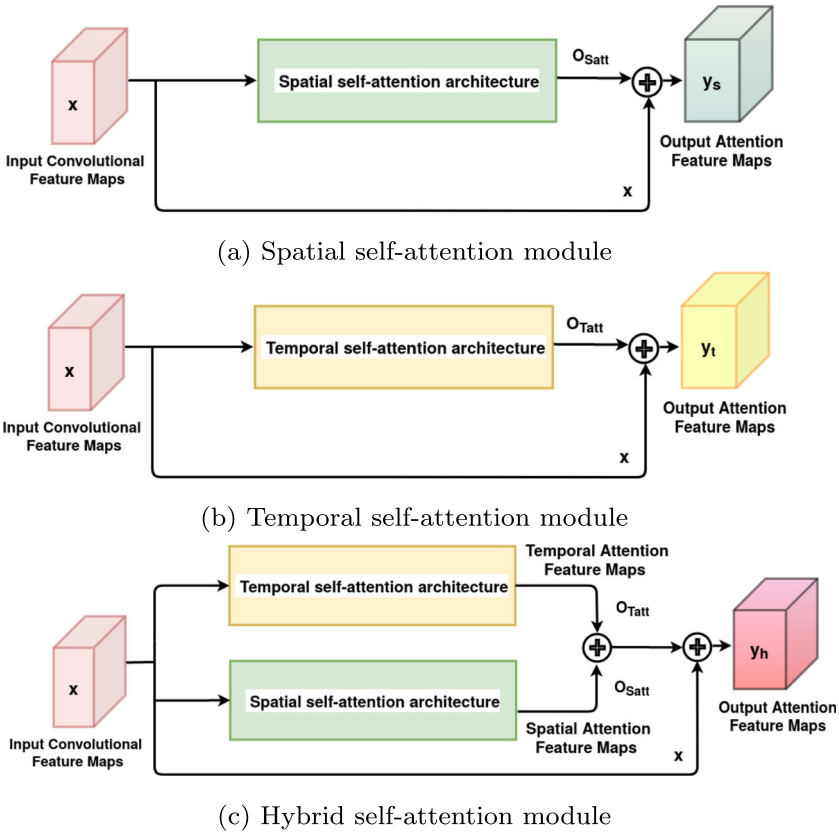
The self-attention blocks' output feature map is subsequently sent to a $1 \times 1 \times 1$ Adaptive Global Pooling layer, and the 1024 flattened features are passed to three sequential fully connected layers, namely FC1, FC2, and Output. The multi-class classification problem as illustrated in Fig. 2A is used for predicting the affective states' all four levels. In



(a) Spatial self-attention architecture

(b) Temporal self-attention architecture

Fig. 3 (a) 3D spatial self-attention architecture (b) 3D temporal self-attention architecture. Where ⊗ denotes element-wise multiplication operation and ⊕ denotes element-wise addition operation. Here, temporal or depth is an interchangeable term



(a) Spatial self-attention module

(b) Temporal self-attention module

(c) Hybrid self-attention module

Fig. 4 The spatial-attention module (which uses the Fig. 3a spatial self-attention architecture), the temporal-attention module (uses the Fig. 3b temporal self-attention architecture), and the hybrid module includes both spatial and temporal self-attention architecture

Fig. 2B all four affective states of the binarized DAiSEE dataset (described in Section 4.1.1) is predicted with a single model. In Section 3.3 and 4.2.2 the loss and hyperparameter settings are described in greater detail.

3.3 Class-balanced (CB) loss

In general, when neural networks are trained on class-imbalanced datasets the classes with a larger number of examples are easier to classify whereas classes with a few samples are more difficult to classify. To handle class-imbalanced datasets mainly two types of techniques have been proposed.

One approach is to use synthetic samples and re-samples the training dataset for imbalanced classes. To some extent, re-sampling can assist to re-balance the distribution of training data, but it can also lead to model over-fitting [65–67]. Furthermore, in the multi-label DAiSEE with 6 : 22 : 2 : 1 imbalance ratio for Boredom: Engagement: Confusion: Frustration affective states, one affective state is linked to another affective state, for example, DAiSEE data label [0, 2, 1, 1] indicates that engagement (majority class), confusion (minority class), and frustration (minority class) are all high at the same time. Hence, to balance the dataset in this situation eliminating samples from classes with large sample numbers is simply incorrect.

While the other approach is to re-weight the training loss based on the imbalance. Re-sampling is the process of oversampling classes with minority samples, under-sampling classes with majority samples and occasionally using both approaches [68, 69]. As a result, the only alternative is to employ the CB loss parameter which provides a weighting factor that is inversely proportional to the effective number of samples that performs better than loss re-weighting by inverse class frequency, to solve the difficulty of training from imbalanced data. Given input **x** and label **y** with **C** total number of classes, the models' estimated class probabilities **p**, and loss denoted by L(**p**,y). Cui et al. [68], introduced $E_{n_y} = (1 - \beta_y^{n_y})/(1 - \beta_y)$ the effective numbers of samples to employ inverse class frequency to alter the CB parameter between non re-weighting and re-weighting smoothly, theoretically β = (N - 1)/N, N is the volume of all possible samples in the feature space of class that is difficult to obtain empirically. Thus, we select the hyper parameters $\beta \in \{0.9, 0.99, 0.999\}$, which was also used by Cui et al. [68] in their experiments. In the above equation, n_y represents the number of samples for a given engagement level. The CB loss is denoted by the following formula:

$$CB(\mathbf{p}, y) = \frac{1}{E_{n_y}} L(p, y) = \frac{1-\beta}{1-\beta^{n_y}} L(p, y). \quad (6)$$

3.3.1 Classification loss

Two distinct versions of the CB classification loss functions are used in the experiment. Firstly, CB-CE, as shown in (7) is used in DAiSEE multi-class classification.

$$CB - CE(\mathbf{p}, y) = -\frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{C} log(Softmax(p_i)). \quad (7)$$

Secondly, we have used CB-FL described in (9) for both the multi-class and multi-label classification. The Focal loss [41] reduces the relative loss for well-identified samples while increasing the relative loss for badly categorized samples by adding a modulating component $(1 - p_t)^{\gamma}$ to the sigmoid cross-entropy loss. In the case of sigmoid cross-entropy loss, the class probabilities are computed with the assumption that each class is independent and not mutually exclusive. Single-label classification and multi-label prediction are combined in sigmoid. This is a useful feature to have since DAiSEE data includes more than one semantic label and to train all the affective states with a single classifier model. One-hot encoding is used to encode the actual labels y and for notational convenience p_t is used.

$$p_t = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{otherwise,} \end{cases} \quad (8)$$

$$CB - FL(\mathbf{p}, y) = -\frac{1-\beta}{1-\beta^{n_y}} \sum_{y=1}^{C} (1-p_t)^{\gamma} log(Sigmoid(p_t)). \quad (9)$$

Here $\gamma \in \{1, 2\}$ is tunable focusing hyper parameter that modulates the original sigmoid cross-entropy loss by $(1 - p_t)^{\gamma}$ factor.

3.3.2 Regression loss

To train the DAiSEE engagement class using the DenseAttNet and to quantify engagement levels in terms of continuous value, the MSE loss function in (10) is used. MSE tries to minimize the difference between the p_i predicted value and y_i ground-truth value of the i-th sample. The definition of the MSE is as follows:

$$MSE(\mathbf{p}, y) = \frac{1}{D} \sum_{i=1}^{D} (p_i - y_i)_2^2. \quad (10)$$

Where D represents the batch size. The $C \in \{0.0, 0.33, 0.66, 1.0\}$ represents the total number of classes. Because the DAiSEE engagement prediction regression task contains just four target values C that the model needs to predict,

the same CB loss parameter is used with MSE, which is defined as:

$$CB - MSE(\mathbf{p}, y) = \frac{1-\beta}{1-\beta^{n_y}} \frac{1}{D} \sum_{y=1}^{C} \sum_{i=1}^{D} (p_i - y_i)_2^2. \quad (11)$$

4 Experiments and results

This section begins with an overview of the DAiSEE and EmotiW-EP datasets. We presented the experimental setup, training strategy as well as the the ablation experiments to select the best model and hyper parameters. On the DAiSEE, we present our experimental results in terms of numerous performance measures, including accuracy, precision, recall, and F1-score [70, 71]. All of the metrics presented above for affective state classification are derived from the confusion matrix of the validation set, which is one of the most intuitive and descriptive measures. In our next set of experiments, the model is used to train on DAiSEE and EmotiW-EP datasets as a regression problem where the labels are pseudo-continuous (mapped in between 0-1). The MSE value is used as an evaluation metric.

4.1 Dataset details

This section describes the DAiSEE and EmotiW-EP datasets, which contain user affective states captured in the wild. These datasets are used to evaluate our models' performance.

4.1.1 DAiSEE

DAiSEE is a collection of 112 peoples' 10-second video snippets captured at a resolution of 1920 × 1080 pixels at 30 frames per second. It includes four affective states, namely boredom, confusion, engagement, and frustration. Each affective state is labelled with one of four levels: very low, low, high, and very high. Each emotional state is crowd-annotated and linked with a gold standard annotation generated by a team of psychologists [21]. Individual DeneseAttNets are trained on each affective state while addressing the DAiSEE as a multi-class classification problem. In the case of multi-label classification, each affective state label is binarized (low and very low to "0", high and very high to "1"). For the regression problem, the four engagement levels are assigned to the values "0.0", "0.33", "0.66", and "1.0".

4.1.2 EmotiW-EP

The EmotiW-EP is a sub-challenge of the EmotiW, which is an engagement in the wild challenge. [42, 43]. Faces of the individuals watching an instructional video (MOOC) are recorded using a Microsoft Life camera at 640 × 480 pixels at 30 frames per second. The video is 5 minutes duration on average. The dataset includes 147, 48, and 67 videos for the train, validation, and test sets, respectively. The engagement levels are split into four categories: 0, 0.33, 0.66, and 1 with "0" denote full disengagement while "1" denote high engagement of the user. EmotiW-EP 5-minute training and validation set are split into several 10-second video clips excluding the initial and final 10 seconds of the video to avoid unnecessary noise and to justify the persons' stabilization time, which is then pre-processed in the same way as DAiSEE for the evaluation of our method and the average results reported on a 10-second segment of a 5-minute video clip.

4.2 Classification experiments

A 30-frames clip is used as input to train the model from scratch with a batch size of 16 clips. During training, we used Stochastic Gradient Descent (SGD) optimizer [72]. The hyper parameters for the SGD are as follows: the initial learning rate (LR) as 0.001, the momentum of 0.9, and a weight decay of $1 \times e^{-5}$ after every 15 epochs. 5-fold cross-validation is used to measure the performance of the model. All the experiments have been implemented on the PyTorch framework and executed on an NVIDIA Tesla V100 with 32 GB GPU memory. Our model contains 19.3M parameters, 53G floating-point operations per second (FLOPs) approximately, and 76.7 MB model size.

4.2.1 Ablation study of DenseAttNet

The performance and results of 3D DenseNet, 3D DenseNet with self-attention, and 3D DenseNet with self-attention and fully connected layers, namely FC1, FC2, and Output for the engagement class using CE loss are shown in Table 1. As described in Section 3.2.2, the attention mechanism is good at modelling long-term dependence of emotional states; the model with the self-attention module outperforms DenseNet without self-attention by 0.04% improvement in the accuracy term, and self-attention becomes more prominent when it is trained with class-balanced losses to address the DaiSEE imbalanced problem discussed in Section 4.2.2. The Grad-CAM visualization in Fig. 10 shows attention maps, adoption of the self-attention module in the DenseAttNet model for more locally-focused attention, highlighting the utility of using self-attention blocks to predict engagement. Furthermore, DenseAttNet with self-attention and FC layers improved accuracy by 0.72% and 0.003 reductions in loss term. The FC layer acquires the activation maps produced by the previous Conv layers, and the FC layer weights

Table 1 Performance evaluation of a four-class engagement classification on DAiSEE for architecture selection

3D DenseNet			3D DenseNet + SA			3D DenseNet + SA + FC		
Loss	F1-Score	Accuracy (%)	Loss	F1-Score	Accuracy (%)	Loss	F1-Score	Accuracy (%)
0.373	0.59	61.39	0.372	0.59	61.43	**0.369**	**0.60**	**62.15**

Where SA and FC stand for self-attention and fully connected layers, respectively

combined with the built-in non-linear activation form a possible stochastic probability representation for each class, resulting in improved performance.

4.2.2 Evaluating models' performance on CB loss functions

DenseAttNet has been trained on DAiSEE's engagement class. The CB-FL parameters are $\beta \in \{0.9, 0.99, 0.999\}$ and $\gamma \in \{1.0, 2.0\}$, while the CB-CE parameter $\beta \in \{0.99\}$. Based on the confusion matrix obtained from the 5-fold cross-validation experiment given in Figs. 5, 6, and 7 and the resultant accuracy presented in Tables 2, 3, and 4 the following observations can be made: Due to the significant imbalance in the sample distribution, our model using CE loss is unable to categorize the "low engagement" and "very low engagement" samples because of the loss caused by misclassifying samples with "high engagement" and "very high engagement" outweighed the loss caused by misclassifying samples with "very low engagement" and "low engagement". Neither the spatial self-attention module nor the temporal self-attention module used in DenseAttNet correctly predicts the "low engagement" samples, and with the spatial self-attention module, not even a single "very low engagement" sample is correctly classified, whereas temporal self-attention does as seen in Fig. 6b.

The hybrid self-attention DenseAttNet using CB-FL loss with $\beta = 0.9$ and $\gamma = 1.0$, along with the predictions for "high engagement" and "very high engagement" levels, the model has shown improvement in the predictions for "low engagement" illustrated in Fig. 7b. The model achieved an all-time state-of-the-art accuracy of 63.59% in the problem of four-class engagement classification displayed in Table 4. The model correctly classifies some "very low engagement" samples with $\beta = 0.99$ and $\gamma = 1.0$, but the overall classifier performance decreases to 62.38%.

The model performance degrades with higher β values, such as $\beta = 0.999$, since the effective number of samples approaches the number of samples, and it is the same as re-weighting by inverse class frequency [68], results in a 7-21 times rise in the "very low engagement" loss term, the

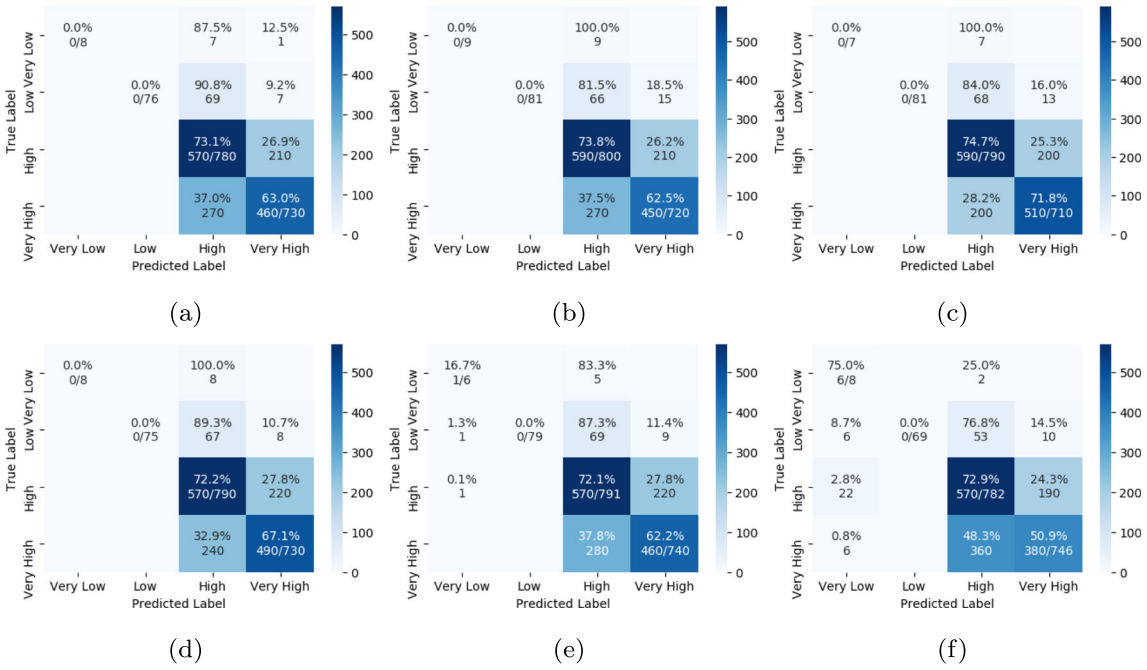


Fig. 5 Confusion matrix of the proposed DenseAttNet with a spatial self-attention module tested on DAiSEE with (**a**) Cross-Entropy loss (CE) (**b**) CB-FL ($\beta = 0.9$, $\gamma = 1.0$) (**c**) CB-FL $\beta = 0.9$, $\gamma = 2.0$) (**d**) CB-FL ($\beta = 0.99$, $\gamma = 1.0$) (**e**) CB-FL ($\beta = 0.99$, $\gamma = 2.0$) (f) CB-FL ($\beta = 0.999$, $\gamma = 1.0$)

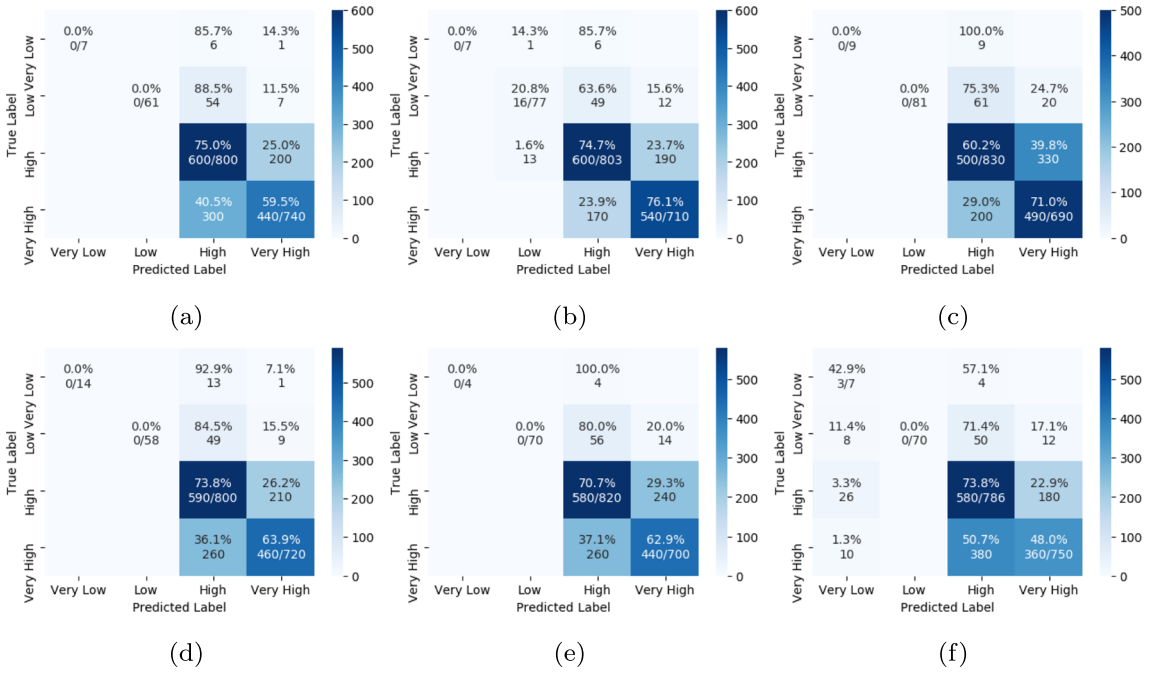


Fig. 6 Confusion matrix of the proposed DenseAttNet with a temporal self-attention module tested on DAiSEE with (**a**) CE (**b**) CB-FL ($\beta = 0.9$, $\gamma = 1.0$) (**c**) CB-FL ($\beta = 0.9$, $\gamma = 2.0$) (**d**) CB-FL ($\beta = 0.99$, $\gamma = 1.0$) (**e**) CB-FL ($\beta = 0.99$, $\gamma = 2.0$) (**f**) CB-FL ($\beta = 0.999$, $\gamma = 1.0$)

classifier over-fits and trains harder on the minority "very low engagement" class lowering overall classifier accuracy to 57.24%, as illustrated in Fig. 7f confusion matrix and in Table 4. In our studies, we observed that $\gamma = 1.0$ is the optimum value. DenseAttNet model utilizing the hybrid self-attention module delivers the best results for the engagement prediction. The overall performance of the classifier is improved by using CB-FL.

The binary conversion of engagement dataset for binary engagement classification resulted in an imbalance ratio

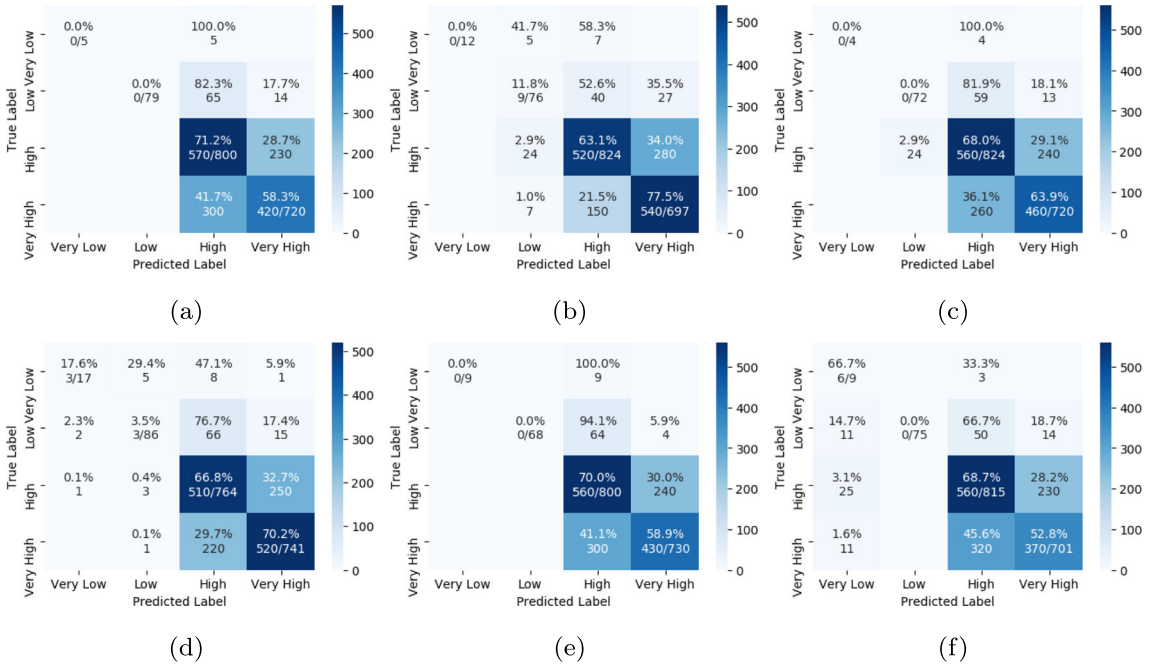


Fig. 7 Confusion matrix of the proposed DenseAttNet with a hybrid self-attention module tested on DAiSEE with (**a**) CE (**b**) CB-FL ($\beta = 0.9$, $\gamma = 1.0$) (**c**) CB-FL ($\beta = 0.9$, $\gamma = 2.0$) (**d**) CB-FL ($\beta = 0.99$, $\gamma = 1.0$) (**e**) CB-FL ($\beta = 0.99$, $\gamma = 2.0$) (**f**) CB-FL ($\beta = 0.999$, $\gamma = 1.0$)

Table 2 Spatial self-attention four-class engagement classification results using CB losses and optimum hyper parameter search

Loss Type	CE	CB-CE	CB-FL	CB-FL	CB-FL	CB-FL	CB-FL
β	-	0.99	0.9	0.99	0.9	0.99	0.999
γ	-	-	1.0	1.0	2.0	2.0	1.0
Precision	0.59	0.58	0.59	0.60	0.60	0.58	0.57
Recall	0.62	0.61	0.62	0.63	0.63	0.60	0.59
F1-Score	0.60	0.59	0.59	0.61	0.61	0.58	0.56
Accuracy (%)	62.15	61.64	61.60	**63.25**	62.89	60.33	58.48

Table 3 Temporal self-attention four-class engagement classification results

Loss Type	CE	CB-CE	CB-FL	CB-FL	CB-FL	CB-FL	CB-FL
β	-	0.99	0.9	0.99	0.9	0.99	0.999
γ	-	-	1.0	1.0	2.0	2.0	1.0
Precision	0.58	0.57	0.58	0.58	0.58	0.58	0.56
Recall	0.60	0.58	0.60	0.61	0.60	0.61	0.57
F1-Score	0.57	0.54	0.59	0.59	0.58	0.59	0.56
Accuracy (%)	60.19	61.15	**62.54**	61.65	60.71	61.10	57.52

Table 4 Hybrid self-attention four-class engagement classification results

Loss Type	CE	CB-CE	CB-FL	CB-FL	CB-FL	CB-FL	CB-FL
β	-	0.99	0.9	0.99	0.9	0.99	0.999
γ	-	-	1.0	1.0	2.0	2.0	1.0
Precision	0.59	0.59	0.61	0.60	0.58	0.59	0.56
Recall	0.61	0.52	0.64	0.63	0.61	0.62	0.57
F1-Score	0.60	0.60	0.62	0.61	0.59	0.60	0.56
Accuracy (%)	61.67	61.73	**63.59**	62.38	61.30	61.77	57.24

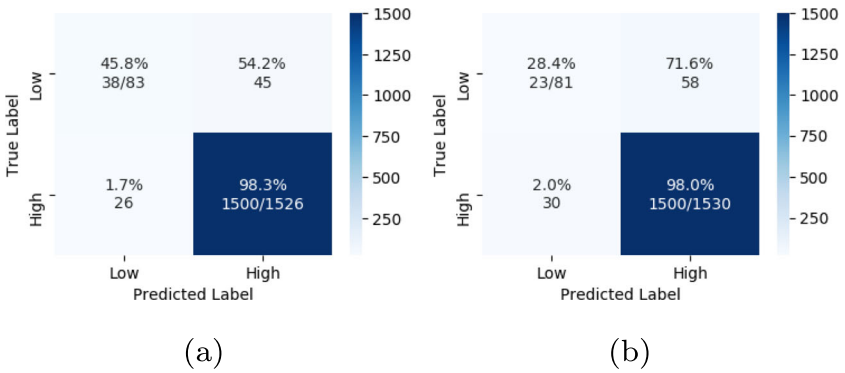


Fig. 8 Confusion matrix for binary engagement prediction on DAiSEE using the proposed DenseAttNet method. We have used hybrid self-attention module with (**a**) CB-FL ($\gamma = 1$, $\beta = 0.9$) (**b**) CB-FL ($\gamma = 1$, $\beta = 0.99$)

of 1:18 between the "not-engaged" and "engaged" classes. We obtained 94.78% accuracy using hybrid self-attention DenseAttNet and CB-FL ($\beta = 0.9$ and $\gamma = 1$) as reported in Table 5. The confusion matrix of the cross-validation result is presented in Fig. 8, and it can be observed that the model is able to predict the "not-engaged" minority class samples and deal with the data imbalance problem. Grad-CAM, a standard gradient-based localization method is used to generate visual explanations for classification decisions to gain a better understanding of the model. We present a mean visualization of DenseAttNet's self-attention layer activation for engagement classification. In Fig. 9 each column represents the temporal relationship of the engagement obtained in the 6th, 12th, 18th, 24th, and 30th frames. Whereas each row represents a level of engagement ranging from "very low" to "very high". It is obvious from Fig. 10 that adding a self-attention module to our architecture has resulted in a finer attention map in the input image, leading to better engagement prediction results.

4.2.3 Comparison with existing methods

Table 6 reports the results and comparison of the proposed DenseAttNet trained on DAiSEE with the current state-of-the-art methods. In four-class classification, the DenseAttNet model utilizing self-attention module with CB-FL ($\beta = 0.9$ and $\gamma = 1$) has outperformed the baseline result attained by LRCN [21]. The proposed model achieved an improvement of 5.35% for the engagement state and 0.57% for the boredom state. For the binary engagement classification, on the other hand, it achieved a boost of 0.35% in accuracy compared to LRCN. Besides, the proposed model has improved the best-published accuracy reported by DERN [27] for the four-class and binary class engagement classification from 60% to 63.59% and 94.2% to 94.78%, respectively. In four-class engagement, the DFSTN [28] method achieves 58.8% accuracy. Furthermore, the I3D model [30] has registered a superior accuracy of 98.82% percent for binary class engagement; it performs poorly in four-class engagement classification and attains an accuracy of 52.4%. Also, the DenseAttNet model using self-attention module with CB-FL ($\beta = 0.9$ and $\gamma = 1$) attained competitive accuracy of 69.22% and 78.58% for confusion and frustration, respectively. The LRCN [21] has recorded the best performance for the confused state, with an accuracy of 72.30%. While for the frustration state, the fine-tuned C3D model [21] with an accuracy of 79.10% has registered the best performance. Nevertheless, compared to LRCN and C3D, the proposed DenseAttNet model is computational efficient and more suitable for real-world applications.

4.2.4 Multi-label classification of DAiSEE

As discussed, in Section 4.2 SGD optimizer and the CB-FL loss function is used to train the multi-label DenseAttNet for all four affective states. In Table 7, the DenseAttNet trained with CB-FL ($\gamma = 1, \beta = 0.99$) as binary multi-label classification for all four classes produces an acceptable accuracy of 81.17% for boredom, 94.85% for engagement (only 0.1% less than multi-class binary classification), 90.96% for confusion, and 95.85% for frustration. As illustrated in Fig. 11, the learned features of the model create clusters that correlate to distinct affective states, such as B - boredom, E - engagement, and so on. The overlap clusters indicate more than one affective state is present; for example, BECF-boredom, engagement, confusion, and frustration states are all high at the same time. The features in each cluster are scattered due to intra-class variations in emotional states. It is clear that our model is capable of distinguishing all possible permutations of a learners' emotional states.

4.3 Regression experiments

The regression task is carried out by modifying the network's last layer and employing the MSE function. The CB parameter β has been introduced into the MSE function, as presented in (11). Setting $\beta = 0.9$ and $\beta = 0.99$, yields weight balancing terms equal to "1.01", "0.99", "0.99", "0.99", and "1.88", "0.71", "0.69", "0.69", for "very low engagement (0)", "low engagement (0.33)", "high engagement (0.66)", and "very high engagement (1)". The model trained with weighted MSE outperforms the model trained just with MSE because weight component allows punishing more on the minority "very low engagement (0)" and "low engagement (0.33)" samples, allowing the model to generalize the engagement prediction better. In Table 8 with a CB parameter $\beta = 0.9$, for the DAiSEE engagement class, our technique has the lowest MSE of 0.0347, outperforming all earlier methods.

Table 5 Binary engagement classification results with CB-FL loss

Loss Type	Precision	Recall	F1-Score	Accuracy (%)
CB-FL ($\gamma = 1.0, \beta = 0.9$)	0.95	0.96	0.95	**94.78**
CB-FL ($\gamma = 1.0, \beta = 0.99$)	0.94	0.95	0.94	94.66

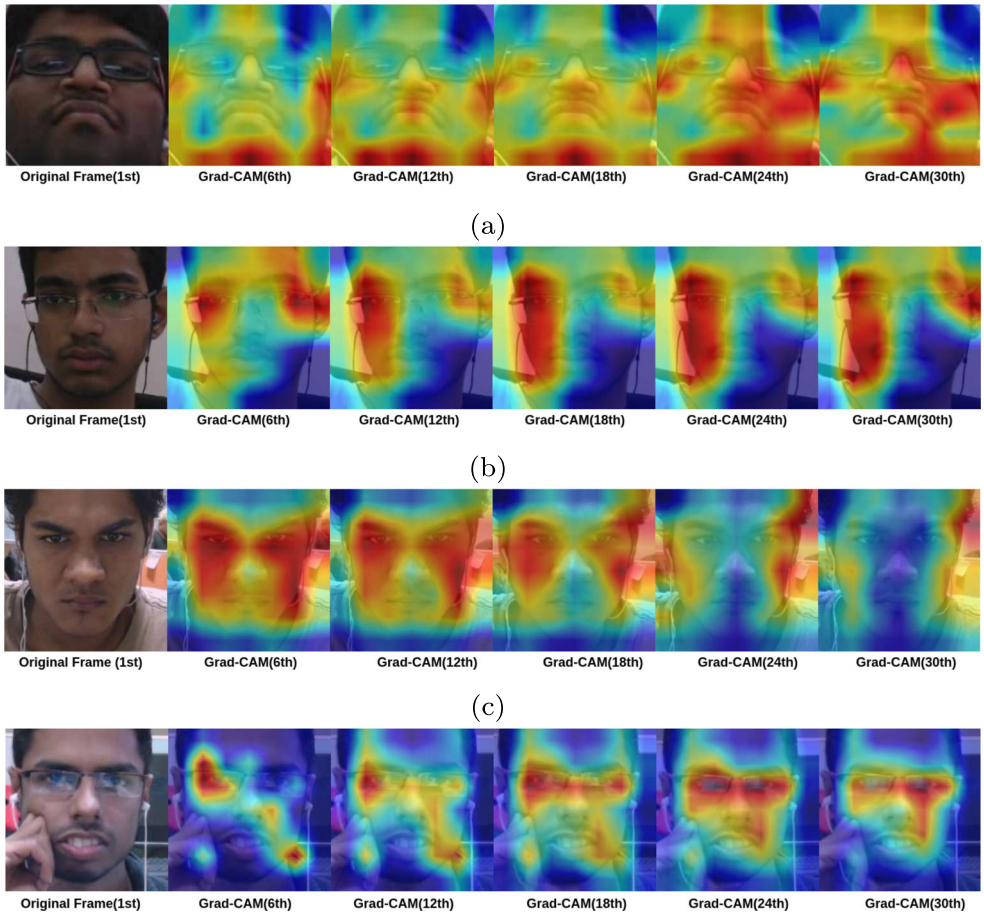


Fig. 9 Grad-CAM results on DAiSEE (**a**) engagement level "0" (**b**) engagement level "1" (**c**) engagement level "2" (**d**) engagement level "3"

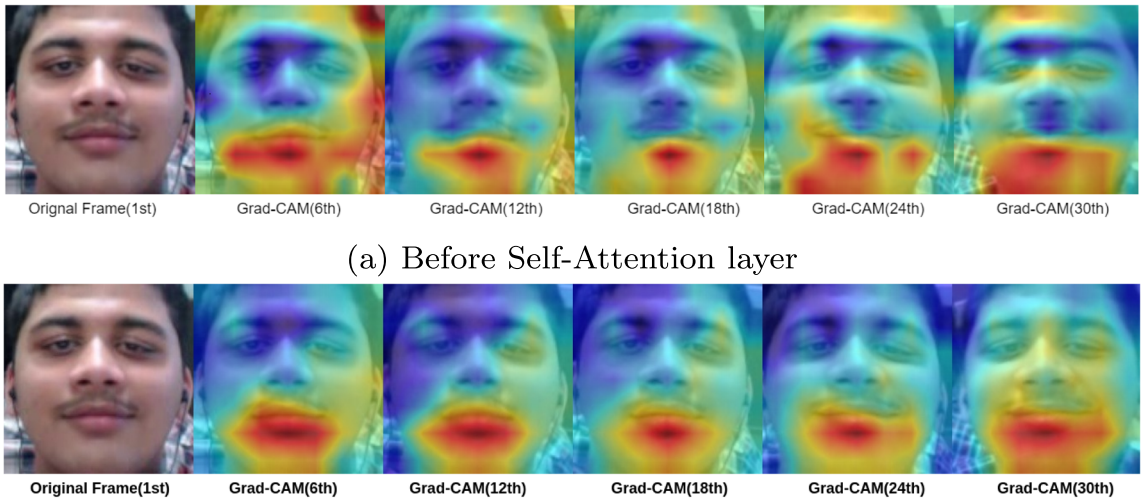


Fig. 10 Models' Grad-CAM visualization of engagement prediction on DAiSEE

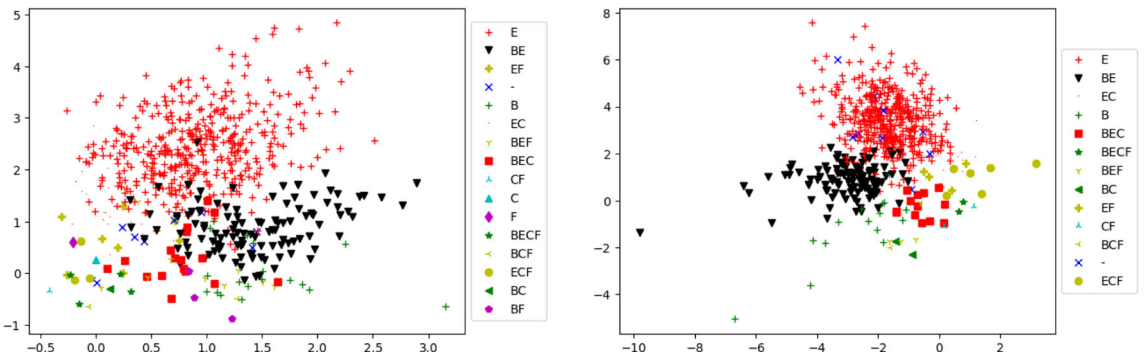
Table 6 Performance comparison with the benchmark results on DAiSEE

Method	Affective State Accuracy (%)				
	Boredom Four	Engagment Four	Binary	Confusion Four	Frustration Four
InceptionNet Frame [21]	36.5	47.1		70.3	78.3
InceptionNet Video [21]	32.3	46.4		66.3	77.3
I3D [30]		52.4	**98.82**		
C3D FineTuning [21]	45.2	56.1		66.3	**79.1**
LRCN [21]	53.7	57.9	94.6	**72.3**	73.5
DFSTN [28]		58.8			
DERN [27]		60.0	94.2		
DenseAttNet (Ours)	**54.27**	**63.59**	94.78	69.22	78.58

The best results are highlighted in bold

Table 7 Multi-label binary classification of DenseAttNet on DAiSEE

Loss Type	Multi-label classification accuracy (%)			
	Boredom	Engagement	Confusion	Frustration
CB-FL ($\gamma = 1, \beta = 0.9$)	78.58	94.81	91.74	96.37
CB-FL ($\gamma = 1, \beta = 0.99$)	81.17	94.85	90.96	95.85



(a) (b)

Fig. 11 An illustration of the multi-label DenseAttNet models' deep features learned from a multi-label DAiSEE dataset. Model trained using focal loss (**a**) $\beta = 0.9$ and $\gamma = 1$ (**b**) $\beta = 0.99$ and $\gamma = 1$. Where boredom (**B**), engagement (**E**), confusion (**C**), frustration (F), and None (-)

Table 8 DenseAttNet model performance comparison on DAiSEE

Method	Metric	Error
C3D (Scratch) [28]	MSE	0.0421
C3D (Fine Tuned) [28]	MSE	0.0442
DFSTN [28]	MSE	0.0422
DenseAttNet (Proposed)	MSE	0.0352
DenseAttNet (Proposed)	CB-MSE ($\beta = 0.9$)	**0.0347**
DenseAttNet (Proposed)	CB-MSE ($\beta = 0.99$)	0.0362

Table 9 DenseAttNet model performance comparison on EmotiW-EP

Method	Metric	Data Splits	Error
Dhall et al. (Baseline) [42]	MSE	val test	0.1 0.15
Yang et al. [33]	MSE	val test	0.0398 0.0626
DFSTN [28]	MSE	train + val	0.0736
DenseAttNet (Proposed)	MSE	train + val	0.0974
DenseAttNet (Proposed)	CB-MSE ($\beta = 0.9$)	train + val	0.0978
DenseAttNet (Proposed)	CB-MSE ($\beta = 0.99$)	train + val	0.0877

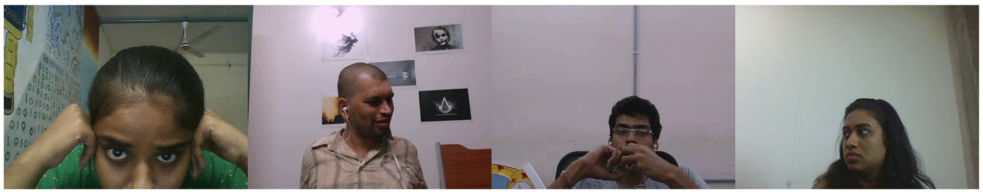


Fig. 12 This figure illustrates instances of non-frontal and occluded faces

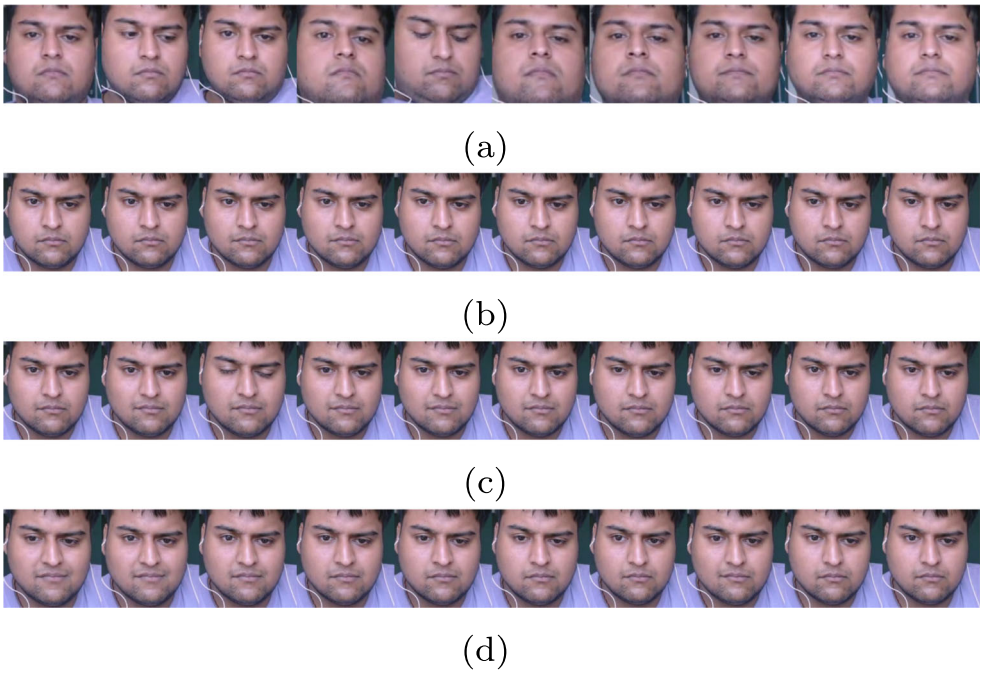


Fig. 13 Ten consecutive sequential frames of the same subject taken from DAiSEE, each labeled differently. **a** Sequence with "very low engagement", **b** sequence with "low engagement", **c** sequence with "high engagement", and **d** sequence with "very high engagement"

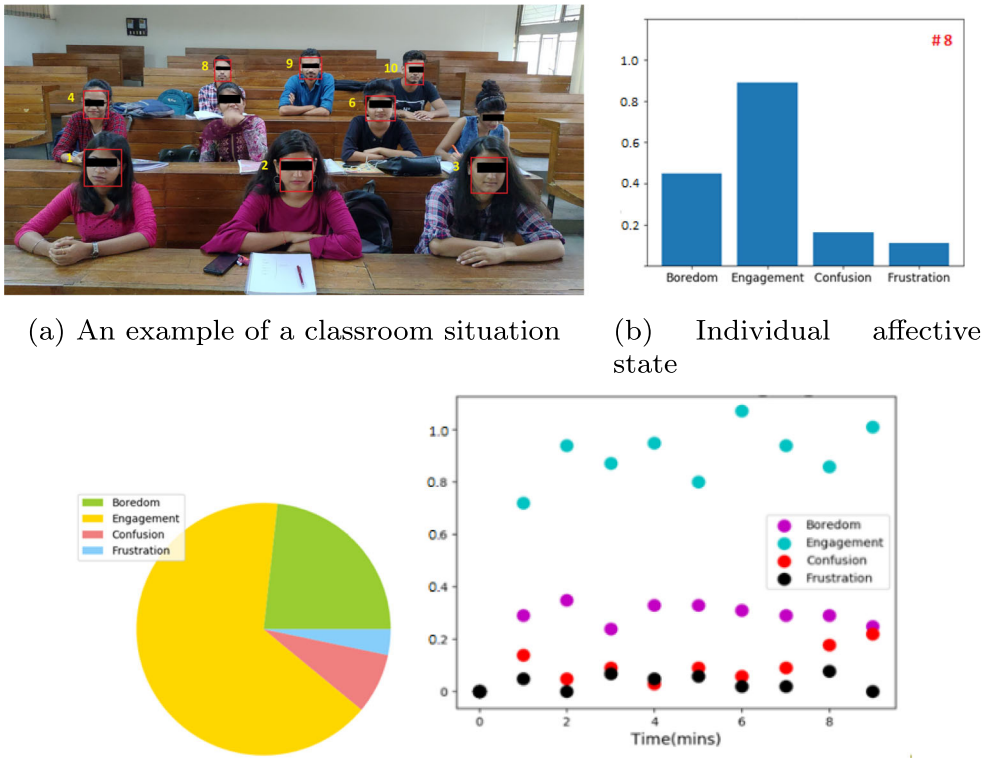


Fig. 14 Multi-label DenseAttNet algorithm-generated graph analytics samples of students' engagement during a lecture in a traditional classroom setting

(a) An example of a classroom situation (b) Individual affective state

(c) Overall affective state (d) Emotional state accumulated over time

To validate the efficiency of our method, similar to Liao et al. approach, DenseAttNet pre-trained on DAiSEE is used to predict engagement on EmotiW-EP. Table 9 compares the performance of the proposed DenseAttNet model to existing methods on the EmowitW-EP training and validation set. The proposed method obtained the best MSE of 0.0877 with $\beta = 0.9$ as compared to the DFSTN [28] method as well as the current state-of-the-art methods based on the fact that DenseAttNet has no prior experience with EmotiW-EP.

4.4 Discussion

We conducted a rigorous analysis of engagement prediction in DAiSEE to establish the optimal model and hyper parameter value for CB loss. The results obtained by the proposed method on DAiSEE engagement classification outperformed all previous methods in terms of accuracy and generalizing ability to predict all four levels, ranging from "very low engagement" to "very high engagement", despite the labels having a high imbalance ratio. Even after removing some strongly hand-over-face gestures and non-frontal faces in Fig. 12 during the data processing step, the models' engagement accuracy is less than 65% because different labels sequences, particularly the "low engagement", "high engagement", and "very high engagement" are quite similar, as shown in Fig. 13.

In addition to finding a method that can separate these similar-appearing different labels of sequences and to learn the hand-over-face gestures features, there is a need for more data with balanced and discriminative labels to improve the accuracy of engagement prediction [28].

Following that, we have trained and tested DenseAttNet on multi-label DAiSEE using CB-FL with $\gamma = 1$ and $\beta = 0.9, 0.99$. As shown in Fig. 11, the feature maps of the last dense layer form a cluster for each emotional state, which means that the classifier correctly predicts the emotion. Thereafter, in the regression experiment, the MSE for DAiSEE and EmotiW-EP is evaluated and it is found that the use of a CB-MSE helped the proposed model to achieve a lower MSE.

We also built a pipeline for the real-time deployment of the multi-label DenseAttNet in a classroom environment. Real-time video sequences of students in the classroom are captured using a Logitech HD webcam. The video clips are sampled every 10-seconds at regular intervals, the students' faces are cropped and aligned, and the processed image cube is provided to the model for prediction. Once obtained, the prediction results are saved in a Comma-separated values (.CSV) file in order to project engagement analytics for individual students as well as the entire class. Using the affective state as a binary label (present/absent) alone may not be the best option, as it may even influence the

instructors' decision in practical applications. For example, if a students' emotional state is halfway between 0.4 and 0.6, the binary classifier will assign it as either "low" or "high". Hence, it appears more logical to define engagement levels in terms of continuous values [28]. To display the likelihood of each affective state, we used the models' last layer raw value. Figure 14 depicts qualitative results of our algorithm-generated graphs, which featured both e-learning and a traditional classroom situation, with bar graphs representing individual levels of involvement, a pie chart indicating overall student involvement, and a scatter plot highlighting when students were most engaged (or not).

5 Conclusions

The paper introduces a three-dimensional DenseNet Self-Attention neural network (3D DenseAttNet) for automatic detection of students' engagement in e-learning platforms. The self-attention module in the proposed 3D DenseAttNet model helps to extract only the relevant high-level intra-inter frame dependency features of videos obtained from the 3D DenseNet block. Evaluated on the DAiSEE dataset, the proposed neural network outperformed the previous state-of-the-art and attained recognition accuracy of 63.59% and 54.27% for four-class engagement and boredom, respectively. The paper employs class-balanced (CB) losses to address the DAiSEE data imbalance problem for engagement prediction, which has never been adequately addressed in previous research. Besides, to test the robustness of the proposed framework, we tested the pre-trained 3D DenseAttNet performance on EmotiW-EP and obtained a competitive MSE of 0.0877. In addition, a multi-label variation of the proposed 3D DenseAttNet model is used, which is an end-to-end feedback system for identifying the student's all four emotional states in e-learning and traditional classroom settings: boredom, engagement, frustration, and confusion. In future work, we will explore more efficient deep learning algorithms and advanced loss functions that can account for data imbalances and multi-class multi-label emotion categorization in DAiSEE.

Acknowledgements The authors would like to thank the Director, CSIR-CEERI, Pilani, India for supporting and encouraging research activities at CSIR-CEERI, Pilani. We also thank Kashish Sapra of CSIR-CEERI, Pilani for proofreading.

Declarations

Conflict of Interest The authors declare that we have no conflict of interest.

References

1. Mahmood S (2021) Instructional strategies for online teaching in covid-19 pandemic. Human Behav Emerg Technol 3(1):199–203
2. Dias SB, Hadjileontiadou SJ, Diniz J, Hadjileontiadis LJ (2020) Deeplms: a deep learning predictive model for supporting online learning in the covid-19 era. Sci Rep 10(1):1–17
3. Singh V, Thurman A (2019) How many ways can we define online learning? a systematic literature review of definitions of online learning (1988-2018). Am J Dist Educ 33(4):289–306
4. Adnan M, Anwar K (2020) Online learning amid the covid-19 pandemic: Students' perspectives. Online Submiss 2(1):45–51
5. Dhawan S (2020) Online learning: A panacea in the time of covid-19 crisis. J Educ Technol Syst 49(1):5–22
6. Lan M, Hew KF (2020) Examining learning engagement in moocs: A self-determination theoretical perspective using mixed method. Int J Educ Technol Higher Educ 17(1):1–24
7. Kuzilek J, Hlosta M, Herrmannova D, Zdrahal Z, Vaclavek J, Wolff A (2015) Ou analyse: analysing at-risk students at the open university. Learn Anal Rev:1–16
8. Dewan MAA, Murshed M, Lin F (2019) Engagement detection in online learning: a review. Smart Learn Environ 6(1):1–20
9. Hussain M, Zhu W, Zhang W, Abidi SMR (2018) Student engagement predictions in an e-learning system and their impact on student course assessment scores. Comput Intell Neurosci
10. Pietarinen J, Soini T, Pyhältö K (2014) Students' emotional and cognitive engagement as the determinants of well-being and achievement in school. Int J Educ Res 67:40–51
11. Pilotti M, Anderson S, Hardy P, Murphy P, Vincent P (2017) Factors related to cognitive, emotional, and behavioral engagement in the online asynchronous classroom. Int J Teach Learn Higher Educ 29(1):145–153
12. Craig S, Graesser A, Sullins J, Gholson B (2004) Affect and learning: an exploratory look into the role of affect in learning with autotutor. J Educ Media 29(3):241–250
13. Jung Y, Lee J (2018) Learning engagement and persistence in massive open online courses (moocs). Comput Educ 122:9–22
14. Kushwaha RC, Singhal A, Chaurasia PK (2015) Study of students' performance in learning management system. Int J Contempor Res Comput Sci Technol (IJCRCST) 1(6):213–217
15. Wang M-T, Willett JB, Eccles JS (2011) The assessment of school engagement: Examining dimensionality and measurement invariance by gender and race/ethnicity. J Sch Psychol 49(4):465–480
16. Bartlett MS, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2005) Recognizing facial expression: machine learning and application to spontaneous behavior. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol 2. IEEE, pp 568–573
17. Guo Y, Tao D, Yu J, Xiong H, Li Y, Tao D (2016) Deep neural networks with relativity learning for facial expression recognition. In: 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, pp 1–6
18. Saurav S, Saini R, Singh S (2021) Emnet: a deep integrated convolutional neural network for facial emotion recognition in the wild. Appl Intell:1–28
19. Yu Z, Zhang C (2015) Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp 435–442
20. Calvo RA, D'Mello S (2010) Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Trans Affect Comput 1(1):18–37

21. Gupta A, D'Cunha A, Awasthi K, Balasubramanian V (2016) Daisee: Towards user engagement recognition in the wild. arXiv:1609.01885
22. Whitehill J, Serpell Z, Foster A, Lin Y-C, Pearson B, Bartlett M, Movellan J (2011) Towards an optimal affect-sensitive instructional system of cognitive skills. In: CVPR 2011 WORKSHOPS. IEEE, pp 20–25
23. Grafsgaard J, Wiggins JB, Boyer KE, Wiebe EN, Lester J (2013) Automatically recognizing facial expression: Predicting engagement and frustration. In: Educational Data Mining 2013
24. Bosch N, D'Mello S, Baker R, Ocumpaugh J, Shute V, Ventura M, Wang L, Zhao W (2015) Automatic detection of learning-centered affective states in the wild. In: Proceedings of the 20th international conference on intelligent user interfaces, pp 379–388
25. Kamath A, Biswas A, Balasubramanian V (2016) A crowdsourced approach to student engagement recognition in e-learning environments. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp 1–9
26. Monkaresi H, Bosch N, Calvo RA, D'Mello SK (2016) Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Trans Affect Comput 8(1):15–28
27. Huang T, Mei Y, Zhang H, Liu S, Yang H (2019) Fine-grained engagement recognition in online learning environment. In: 2019 IEEE 9th international conference on electronics information and emergency communication (ICEIEC). IEEE, pp 338–341
28. Liao J, Liang Y, Pan J (2021) Deep facial spatiotemporal network for engagement prediction in online learning. Appl Intell:1–13
29. Wang Y, Kotha A, Hong P, Qiu M (2020) Automated student engagement monitoring and evaluation during learning in the wild. In: 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom). IEEE, pp 270–275
30. Zhang H, Xiao X, Huang T, Liu S, Xia Y, Li J (2019) An novel end-to-end network for automatic student engagement recognition. In: 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC). IEEE, pp 342–345
31. Zhang S, Pan X, Cui Y, Zhao X, Liu L (2019) Learning affective video features for facial expression recognition via hybrid deep learning. IEEE Access 7:32297–32304
32. Saurav S, Gidde P, Saini R, Singh S (2021) Dual integrated convolutional neural network for real-time facial expression recognition in the wild. Vis Comput:1–14
33. Yang J, Wang K, Peng X, Qiao Y (2018) Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction. In: Proceedings of the 20th ACM international conference on multimodal interaction, pp 594–598
34. Murshed M, Dewan MAA, Lin F, Wen D (2019) Engagement detection in e-learning environments using convolutional neural networks. In: 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
35. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4489–4497
36. Uemura T, Näppi JJ, Hironaka T, Kim H, Yoshida H (2020) Comparative performance of 3d-densenet, 3d-resnet, and 3d-vgg models in polyp detection for ct colonography. In: Medical Imaging 2020: Computer-Aided Diagnosis, vol 11314. International Society for Optics and Photonics, p 1131435
37. Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp 6546–6555
38. Ruiz J, Mahmud M, Modasshir M, Kaiser MS, Alzheimer's Disease Neuroimaging Initiative ft et al (2020) 3d densenet ensemble in 4-way classification of alzheimer's disease. In: International Conference on Brain Informatics. Springer, pp 85–96
39. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
40. Zhang Z, Sabuncu MR (2018) Generalized cross entropy loss for training deep neural networks with noisy labels. In: 32nd Conference on Neural Information Processing Systems (NeurIPS)
41. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
42. Dhall A (2019) Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In: 2019 International Conference on Multimodal Interaction, pp 546–550
43. Dhall A, Kaur A, Goecke R, Gedeon T (2018) Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp 653–656
44. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
45. Lim R, MJT Reinders T (2000) Facial landmark detection using a gabor filter representation and a genetic search algorithm. In: PROCEEDING,(SITIA'2000), GRAHA INSTITUT TEKNOLOGI SEPULUH NOPEMBER. Citeseer
46. Sathik M, Jonathan SG (2013) Effect of facial expressions on student's comprehension recognition in virtual educational environments. SpringerPlus 2(1):1–9
47. Liu P, Lin Y, Meng Z, Lu L, Deng W, Zhou JT, Yang Y (2021) Point adversarial self-mining: A simple method for facial expression recognition. IEEE Transactions on Cybernetics
48. Tonguç G, Ozkara BO (2020) Automatic recognition of student emotions from facial expressions during a lecture. Comput Educ 148:103797
49. Bhardwaj P, Gupta PK, Panwar H, Siddiqui MK, Morales-Menendez R, Bhaik A (2021) Application of deep learning on student engagement in e-learning environments. Comput Electr Eng 93:107277
50. Pan M, Wang J, Luo Z (2018) Modelling study on learning affects for classroom teaching/learning auto-evaluation. Science 6(3):81–86
51. Thomas C (2018) Multimodal teaching and learning analytics for classroom and online educational settings. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp 542–545
52. El Kerdawy M, El Halaby M, Hassan A, Maher M, Fayed H, Shawky D, Badawi A (2020) The automatic detection of cognition using eeg and facial expressions. Sensors 20(12):3516
53. Hu X, Chen J, Wang F, Zhang D (2019) Ten challenges for eeg-based affective computing. Brain Sci Adv 5(1):1–20
54. Khedher AB, Jraidi I, Frasson C et al (2019) Tracking students' mental engagement using eeg signals during an interaction with a virtual learning environment. J Intell Learn Syst Appl 11(01):1
55. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In:

Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
56. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634
57. Geng L, Xu M, Wei Z, Zhou X (2019) Learning deep spatiotemporal feature for engagement recognition of online courses. In: 2019 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, pp 442–447
58. Niu X, Han H, Zeng J, Sun X, Shan S, Huang Y, Yang S, Chen X (2018) Automatic engagement prediction with gap feature. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp 599–603
59. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823
60. Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, He X (2018) Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1316–1324
61. Hoogi A, Wilcox B, Gupta Y, Rubin DL (2019) Self-attention capsule networks for object classification. arXiv:1904.12483
62. Li M, Hsu W, Xie X, Cong J, Gao W (2020) Sacnn: Self-attention convolutional neural network for low-dose ct denoising with self-supervised perceptual loss network. IEEE Trans Med Imaging 39(7):2289–2301
63. Zhang H, Goodfellow I, Metaxas D, Odena A (2019) Self-attention generative adversarial networks. In: International conference on machine learning. PMLR, pp 7354–7363
64. Zhang X, Han L, Zhu W, Sun L, Zhang D (2021) An explainable 3d residual self-attention deep neural network for joint atrophy localization and alzheimer's disease diagnosis using structural mri. IEEE Journal of Biomedical and Health Informatics
65. Drummond C, Holte RC et al (2003) C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In: Workshop on learning from imbalanced datasets II, vol 11. Citeseer, pp 1–8
66. Huang C, Li Y, Loy CC, Tang X (2016) Learning deep representation for imbalanced classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5375–5384
67. Khan SH, Hayat M, Bennamoun M, Sohel FA, Togneri R (2017) Cost-sensitive learning of deep feature representations from imbalanced data. IEEE Trans Neural Netw Learn Syst 29(8):3573–3587
68. Cui Y, Jia M, Lin T-Y, Song Y, Belongie S (2019) Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9268–9277
69. Wang L, Wang C, Sun Z, Cheng S, Guo L (2020) Class balanced loss for image classification. IEEE Access 8:81142–81153
70. Saurav S, Saini R, Singh S (2021) A dual-stream fused neural network for fall detection in multi-camera and 360° videos. Neural Comput Appl:1–28
71. Grandini M, Bagli E, Visani G (2020) Metrics for multi-class classification: an overview. arXiv:2008.05756
72. Bottou L (2012) Stochastic gradient descent tricks. In: Neural networks: Tricks of the trade. Springer, pp 421–436

Naval Kishore Mehta is currently an Integrated Dual Degree Ph.D (IDDP) student at Academy of Scientific and Innovative Research (AcSIR), CSIR - Central Electronics Engineering Research Institute (CSIR-CEERI) Campus, Pilani, India. His research interests include computer vision, machine learning, and deep learning, with a focus on activity recognition in video and its hardware implementation under resource constraints.

Shyam Sunder Prasad is enrolled in the integrated M.Tech. Ph.D program under Academy of Scientific and Innovative Research (AcSIR) at CSIR - Central Electronics Engineering Research Institute (CSIR-CEERI), Pilani, India since 2015. He completed his M.Tech. in year 2017 and is currently pursuing his Ph.D from the Intelligent Systems Group. His research areas include computer vision, deep learning and spiking neural networks.

Sumeet Saurav is working as a Senior Scientist in Intelligent Systems Group at CSIR - Central Electronics Engineering Research Institute (CSIR-CEERI), Pilani, India. His research interest includes computer vision, machine learning, deep learning architectures for vision-based applications, and embedded real-time implementation of computer vision algorithms. He is also pursuing his Ph.D in the related areas.

Ravi Saini is working as a Principal Scientist and is currently heading the Intelligent Systems Group at CSIR - Central Electronics Engineering Research Institute (CSIR-CEERI), Pilani, India. His research interest includes hardware accelerators for computer vision, machine learning and deep learning based applications, ASIC and ASIP Design, HDLs, and FPGA Prototyping.

Sanjay Singh Sanjay Singh received the B.Sc. (Electronics and Computer Science), M.Sc. (Electronic Science), M.Tech. (Microelectronics and VLSI Design), and Ph.D. Degrees from Kurukshetra University, Kurukshetra, India. He joined CSIR - Central Electronics Engineering Research Institute (CSIR-CEERI), Pilani as Scientist Fellow in 2009. Now he is Principal Scientist in Intelligent Systems Group at CSIR-CEERI, Pilani, and is also an Associate Professor in Academy of Scientific and Innovative Research (AcSIR). His research interests include Image Processing, Computer Vision, and Artificial Intelligence.