Social Network Analysis and Mining (2024) 14:43
https://doi.org/10.1007/s13278-023-01181-x

ORIGINAL ARTICLE

Online learners' engagement detection via facial emotion recognition in online learning context using hybrid classification model

Rama Bhadra Rao Maddu[1] · **S. Murugappan**[1]

Received: 1 August 2023 / Accepted: 2 December 2023

Abstract

Writing, reading, viewing video lectures, completing online examinations, and attending online meetings are all activities that students participate in through the internet. While participating in these educational activities, they demonstrate various degrees of interest, including boredom, aggravation, delight, indifference, confusion, and learning gain. Online educators must accurately and efficiently monitor the degree of engagement of their online learners with the goal of giving focused pedagogical assistance to them through interventions. The objective of this paper is to propose a novel students engagement prediction model for online learners based on facial emotion, which will include four basic phases: (a) preprocessing, (b) feature extraction, (c) emotion recognition, and (d) student engagement prediction. The preprocessing step is the first phase in which the Face detection process is followed. Following the preprocessing step, the feature extraction phase proceeds with the extraction of the Improved Active Appearance Model (AAM), Shape Local Binary Texture (SLBT), Global Binary Pattern (GBP), and ResNet features. The retrieved characteristics are subsequently subjected to emotion recognition via the Hybrid Classification model, which incorporates models including Improved Deep Belief Network (IDBN) and Convolutional Neural Network (CNN). The student's involvement or engagement is identified based on the emotions recognized, as well as the way they performed via the enhanced entropy-based process**.** The execution of the suggested hybrid IDBN + CNN model is evaluated over the extant methods like DBN, SVM, CNN, LSTM-CNN, LSTM, and RF under various measures for two datasets. The hybrid model had the greatest accuracy of 0.95 at a learning percentage of 80% for the CK+ dataset. Also, the hybrid model has a higher sensitivity of 60% for FER-2013 datasets.

Keywords Student engagement prediction · Online learning context · Improved AAM · Emotion recognition phase · Hybrid classifier

Abbreviations

CNN	Convolutional neural network
DFSTN	Deep facial spatiotemporal network
SVM	Support vector machines
LSTM	Long short-term memory
GALN	LSTM network with global attention
ML	Machine learning
FER	Facial Expression Recognition
DL	Deep learning
GPU	Graphics processing unit
DBN	Deep Belief Network
KNN	K-nearest neighbor
FS	Face-sensitive
EI	Engagement index
PCA	Principal component analysis
AAM	Active Appearance Model
GBP	Global Binary Pattern
IDBN	Improved Deep Belief Network
DT	Decision tree
SLBT	Shape Local Binary Texture
EEG	Electroencephalogram
GPU	Graphics processing unit
DNN	Deep neural network
AI	Artificial intelligence
VGG	Visual Geometry Group

✉ Rama Bhadra Rao Maddu
ramamaddu7288@gmail.com

[1] Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalainagar, Chennai, Tamilnadu 608002, India

Published online: 21 February 2024

1 Introduction

Facial emotion recognition is defined as the process of anticipating bodily emotions such as facial expressions or brain impulses, and it is distinguished by qualities that allow us to distinguish one emotion from the others (Buono et al. 2022). As a result, the current research focuses on the automated identification of facial expressions in online learners. In a real-time scenario, to determine the diverse emotions, the learners' facial expressions are evaluated on the e-learning platform. Facial expressions are the movements of physical muscles caused by emotional impulses, including lip curling, brow raising, or brow wrinkling. A lot of information regarding online learners' emotional states may be determined by automatically watching the change in facial expressions (Savchenko and Makarov 2022).

The face intuitively contains a vast number of emotional information and has been directly connected to one's perceived involvement (Hassouneh et al. 2020). Recently, the study of physiological data acquired by sensors such as galvanic skin reaction, heart rate, and electroencephalogram (EEG) has been examined for assessing engagement. These sensors, however, were more obtrusive since users had to physically wear them, and they were also more costly than a camera (Ngai et al. 2022). Despite several research efforts for Facial Expression Recognition (FER) throughout the last decades, identifying learners' emotional states using facial expressions remained a difficult challenge in practical applications. Facial expressions come in a variety of intensities, ranging from modest micro-expressions to dramatic emotions (Hung et al. 2019) (Ninaus et al. 2019). The existing system detects the perceived and non-perceived behavior states which makes online learners more frustrated and causes dropout rates in online courses (Dewan, et al. 2018).

AI techniques including ML techniques and deep architectures prove their betterment in distinguishing the aspects based on the proper training that is employed to predict the students' mood at the present state (Mehta et al. 2022) (Ashwin and Guddeti 2020a). SVM, KNN, DT learning, association rule learning, rule-based ML, and others are the classic ML techniques that categorize the provided input data using handcrafted/predefined features (Zhang et al. 2020). However, the amount of processing power required for an NN is determined by the size of your data as well as the depth and complexity of the network. CNN design makes extensive use of computer resources, such as graphics processing units (GPU) (Yuan 2022) (Vidyadhari and Mishra 2020) (Ding and Xing 2022) to overcome the problem defined in Ml techniques. Also, the combinations of different DL statistics ensure precise collaboration in predicting things. This paper proposes the usefulness of deep learning methods in spotting learners' engagement through facial expression study. Our objective is to determine how engaged online students are regarded to be by an outsider. Since professors in traditional classrooms rely on perceived involvement to implement their teaching methods, automated perceived engagement detection is probably advantageous for giving students support to online education. Thus, this work proposes the deep hybrid model for emotion recognition and detecting the online learner's engagement.

The contributions of the work are as follows:

- Develops a hybrid classification model (CNN + IDBN) for recognizing facial expressions.
- Introduces Improved Active Appearance Model (AAM), Shape Local Binary Texture (SLBT), Global Binary Pattern (GBP), and ResNet features for feature extraction.
- Proposes improved entropy for the detection of student engagement in online courses.

The following of this work is planned: The second section depicts a review of student engagement prediction. Section 3 offers an overview of the given study on student engagement prediction for online learners. Section 4 describes the preprocessing and feature extraction steps in the student engagement process. Section 5 depicts the emotion recognition phase as well as the student performance prediction phase. Sections 6 and 7 denote the results and conclusion.

2 Literature review

2.1 Related works

Ashwin and Guddeti (2020b) presented a unique hybrid CNN architecture in 2020 for evaluating students' emotional states in a classroom setting. This suggested architecture includes two separate techniques: the initial one (CNN-1) was meant to assess the one student's emotional state in a single picture frame, while the subsequent one (CNN-2) was intended to evaluate the affective states of multiple learners in one picture frame. The suggested architecture assesses the students' emotional states using hand gestures, body postures, and facial expressions.

Using a hybrid deep neural network, Zhu and Chen (2020) in 2020 suggested a novel dual-modality Identifying face expression in e-learning through the learning of spatiotemporal feature representations. In our work, sample expression states were used for expression recognition rather than facial expression class information. A hybrid DNN was used to learn spatial–temporal appearance feature representations and spatial–temporal geometrical feature

representations. To detect face expressions, dual-modality feature fusion representations were utilized.

Cabada et al. (2020) will publish their findings in 2020. The author presented a method for optimizing the hyperparameters of a CNN, which is used to determine a person's emotional state, using evolutionary algorithms. They also exhibited the optimized network in a mobile phone-based intelligent teaching method. The CNN was trained on a PC equipped with a GPU before being delivered to a mobile device.

Schoneveld et al. (2021) developed a novel deep learning-based method for recognizing audio-visual emotions. The approach makes use of fresh advances in DL, such as high-performance deep architectures as well as knowledge distillation. The features of deep representations of the visual modalities and auditory were integrated using a model-level fusion approach.

Gupta et al. (2022) created a deep learning-based system that uses facial expressions to determine online learners' real-time involvement. This was accomplished by studying the students' facial expressions throughout the online learning session to identify their moods. The EI was calculated using the facial expression detection information to forecast two engagement states: "engaged" and "disengaged." Various DL algorithms, including VGG-19, ResNet-50, and Inception-V3, were researched and evaluated to determine the most excellent predictive classification algorithm.

Said and Barr (2021) presented a face-sensitive FS-CNN for recognizing human emotions. The adopted FS-CNN on large-scale images was determined for emotion recognition to identify faces followed by a facial landmark analysis to forecast emotions. The FS-CNN was created using CNN and patch cropping. The initial stage was to recognize and crop faces for subsequent processing in high-resolution photographs. The employment of a CNN based on landmark analytics to predict facial expressions was then used on pyramid pictures to process scale invariance.

Liao et al. (2021) published their findings in 2020. The DFSTN, a unique model to anticipate participation, was introduced by the author. The model consists of two procedures: pre-trained SE-ResNet-50 for deriving face spatial characteristics and GALN for constructing an attentional hidden state. As the performance metric changed, so did the model's training approach. By gathering facial spatial and temporal information, the DFSTN can detect fine-grained engaged states and increase engagement prediction ability.

Xu et al. (2020) published their findings in 2020. The author suggested a model for emotion-sensitive cognitive state analysis that non-intrusively evaluates learners' attention based on head position and mood according to facial expression. The evaluated head position and landmarks are utilized to identify the learner's visual center of focus. The emotions of the students were then assessed using their facial expressions.

Lasri and Riadsolh (2023) developed a method in 2023, for detecting the facial expressions of deaf and hard-of-hearing pupils to determine their level of participation ('highly engaged,' 'nominally engaged,' and 'not engaged'), using a deep CNN (DCNN) model and transfer learning (TL) technique. On the KDEF dataset and the JAFFE dataset, a pre-trained VGG-16 model is used and modified.

Benabbes et al. (2023) established a BiLSTM approach in 2023, with FastText word embedding employed in this work to identify the emotions of participants in forum discussions. Using an unsupervised clustering method built on the new dataset, the learners were then divided into groups according to their level of involvement. The performance of each supervised classification algorithm was evaluated after training using a variety of accuracy measures and cross-validation techniques. The decision tree rule technique was more accurate than the existing methods, with an AUC score of 0.97 and a 98% accuracy rate. Table 1 shows the review on the existing models.

2.2 Review

Because of a number of difficulties, including differences in face shapes from person to person, difficulty recognizing dynamic facial features, poor picture quality, and others, the computer vision community views the identification of human emotion based on facial expression as a problematic topic. The main challenge is finding a way to derive those expressions from high-resolution photographs where the face only takes up 10% of the whole image. The intricacy of visual backdrops, interclass and intraclass variance, differences in point of view, and many other aspects make developing a strong emotion recognition system a difficult task. Learning non-invasive cognitive state analysis is a difficult endeavor, and the connected problems of emotion prediction and attention are 2 key research issues from time series data. For gaining a thorough understanding of the relationship between learning and affect, an examination of the larger collection of facial action units monitored by CERT will be required. To improve the accuracy of the predictions acquired for the samples, enhancement approaches that can optimize the prediction model and improve the findings, which can then boost the accuracy levels gained via the analysis of the Biometric features, might be used.

2.3 Objectives

To obtain accurate recognition results based on performance measures.

To achieve better stability in online learners' engagement detection.

Table 1 Features and challenges of traditional schemes

Author	Method	Merits	Demerits
Ashwin and Guddeti (2020b)	Hybrid CNN	Maximal accuracy was obtained	The drawback is it doesn't analyze the cognitive engagement of students
Zhu and Chen (2020)	Hybrid DNN	Accuracy was greater	It requires a very large amount of data
Cabada et al. (2020)	CNN	Higher accuracy was acquired	Need to test the optimized CNN with opinion mining and sentiment analysis
Schoneveld et al. (2021)	DL-based approach	Accuracy was maximized	Continuous emotion encoding is difficult
Gupta et al. (2022)	DL-based approach	Greater accuracy was attained	Train the test of the model with a large dataset
Said and Barr (2021)	FS-CNN	Precision was greater	Overfitting
Liao et al. (2021)	DFSTN	MSE was minimized and the accuracy was maximized	The problems that occurred in this model were data deficiencies and data imbalances
Xu et al. (2020)	Emotion-sensitive learning cognitive state analysis	A higher correctness rate was attained	For learners' cognitive states, it is necessary to combine the multimodal aspects
Lasri and Riadsolh (2023)	DCNN	Better accuracy	More amount of resources is required
Benabbes et al. (2023)	BiLSTM	Better performance in facial recognition	A large amount of data is required

To attain great precision in the facial emotion recognition system.

3 Overall description of presented work on student engagement prediction

Improving learner engagement with instructional activities is a critical challenge in online learning. It is well accepted that higher productivity and learning gain is connected to learners' engagement and their emotions. Some research claim that learner engagement is flexible and that it may be increased through the use of effective pedagogical interventions, instructional strategies, and feedback. However, detecting learner participation via emotions has become crucial in online education, and a lot of studies reveal that the usage of AI technologies could manage the capturing of accurate emotions.

This study proposes a novel students performance prediction based on facial emotion with four important phases: (a) preprocessing, (b) feature extraction, (c) emotion recognition, and d) student performance prediction. The preprocessing phase is the initial stage, in which the Face detection process takes place. The face detection is carried out by Viola–Jones model. The Improved AAM, SLBT, GBP and ResNet-based features are extracted after the preprocessing step. The hybrid recognition model is developed to be trained with the extracted feature set. The suggested hybrid classification model is a mixture of CNN and IDBN. The student's engagement will be detected based on an improved entropy process, which estimates the performance of students by facial emotion recognition. Figure 1 displays the architecture of the suggested hybrid IDBN-CNN model.

4 Student engagement process: preprocessing and feature extraction

In addition, the input image Γ is given to preprocessing stages, where the face detection process is carried out by the Viola–Jones model.

4.1 Preprocessing

The initial stage after data collection that will be utilized to train the classifier in a FER system is preprocessing. In order to prevent fluctuations in face expressions and emotions from impairing the recognition process, some of these preprocessing approaches are utilized for feature extraction, while others are used to normalize the features. Here, the input image Γ is preprocessed by a face detection approach and the Viola–Jones model.

Face identification is challenging because there are so many factors involved in picture emergence, such as image orientation, facial expression, position changes, lighting conditions, and occlusion. In this design, the Viola–Jones face detection was used.

Viola–Jones model (Barnouti et al. 2016): The Viola–Jones object identification system, widely considered the first to deliver useful object recognition in real

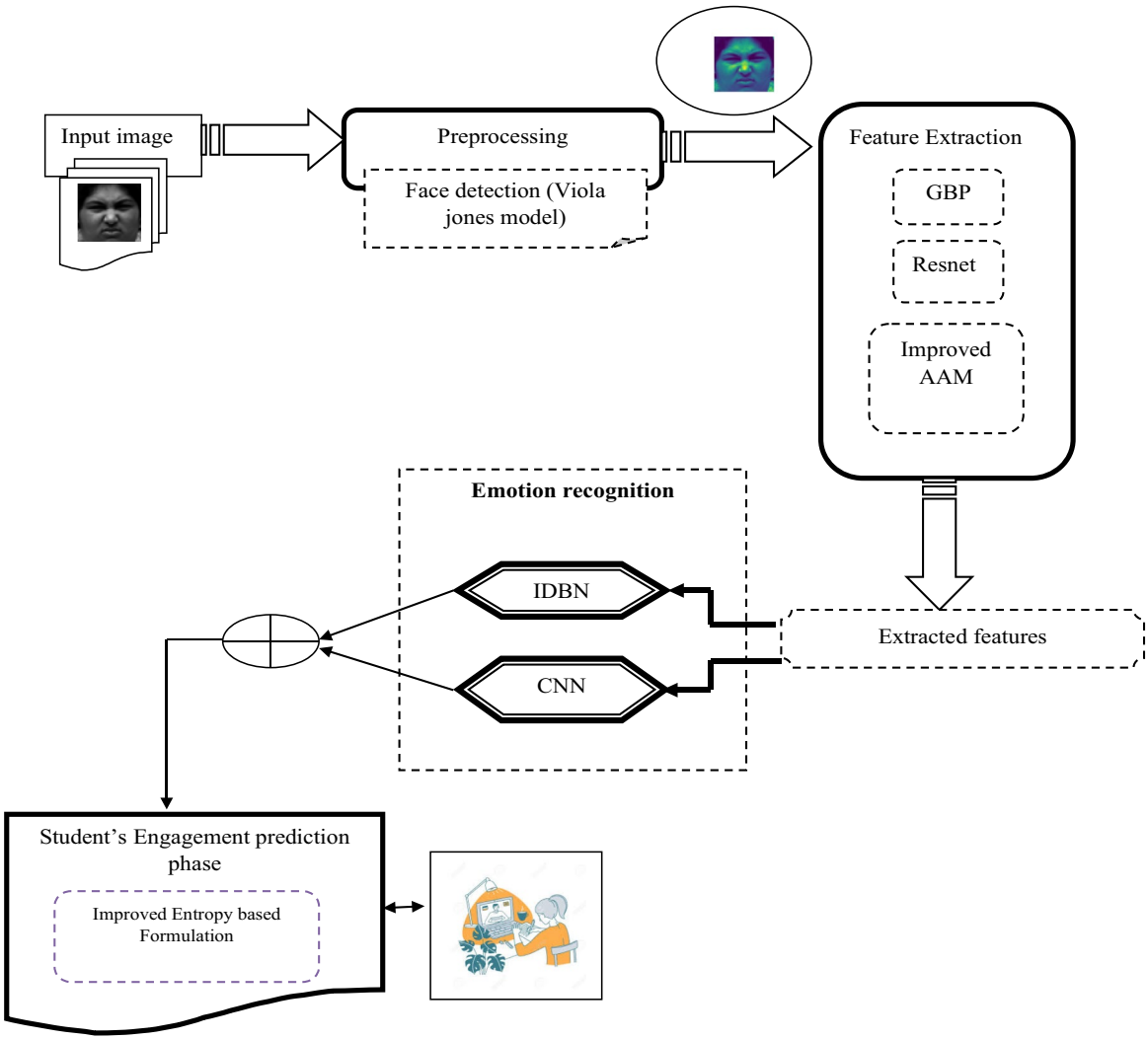

Fig. 1 Architecture of the proposed hybrid IDBN-CNN model

time, had the greatest influence in the 2000s. Viola–Jones needs full eyesight in the front upright face. With this technique, the features of a human face are searched for in a window-sized input image. The type of window in the image is recognized as a face when other qualities are discovered. The window will be extended and the process repeated to obtain different face sizes. The approach is applied independently of various scales for each window scale. Each window should be evaluated and put through a number of stages in order to reduce the number of features. Early levels have fewer attributes to verify, so later levels contain more features and are simpler to pass. Each level evaluates the features; if the total value is less than a certain threshold, the stage fails and the window is not identified as a face. Furthermore, the preprocessed results are designated as Γ_{pre}.

4.2 Feature extraction

By extracting features from the input data, feature extraction enhances the suggested models' accuracy. By removing unnecessary data, it reduces the measurements of the data. It increases the pace of training and inference. The CNN neural network derives the features of the input image, and various neural network classifies additional image features.

Furthermore, at this step, the following input Γ_{pre} is used to extract the features that include:

- Improved AAM
- ResNet
- SLBT
- GBP

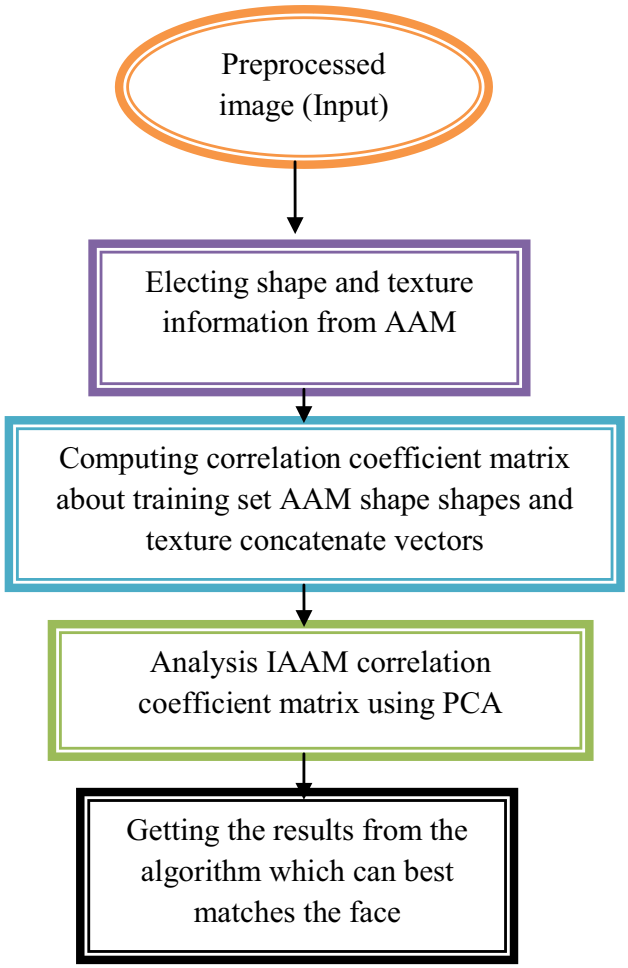

Fig. 2 Framework of IAAM scheme

(i) Improved AAM (IAAM): AAM (Iqtait et al. 2018): The AAM model originated from the ASM model. Triangles whose vertices are the ASM landmark points are used to divide the entire face into smaller sections. In order to analyze using PCA once more, the active appearance model combines shape and texture PCA parameters for one categorize of vectors and a weight vector. Figure 2 illustrates the framework of the IAAM scheme.

The algorithm of the AAM model includes three phases: (i) Shape and texture vectors are jointly combined in the similar vector by connecting them to every AAM in the training set: $\hat{C}_1 = \left(\hat{T}_1, Y_1\right)^{\mathrm{T}}$ where $\hat{T}_1$ denotes pixel texture vector, $\left(Y_1\right)$ indicates the location of a landmark. (ii) Calculating the training set's connected shape and texture vectors' correlation coefficient matrix. (iii) Using PCA to analyze the correlation coefficient matrix. Conventionally, the PCA-based correlation is determined in Eq. (1).

$$\varphi_1 = \left(Y_1 - \mu\right) \tag{1}$$

In order to reduce the high dimension in image pixels, the IAAM is used. As per the IAAM concept, find the coefficient of variation for Y_1; therefore, the final equation is specified as Eq. (2)

$$\varphi_1 = \left(\frac{\sigma\left(Y_1\right) - \mu^2\left(Y_1\right)}{\mu\left(Y_1\right)}\right) \tag{2}$$

where

$$\mathrm{CV} = \frac{\sigma}{\mu} \tag{3}$$

Here, μ is the mean of $\left(Y_1\right)$, and σ denotes the standard deviation of $\left(Y_1\right)$. AAM feature is denoted as Γ_{IAAM} In Eq. (3).

(ii) ResNet:

ResNet-based feature $\Gamma_{\mathrm{Res\ -\ net}}$ is obtained from the input Γ_{pre}. ResNet (Mhapsekar et al. 2015) architecture involved the insertion of shortcut connections in turning a plain network into its residual network counterpart. In this instance, the convolutional networks had 33 filters, while the simple network was taken by Visual Geometry Group (VGG) neural networks (VGG-16 and VGG-19). ResNets, however, are simpler and contain fewer filters than VGGNets. The performance of the 34-layer ResNet is 3.6 billion FLOPs, as opposed to 1.8 billion FLOPs for the smaller 18-layer ResNets. It also followed two simple design principles: the layers had the same number of filters for the same output feature map size, and the number of filters increased for smaller feature map sizes in order to preserve the time complexity per layer. There were 34 weighted layers in it. From this architecture, the ResNet features are considered as $\Gamma_{\mathrm{Res}-net}$.

(iii) SLBT: SLBT (Lakshmiprabha and Majumder 2012) functions combine shape and texture information.

Let $\Gamma_{\mathrm{pre}} = \left[\Gamma_{\mathrm{pre}\ 1}, \Gamma_{\mathrm{pre}\ 2}, \ldots, \Gamma_{\mathrm{pre}\ M}\right]$ describe M training set images with $_{Z = [Z_1, Z_2, \ldots\ldots, Z_M]}$ as landmarks in its shape. Every shape vector Z in the training set can be modeled using Eq. (4).

$$\begin{aligned} Z &\approx \overline{Z} + V_{\mathrm{s}} A_{\mathrm{s}}, \\ A_{\mathrm{s}} &= V_{\mathrm{s}}^{\mathrm{T}}\left(Z - \overline{Z}\right) \end{aligned} \tag{4}$$

where $V_{\mathrm{s}}\rightarrow$ greatest eigenvalues' eigenvectors, $\overline{Z}\rightarrow$ shape of mean, and $A_{\mathrm{s}}\rightarrow$ model of shape or parameters of weights. Take a 3×3 window with center pixel $\left(u_{\mathrm{c}}, v_{\mathrm{c}}\right)$ intensity number be g_{c} and local texture as $Q = q\left(g_i\right)$ where $g_i(i = 0, 1, 2, 3, 4, 5, 6, 7)$ correlates to the gray numbers of the eight pixels around it. The center pixel of LBP pattern gc is described using Eq. (6). The function $a(n)$ is defined Eq. (5).

$$\begin{cases} 1, & n > 0 \\ 0, & n \le 0 \end{cases} = a(n) \tag{5}$$

$$\sum_{i=0}^{7} a\left(g_i - g_c\right)2^i = \mathrm{LBP}\left(u_c, v_c\right) \tag{6}$$

Let $K = \left[K_1, K_2, ..., K_M\right]$. The texture modeling is executed via PCA as in Eq. (7)

$$A_t = V_t^{\mathrm{T}}\left(K - \overline{K}\right) \tag{7}$$

The whole shape and texture parameter vector is computed by Eq. (8). By running PCA on the connected parameter vector as shown in Eq. (9), the parameter of shape texture controlling texture, global shape, and local shape may be created.

$$\begin{pmatrix} W_s A_s \\ A_t \end{pmatrix} = A_{st} \tag{8}$$

$$V_s^t\left(A_{st} - \overline{A_{sr}}\right) = C \tag{9}$$

where $C \rightarrow$ shape texture parameter, $V_{st} \rightarrow$ eigenvectors and $\overline{A_{st}} \rightarrow$ mean vector. The obtaining SLBT feature is denoted as Γ_{SLBT}.

(iv) GBP: The preprocessed image Γ_{pre} is converted to a binary image B to generate the GBP (Sivri and Kalkan 2013), which is a series of bit strings for each direction of a binary image and interprets them as binary integers to provide a global descriptor. GBP of a row r of a binary image B is described in Eq. (10), its most basic form as follows.

$$\Gamma_d^{\mathrm{GBP}}(r) = \sum_{e=0}^{E-e_m^d} B\left(r, e + e_m^d\right)2^{-e} + \sum_{e=0}^{E-e_m^d} B\left(r, e_m^d + e\right)2^{-je} \tag{10}$$

where$E \rightarrow$ the number of columns and $e_m^d \rightarrow$ center of mass along the horizontal direction. Γ_{GBP}^x is defined in Eq. (11):

$$\Gamma_{GBP}^x(e) = \sum_{r=0}^{R-e_m^x} B\left(r + e_m^x, e\right)2^{-r} + \sum_{r=0}^{e_m^x} B\left(e_m^x - r, e\right)2^{-jr} \tag{11}$$

where $e_m^x \rightarrow$ the center of mass is along the vertical direction and $R \rightarrow$ number of rows. Γ_{GBP}^{θ} for the image B is then defined in Eq. (12):

$$\Gamma_{GBP}^{\theta}(k) = \sum_{p \in B^-} \delta\left(O_\gamma - k\right)B(p)2^{-jb^p} + \sum_{p \in B^+} \delta\left(O_\gamma - k\right)B(p)2^{-jb^p} \tag{12}$$

where $b^p \rightarrow$ the point-to-line distance between the pixel p and the line $h_\theta^{p_0}$, B^- and $B^+ \rightarrow$ set of pixels on the negative and positive side of the center of mass, $y^p \rightarrow$ projection of the pixel p onto the line $h_\theta^{p_0}$, $O \rightarrow$ desired length of the GBP_θ descriptor, $\gamma \rightarrow \frac{y^p - \min\left(y^{p\hat{y}}\right)}{\max\left(y^{p\hat{y}}\right) - \min\left(y^{p\hat{y}}\right)}$), and $\delta(.) \rightarrow$ Kronecker delta defined in Eq. (13):

$$\delta(y) = \begin{cases} 1, & \text{if } y = 0 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

The overall features Γ_{FE} are evaluated in Eq. (14).

$$\Gamma_{FE} = \left[\Gamma_{IAAM} \;\; \Gamma_{Res-net} \; \Gamma_{SLBT} \; \Gamma_{GBP}\right] \tag{14}$$

5 Emotion recognition phase and student engagement prediction phase

5.1 Emotion recognition phase

Hybrid classifier (***HC***): Different machine learning (and decision-making) models are integrated or combined in hybrid machine learning systems. Each machine learning technique operates differently and targets a distinct area of the problem (input) space, typically with a different feature set. Consequently, combining or integrating these techniques typically yields better results than utilizing each machine learning or decision-making model independently. Hybrid models can take use of the various generalization methods of basic models while mitigating their unique shortcomings. In this work, an HC method is created by combining the IDBN and CNN schemes to categorize the emotions according to the extracted feature set Γ_{FE}.

There are two primary reasons to use IDBN for engagement detection. Since pixel-based classification is not required in facial expression detection applications, IDBNs have been demonstrated to be effective in these settings. In the past several years, IDBNs have been effectively applied to numerous image classification applications. They are capable of automatically learning key data features from tiny samples without the need for artificial selection. Also, CNN's classification capabilities are used in the sentiment analysis operation. The process is as follows: The obtained Γ_{FE} is fed into IDBN and CNN classifiers. The output of both the IDBN and CNN classifiers is averaged to generate the recognized result. The emotions categorized are 0 = Angry, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Sad, 5 = Surprise, 6 = Neutral. This suggested method can enhance the way that students learn online, whether they are reading, writing, watching tutorial videos, or taking part in virtual meetings. Additionally, this aids in accurately identifying the emotions in students' facial expressions. Their expressions help to know about the interest of students in online courses

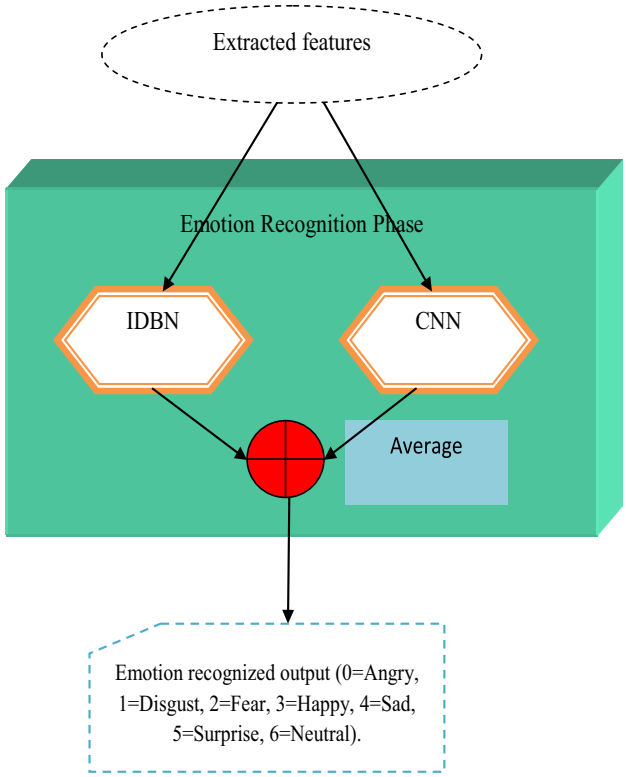

Fig. 3 Emotion recognition phase via hybrid classifier

and reduce the dropout rates. The emotion recognition step is shown in Fig. 3.

IDBN: IDBN is given the extracted features Γ_{FE} as a source. This structure comprises several levels of DBN, and each layer contains both visible and buried neurons (Wang et al. 2016).

In Boltzmann networks, Stochastic neurons produce probabilistic outcome (O). The probability function $prob(\eta)$ for the binary output (o) is given quantitatively in Eq. (15). As the pseudo-temperature ψ schemes 0 is described in Eq. (16), the stochastic method becomes deterministic.

$$prob(\eta) = \frac{1}{1 + e^{\frac{-\eta}{\psi}}} \tag{15}$$

$$o = \left\{ \begin{array}{l} 0 \text{ with } 1 - \text{prob}(\eta) \\ 1 \text{ with } \text{prob}(\eta) \end{array} \right\} \tag{16}$$

Based on the state of configuration H of the neuron, Eq. (17) determines the energy function of Boltzmann machine.

$$G(\tilde{k}) = -\sum_{\tilde{v}<\tilde{k}} H_{\tilde{v}} H_z \tilde{W}_{\tilde{v},z} - \sum_{\tilde{v}} \zeta_f H_{\tilde{v}} \tag{17}$$

Equations (18), (19), and (20) determine the RBM's hidden G and visible units U. The visible $\tilde{v}$ and hidden units'$\tilde{k}$ binary state is $U_{\tilde{v}}$ and $G_{\tilde{k}}$. The DBN's weight function $\tilde{W}$ and layers are determined by the formulae below.

$$J\left(U, \vec{G}\right) = \sum_{(\tilde{v},\tilde{k})} \tilde{W}_{\tilde{v},\tilde{k}} G_{\tilde{k}}.U_{\tilde{v}} - \sum_{f} U_{\tilde{v}} L_{\tilde{v}} - \sum_{\tilde{k}} G_{\tilde{k}} \hat{e}_{\tilde{k}} \tag{18}$$

$$J\left(U_{\tilde{v}}, \vec{G}\right) = \sum_{\tilde{k}} \tilde{W}_{\tilde{v},\tilde{k}}.G_{\tilde{k}} + L_{\tilde{v}} \tag{19}$$

$$J\left(U, G_{\tilde{k}}\right) = \sum_{\tilde{v}} \tilde{W}_{\tilde{v},\tilde{k}}.U_{\tilde{v}} + \hat{e}_{\tilde{k}} \tag{20}$$

Equation (21) specifies the $\tilde{W}_z$. Equation (22) determines the energy function of visible and hidden neurons. $N*$ specified in Eq. (23).

$$\tilde{W}_z = \max_{\Re} \prod_{U \in T} P(U) \tag{21}$$

$$P\left(U, \vec{D}\right) = \frac{1}{N*} e^{-EN\left(U,\vec{D}\right)} \tag{22}$$

$$N* = \sum_{U,\vec{G}} e^{-EN\left(U,\vec{G}\right)} \tag{23}$$

Entire weights are computed as $W'_{\tilde{v},\tilde{k}} = \Delta\tilde{W}_{\tilde{v},\tilde{k}} + \tilde{W}_{\tilde{v},\tilde{k}}$.

As per IDBN, the loss function is applied to find the loss of errors in data. Here, we used inverse binary cross entropy in Eq. (24).

$$LF(\hat{y}, \tilde{y}) = \frac{1}{\left(-\frac{1}{\tilde{N}} \sum_{\tilde{i}}^{\tilde{N}} \left[\tilde{y}_{\tilde{i}} \log \hat{y}_{\tilde{i}} + \left(1 - \tilde{y}_{\tilde{i}}\right) \log \left(1 - \hat{y}_{\tilde{i}}\right) \right] \right)} \tag{24}$$

It predicts the accuracy of loss errors for actual and predicted data. Here, $\tilde{y}$ indicates the real value, $\hat{y}$ denotes the predicted value, and $\tilde{N}$ indicates the overall count of data points. The IDBN output is displayed as Γ_{IDBN}.

CNN classifier: CNN provides the derived features Γ_{FE} as a source. The visual system of living organisms is modeled by CNN (Ghosh et al. 2020), a unique multilayer NN or DL architecture. Many NLP and computer vision applications benefit from CNN. Furthermore, CNN uses fully linked, convolutional, and pooling layers in its well-known deep learning architecture.

Convolutional Layer: The values of feature are evaluated using Eq. (25) at the offered position $(\hat{x}, \hat{z})$ inside the appropriate layer $\tilde{h}$th feature map.

$$Z^{S}_{\hat{x},\hat{z},\tilde{h}} = w^{S\hat{T}}_{\tilde{h}} \Gamma_{\mathrm{FE}} + F^{S}_{\tilde{h}} \tag{25}$$

where $w^{S}_{\tilde{h}} \rightarrow$ weight is taken in training and $_{F^{S}_{\tilde{h}}} \rightarrow$ biased layer-specific term $\tilde{h}$th filter. $\Gamma_{\mathrm{FE}} \rightarrow$ the repaired input in the layer's $(\hat{x}, \hat{z})$ middle Sth. Equation (26) describes the computation of the activation value $\left(\mathfrak{F}^{S}_{\hat{x},\hat{z},\tilde{h}}\right)$ for the convolutional features $\left(Z^{S}_{\hat{x},\hat{z},\tilde{h}}\right)$.

$$Z^{S}_{\hat{x},\hat{z},\tilde{h}} = \mathfrak{F}\left(Z^{S}_{\hat{x},\hat{z},\tilde{h}}\right) \tag{26}$$

Dropout layer: CNN also features of Dropout layer. The Dropout layer serves as a mask by preventing some neurons from contributing to the subsequent layer while keeping all other neurons active. The dropout layer can be used to remove some of the properties of an input vector or some of the neurons in a hidden layer. Dropout layers are essential while training CNNs to reduce overfitting on training data.

Pooling layer: The rate $\tilde{J}^{S}_{\hat{x},\hat{z},\tilde{h}}$ is evaluated for every pooling function $\xi(\bullet)$ that fixes $\mathfrak{F}^{S}_{\hat{x},\hat{z},\tilde{h}}$, based on Eq. (27).

$$\tilde{J}^{S}_{\hat{x},\hat{z},\tilde{h}} = \xi\left(\mathfrak{F}^{S}_{\tilde{m},\tilde{n},\tilde{h}}\right), \forall(\tilde{m},\tilde{n}) \in \Gamma_{\mathrm{FE}\ \hat{x},\hat{z}} \tag{27}$$

Flatten layer: An input's spatial dimensions are reduced to its channel dimension in a flattened layer. CNN's loss is calculated by Eq. (28).

$$Loss = \frac{1}{\varpi} \sum_{\tilde{n}=1}^{\varpi} \tilde{k}\left(\vartheta; \tilde{J}^{(\tilde{n})}, \tilde{c}^{(\tilde{n})}\right) \tag{28}$$

The input–output relationships ϖ are $\tilde{c}^{(\tilde{n})}$, $\tilde{J}^{(\tilde{n})}$ and $\Gamma^{(\tilde{n})}_{\mathrm{FE}}$ signifying the output of CNN, associated target labels, and input feature. Here $\left\{\left(\aleph^{(S)}, \tilde{Y}^{(S)}\right) ; S \in [1, \cdots, \varpi]\right\}$. The outcome of CNN is described as $\tilde{c}^{(\tilde{n})}$. The emotion recognition phase output $O*$ is specified in Eq. (29).

$$O* = \frac{\Gamma_{\mathrm{IDBN}} + \tilde{c}^{(\tilde{n})}}{2} \tag{29}$$

5.2 Objective function

Equation (30), which calculates the loss between the actual and predicted numbers by averaging the CNN loss and the IDBN loss, describes the objective function.

$$\mathrm{Obj} = \min\left(\frac{\mathrm{LF}(\hat{y}, \tilde{y}) + \mathrm{Loss}}{2}\right) \tag{30}$$

5.3 Student engagement prediction process

Online students engage in a range of learning activities, such as writing, reading, watching tutorial videos, taking online tests, and attending online conferences. They exhibit a range of involvement levels while taking part in these instructional activities, including neutral, perplexity, delight, boredom, and frustration. It is critical for online educators to accurately and effectively assess the engagement level of their online learners in order to offer individualized pedagogical support through interventions. Real-time engagement signals allow online learning environments to modify their approach to teaching in ways that a skilled instructor would, such as recommending some helpful readings, altering the course material and practice problems, and customizing exam and assignment questions. It will make it possible to develop learning management systems for online courses that are affect-sensitive. Teachers in online learning settings have the ability to provide individualized support since they receive real-time feedback regarding the engagement level of their students.

In our work, based on the emotions recognized, the student's engagement will be detected (with the improved process). It shows the performance as well. The engagement index is formulated as in Eq. (31).

$$\mathrm{EI} = \mathrm{score} \times \mathrm{IE} \tag{31}$$

where EI is the engagement index and IE is the improved entropy in Eq. (32).

$$\mathrm{IE}(\hat{g}) = -\sum_{\hat{b} \le \rho} \hat{g}(A*) \log\left(\frac{\hat{g}(A*)}{2^{|A*|}-1}.e^{\frac{|A*|-1}{|\hat{s}|}}\right) * \mathrm{CV} \tag{32}$$

Improved entropy is used to gauge the level of uncertainty in a data set. It can be used to quantify the internal information qualities of the signal and to define the uncertainty distribution and complexity properties of the signal.

$$\mathrm{CV} = \frac{Standard\ deviation}{mean} \tag{33}$$

Here, CV is the coefficient of variance in Eq. (33), $|A*|$ is the cardinality of the focal element $A*$, and $\hat{g}(A*)$ is the mass function.

Steps for finding the score:

Consider an example with proposed target values as 0, 5, 6, 0, 1, 4, 0, 2, 4, 6. Here, 0 is anger, 1 is contempt, 2 is disgust, 3 is fear, 4 is happy, 5 is sadness, and 6 is surprise.

If the proposed target value [0] = = 0, then val = anger. For each of the goal values, the exact same process is performed.

Step 1 Score that goal

$$Finalscore = score(anger)$$

For example, the last score values are [0.59, 0.3, 0.12, 0.61].
Step 2 Median value of last score values.
Step 3 Fix a threshold as the median value.
Step 4 Then check, if final score values > threshold, students are engaged. Otherwise, not engaged.

Table 2 describes the attributes used in the proposed model. Table 3 shows the parameters of the proposed model.

6 Outcomes and discussions

6.1 Simulation setup

The hybrid method for emotion recognition was evaluated to extant systems using the PYTHON 3.7.9, and its outcomes were analyzed. The performance of the hybrid model was compared against prior techniques such as CNN, DBN, CNN (Cabada et al. 2020), LSTM-CNN (Hassouneh et al. 2020), SVM, LSTM, RF, DCNN, and BiLSTM approaches based on positive, negative, and other metrics using the CK+ dataset and FER-2013 dataset for various learning percentages of 60, 70, 80, and 90, respectively. The dataset was also obtained from https://www.kaggle.com/datasets/msambare/fer2013 and https://www.kaggle.com/datasets/shawon10/ckplus. Figure 4 depicts a sample image representation for the CK+ dataset and the FER-2013 dataset. The processors carried out in the system model are listed in Table 4.

Table 2 Notations and descriptions

Notations	Descriptions
Γ_{pre}	Preprocessed data
$\hat{T}_1$	Texture vector for pixel
(Y_1)	Position of n landmark point
μ	Mean of (Y_1)
σ	Standard deviation of (Y_1)
Γ_{IAAM}	AAM feature
$\Gamma_{Res-net}$	ResNet features
Z	Shape vector
$\overline{Z}$	Mean shape
V_s	Eigen vectors of largest eigen values
A_s	Shape model
C	Shape texture parameter
V_{st}	Eigenvectors
$\overline{A_{st}}$	Mean vector
Γ_{SLBT}	SLBT feature
Γ^x_{GBP}	Global binary pattern
Γ_{FE}	Extracted feature set
$U_{\tilde{v}}$	Visible binary state
$G_{\tilde{k}}$	Hidden units
$\tilde{W}$	Weight function
$\tilde{y}$	Actual value
$\hat{y}$	Predicted value
Γ_{IDBN}	Output of IDBN
IE	Improved entropy
EI	Engagement index
CV	Coefficient of variance

Table 3 Parameters of the proposed model

Parameters	Values
CNN	
Epoch	25
Loss	Categorical cross entropy
Optimizer	adam
Activation	softmax
Batch size	32
Dropout	0.5 (units)
DBN	
Hidden layers structure	Zhang et al. (2020)
Learning rate rbm	0.05
Learning rate	0.1
n epochs rbm	1
n iter backprop	1
Batch size	32
Activation function	'sigmoid'
Dropout_p	0.2

6.2 Dataset description

The two datasets as CK+ dataset and the FER-2013 dataset are used in this work. "The FER- 2013 dataset contains 48 × 48 pixel grayscale photos of faces. The faces have been automatically registered such that the face is more or less in the center of each image and occupies roughly the same amount of area. The aim is to put each face into one of seven categories depending on the emotion portrayed in the facial expression (0 = Angry, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Sad, 5 = Surprise, 6 = Neutral). The total number of photos in the FER-2013 dataset is 1400. The aim of the CK+ dataset is to identify each face into one of seven categories depending on the emotion expressed in the facial expression, such as anger, contempt, disgust, fear, gladness, sorrow, and surprise. The total number of images utilized in the CK+ dataset is 981". Table 5 describes the training and testing pictures for the FER-2013 dataset and the CK+ dataset.

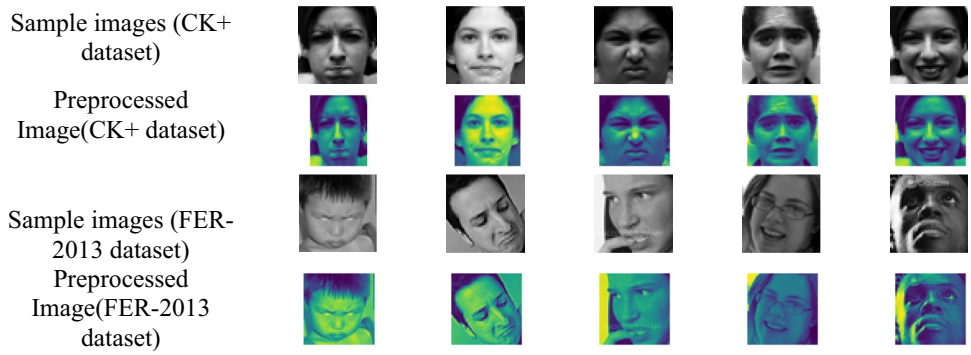

Fig. 4 Sample image representation for CK+ dataset and FER-2013 dataset

Table 4 System configuration

Device specifications	
Processor	11th Gen Intel(R) Core (TM) i5-1135G7 @ 2.40 GHz 2.42 GHz
Installed RAM	16.0 GB (15.7 GB usable)
System type	64-bit operating system, × 64-based processor
Windows specifications	
Edition	Windows 11 Home Single Language
Version	21H2

Table 5 Testing and training images for CK+ dataset and FER-2013 dataset

Learning percentage	Testing images	Training images
CK+ dataset		
60%	588	393
70%	686	294
80%	784	196
90%	882	99
FER-2013 dataset		
60%	560	840
70%	421	979
80%	280	1120
90%	140	1260

6.3 Analysis of positive metrics of the hybrid model based on the emotion recognition process over previous approaches for the CK+ dataset and FER-2013 dataset

Figures 5 and 6 show the analysis of a hybrid model on emotion recognition to existing approaches using positive metrics like sensitivity, precision, accuracy, and specificity for the CK+ and FER-2013 datasets. Furthermore, when compared to previous schemes such as CNN, DBN, CNN (Hassouneh et al. 2020; Cabada et al. 2020), LSTM-CNN (Hassouneh et al. 2020), SVM, LSTM, RF, DCNN, and BiLSTM for emotion recognition, the hybrid model had the greatest accuracy (0.95) at a learning percentage of 80% for the CK+ dataset, while other models achieve less accuracy of 0.81, 0.83, 0.83, 0.83, 0.86, 0.85, 0.85, 0.84, and 0.85. In comparison to a learning percentage of 80%, the hybrid model has a higher sensitivity of 60% for FER-2013 datasets. Furthermore, in 90% learning, the hybrid model of emotion recognition surpassed other traditional approaches with specificity (0.92). The hybrid model on emotion recognition offers higher accuracy results at a learning percentage of 60% than other techniques such as CNN, DBN, CNN (Cabada et al. 2020), LSTM-CNN (Hassouneh et al. 2020), SVM, LSTM, RF, DCNN, and BiLSTM. The results analysis proves the influence of the proposed concept on Improved AAM features to train the model for accurate recognition.

6.4 Negative measure analysis of hybrid model over previous approaches for CK+ dataset and FER-2013 dataset

Figures 7 and 8 specify the analysis of the hybrid model on emotion recognition to previous schemes like CNN, DBN, CNN (Cabada et al. 2020), LSTM-CNN (Hassouneh et al. 2020), SVM, LSTM, RF, DCNN, and BiLSTM correspondingly for CK+ dataset and FER-2013 dataset based on negative metrics (FNR and FPR). The hybrid model on emotion recognition has attained the minimal FPR (0.070) with the better outcomes at 80% learning than at a 70% for the CK+ dataset, while other models such as CNN, DBN, CNN (Cabada et al. 2020), LSTM-CNN (Hassouneh et al. 2020), SVM, LSTM, RF, DCNN, and BiLSTM receive the highest ratings of 0.178, 0.165, 0.164, 0.166,0.157, 0.167, 0.166, 0.163, 0.164, and 0.1622 at an 80 learning percentage. Furthermore, the hybrid model on emotion recognition outperformed previous schemes including CNN, DBN, CNN (Cabada et al. 2020), LSTM-CNN (Hassouneh et al. 2020), SVM, LSTM, RF, DCNN, and BiLSTM correspondingly for the FER-2013 dataset at a learning percentage of 90%, in terms of minimal FNR values (0.012). Finally, as compared to earlier schemes, the hybrid model gives its lowest error

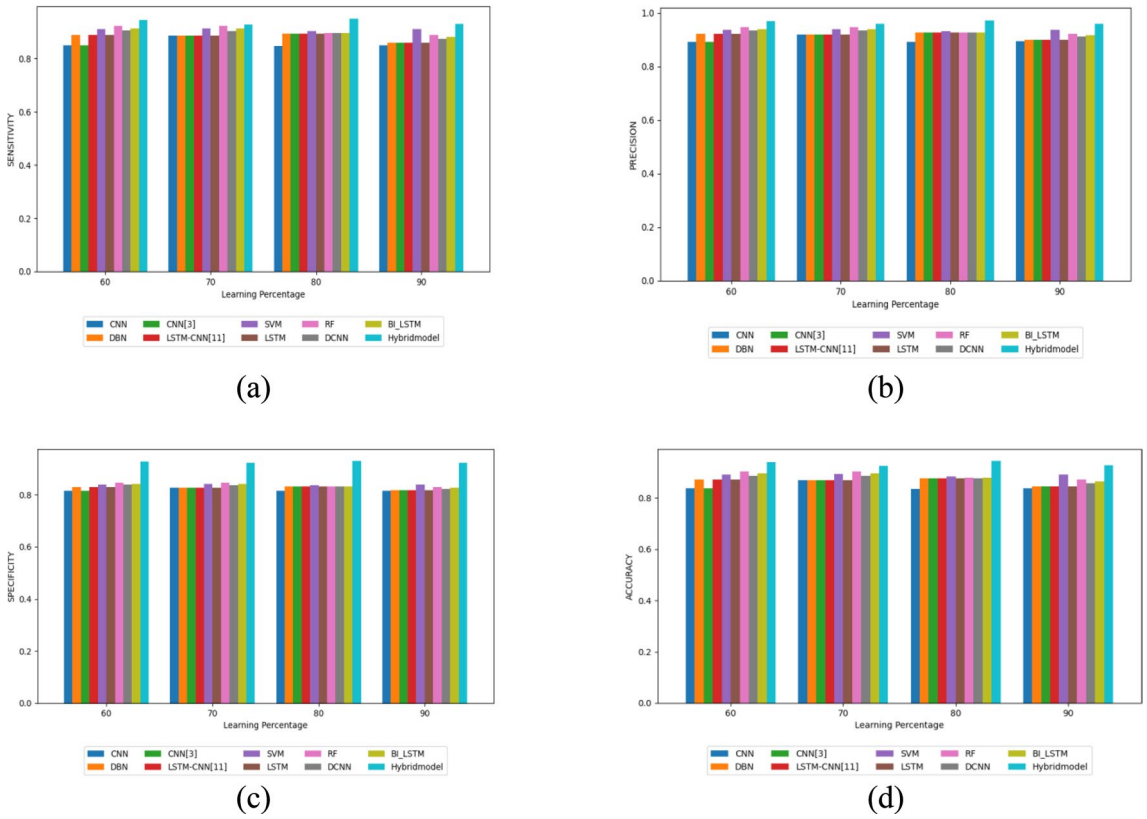

Fig. 5 Analysis of the hybrid model over the previous schemes for **a** Sensitivity **b** Precision **c** Specificity **d** Accuracy for the CK+ dataset

value than the conventional schemes that show the impact of IDBN with novel loss evaluation during training.

6.5 Analysis of other measures of the hybrid model over previous approaches for the CK+ dataset and FER-2013 dataset

Figures 9 and 10 determine the hybrid model on emotion recognition to conventional schemes CNN, DBN, CNN (Cabada et al. 2020), LSTM-CNN (Hassouneh et al. 2020), SVM, LSTM, RF, DCNN and BiLSTM correspondingly for CK+ dataset and FER-2013 dataset of other metrics. The F-measure for the hybrid model on emotion recognition was higher (0.94) at 80% LP than at 70% LP for the CK+ dataset. Furthermore, after 90% of learning, the NPV of the hybrid model on emotion recognition was considerably higher (0.91) than CNN, DBN, CNN (Hassouneh et al. 2020; Cabada et al. 2020), SVM, LSTM, RF, DCNN, and BiLSTM which attain 0.75, 0.74, 0.76, 0.80, 0.77, 0.78, 0.76, and 0.76 for the FER-2013 dataset. When the learning rate is 70%, the hybrid model on emotion recognition achieved a higher MCC value for the CK+ dataset; however, the previous methods like CNN, DBN, CNN (Cabada et al. 2020), LSTM-CNN (Hassouneh et al. 2020), SVM, LSTM, RF, DCNN, and BiLSTM correspondingly hold less MCC value. Thereby, the impact of the suggested hybrid model enriches its prediction in terms of other measures.

6.6 Impact of the hybrid model over previous approaches for the CK+ dataset and FER-2013 dataset

Tables 6 and 7 illustrate the impact of hybrid model-based emotion recognition processes over previous approaches like method with extant AAM, method with extant DBN, and method without feature extraction, correspondingly for both datasets. Moreover, the model without feature extraction yields low-precision results (0.922), while the precision of the proposed hybrid model on student emotion recognition yields maximum results (0.959) for the CK+ dataset. Further, the proposed hybrid model on emotion recognition with an NPV value is 0.854 greater; nevertheless, the method with extant AAM (0.735), method with extant DBN (0.729), and method without feature extraction (0.731) achieves lower NPV values for FER-2013 dataset. Thus, the proposed hybrid model has attained better outcomes on emotion recognition.

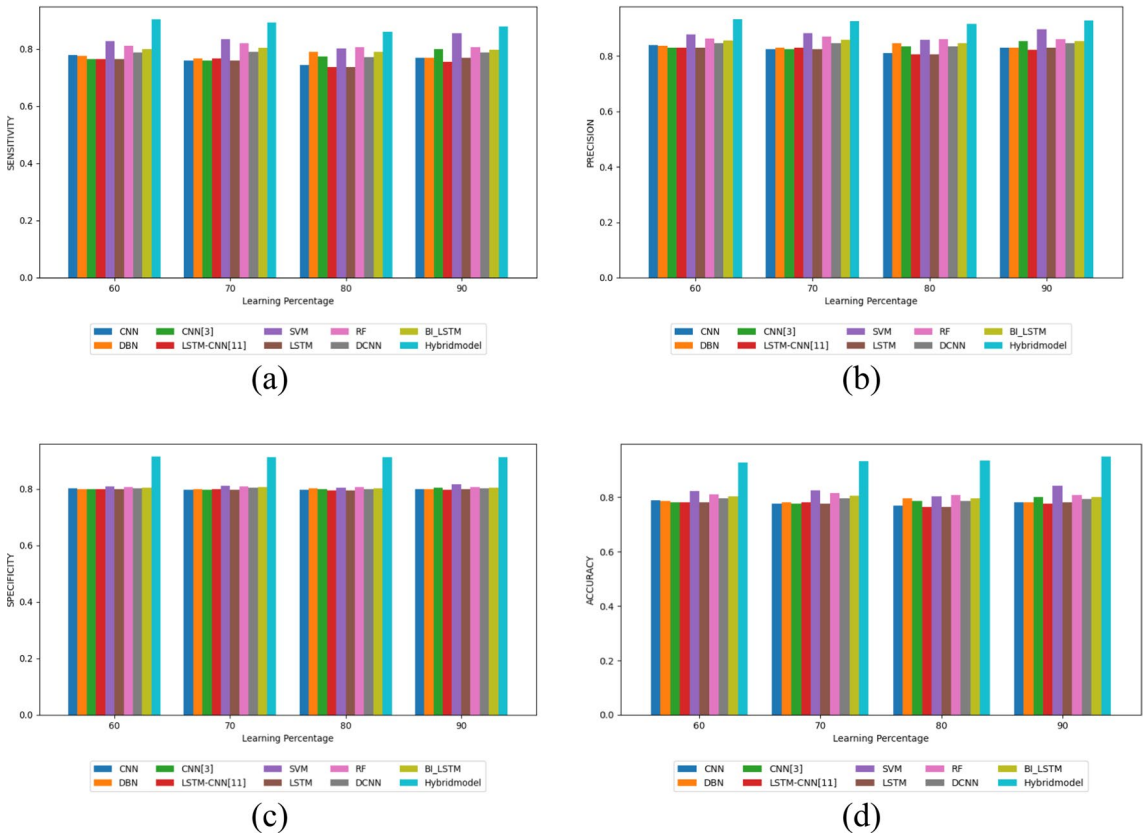

Fig. 6 Analysis of the hybrid model over previous schemes for **a** Sensitivity **b** Precision **c** Specificity **d** Accuracy for the FER-2013 dataset

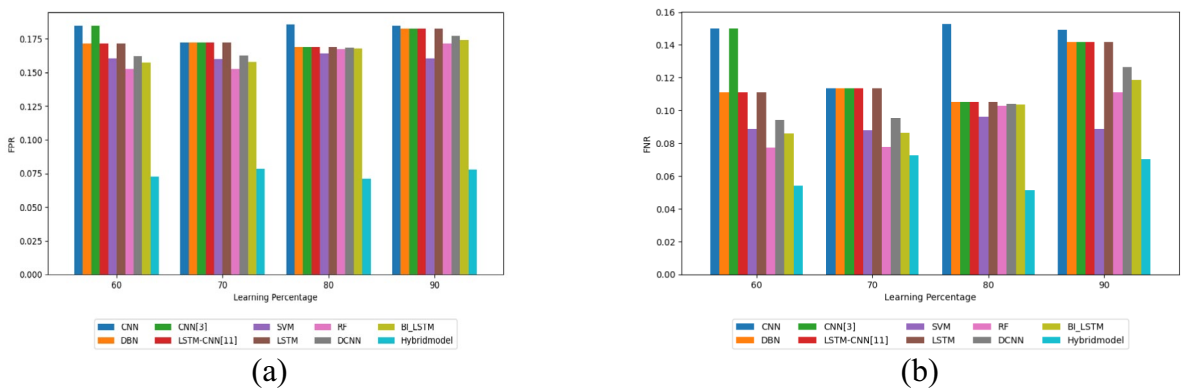

Fig. 7 Analysis of the hybrid model over previous schemes for **a** FPR and **b** FNR for the CK+ dataset

6.7 Statistical analysis of hybrid model over previous approaches for CK+ dataset and FER-2013 dataset

Tables 8 and 9 examine the statistical analysis of the hybrid model-based emotion recognition process over previous approaches like CNN, DBN, CNN (Cabada et al. 2020), LSTM-CNN (Hassouneh et al. 2020), SVM, LSTM, RF, DCNN, and BiLSTM correspondingly for CK+ dataset and FER-2013 dataset. Despite the unpredictability of the optimization model, statistical analysis is frequently used to assess the whole performance after several cycles. Compared with the other methods, the hybrid classification has attained a higher rate (0.942) under the Max

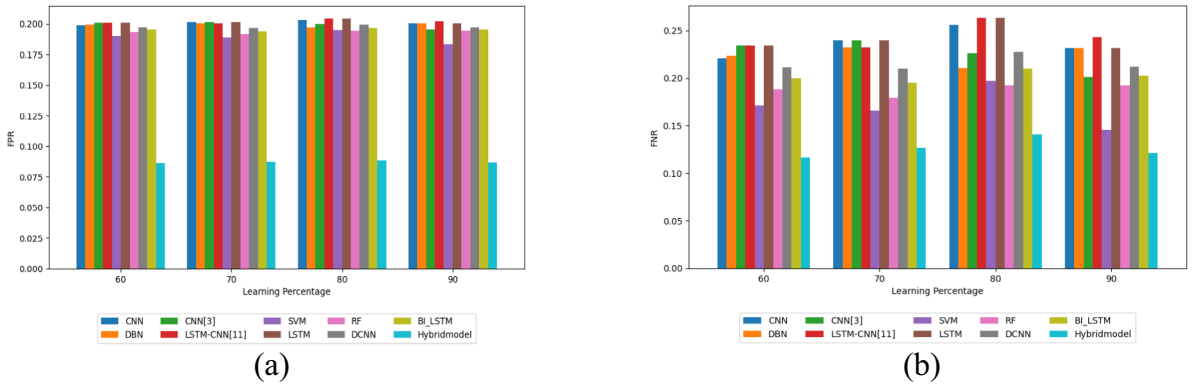

Fig. 8 Analysis of the hybrid model over traditional methods for **a** FPR and **b** FNR for FER-2013 dataset

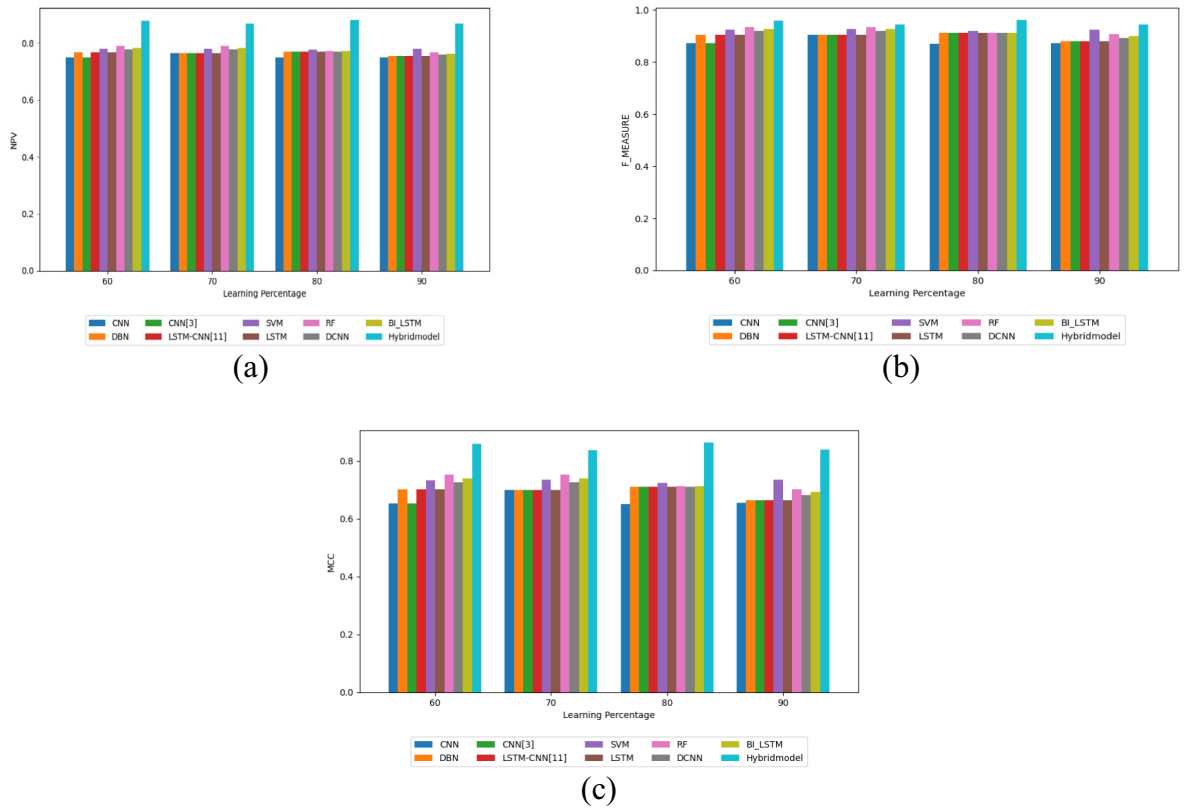

Fig. 9 Analysis of the hybrid model over previous schemes for **a** NPV **b** F-measure **c** MCC for the CK+ dataset

case on student engagement prediction for online learners for the CK+ dataset. Moreover, the hybrid model outperformed other models including the CNN, DBN, CNN (Cabada et al. 2020), LSTM-CNN (Hassouneh et al. 2020), SVM, LSTM, RF, DCNN, and BiLSTM correspondingly in terms of best mean case values (0.935) for the FER-2013 dataset and excellent outcome for emotion recognition. In all case scenarios, the hybrid model execute superior outcomes than the other models for online learners' engagement detection via facial emotion recognition in an online learning context.

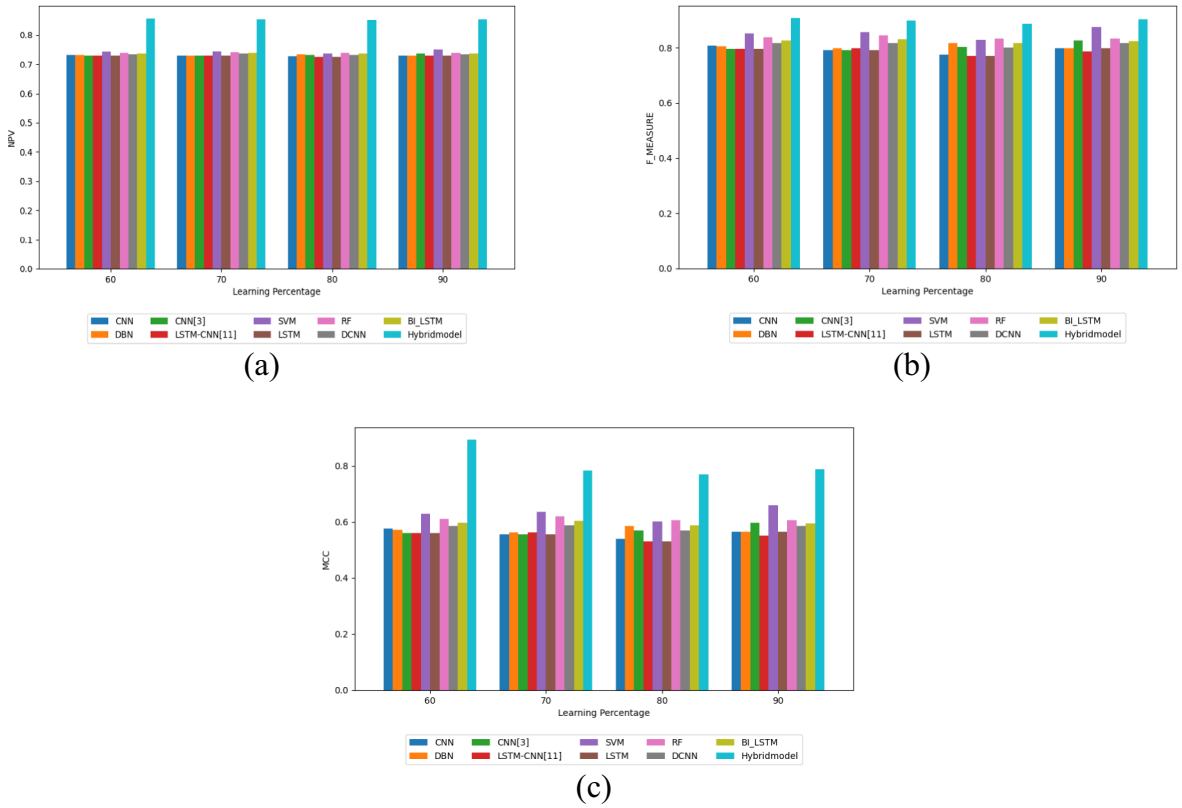

Fig. 10 Analysis of the hybrid model over previous schemes for **a** NPV **b** F-measure **c** MCC for the FER-2013 dataset

Table 6 Impact on hybrid model-based emotion recognition process over previous approaches for CK+ dataset

Metrics	Method with extant AAM (%)	Method with extant DBN (%)	Method without feature extraction (%)	Suggested hybrid model (%)
Accuracy	87.192	87.050	87.121	92.708
Sensitivity	89.046	88.890	88.968	92.953
MCC	70.472	70.261	70.367	83.986
Specificity	82.926	82.859	82.893	92.223
Precision	92.306	92.192	92.249	95.943
NPV	76.696	76.611	76.654	86.868
FPR	17.073	17.140	17.107	7.776
FNR	10.953	11.109	11.031	7.046
F-measure	90.647	90.511	90.579	94.425

6.8 Computational time analysis of the hybrid model over previous approaches for CK+ dataset and FER -2013 dataset

Tables 10 and 11 describe the computational time analysis of the hybrid model-based emotion recognition process over previous approaches like CNN, DBN, CNN (Cabada et al. 2020), LSTM-CNN (Hassouneh et al. 2020), SVM, LSTM, RF, DCNN, and BiLSTM correspondingly for CK+ dataset and FER-2013 dataset. On observing the analysis outcomes, the computational time for the proposed hybrid model has achieved minimal values (31.2815) and (27.48559) for the CK+ dataset and FER-2013 dataset over the existing schemes such as CNN, DBN, CNN (Cabada et al. 2020), LSTM-CNN (Hassouneh et al. 2020), SVM, LSTM, RF, DCNN, and BiLSTM respectively. Thus, the adopted hybrid method has less time for computation.

Table 7 Impact on hybrid model-based emotion recognition process over previous approaches for the FER-2013 dataset

Metrics	Method with extant AAM (%)	Method with extant DBN (%)	Method without feature extraction (%)	Proposed hybrid model (%)
Accuracy	80.223	77.501	78.862	94.838
Sensitivity	80.070	75.629	77.849	87.848
Specificity	80.446	79.793	80.119	91.342
Precision	85.568	82.077	83.823	92.837
F-measure	82.727	78.721	80.724	90.274
MCC	59.837	55.146	57.492	78.751
NPV	73.59	72.794	73.196	85.476
FPR	19.553	20.206	19.880	8.657
FNR	19.929	24.370	22.150	12.151

Table 8 Statistical analysis of the hybrid model over previous approaches for the CK+ dataset

	CNN	DBN	CNN (Cabada et al. 2020)	LSTM-CNN (Hassouneh et al. 2020)	SVM	LSTM	RF	DCNN	BI_LSTM	Hybrid model
Mean	0.8449	0.8646	0.8564	0.8646	0.8900	0.8646	0.8887	0.8767	0.8827	0.9339
Median	0.8378	0.8693	0.8562	0.8693	0.8916	0.8693	0.8905	0.8813	0.8859	0.9336
Std	0.0135	0.0121	0.0159	0.0121	0.0032	0.0121	0.0146	0.0117	0.0128	0.0077
Min	0.8356	0.8442	0.8375	0.8442	0.8843	0.8442	0.8706	0.8574	0.8640	0.9254
Max	0.8682	0.8758	0.8758	0.8758	0.8924	0.8758	0.9032	0.8868	0.8950	0.9429

Table 9 Statistical analysis of the hybrid model over previous approaches for the FER-2013 dataset

	CNN	DBN	CNN (Cabada et al. 2020)	LSTM-CNN (Hassouneh et al. 2020)	SVM	LSTM	RF	DCNN	BI_LSTM	Hybrid model
Mean	0.7790	0.7864	0.7860	0.7754	0.8231	0.7760	0.8096	0.7928	0.8012	0.9354
Median	0.7796	0.7845	0.7829	0.7778	0.8235	0.7788	0.8081	0.7947	0.8014	0.9334
Std	0.0075	0.0053	0.0091	0.0068	0.0133	0.0069	0.0035	0.0042	0.0035	0.0079
Min	0.7680	0.7818	0.7772	0.7642	0.8039	0.7642	0.8067	0.7855	0.7961	0.9265
Max	0.7886	0.7949	0.8009	0.7818	0.8413	0.7821	0.8155	0.7964	0.8059	0.9483

Table 10 Computational time analysis of the hybrid model over previous approaches for the CK+ dataset

Methods	Time
CNN	58.97222
DBN	65.46968
CNN (Cabada et al. 2020)	65.05123
LSTM-CNN (Hassouneh et al. 2020)	58.20244
SVM	57.33552
LSTM	65.49367
RF	63.62359
DCNN	61.57225
BI_LSTM	64.27183
Hybrid model	31.2815

Table 11 Computational time analysis of the hybrid model over previous approaches for the FER-2013 dataset

Methods	Time
CNN	53.61284
DBN	67.36283
CNN (Cabada et al. 2020)	65.94747
LSTM-CNN (Hassouneh et al. 2020)	68.2372
SVM	63.02423
LSTM	65.91891
RF	48.76416
DCNN	45.0245
BI_LSTM	50.5677
Hybrid model	27.48559

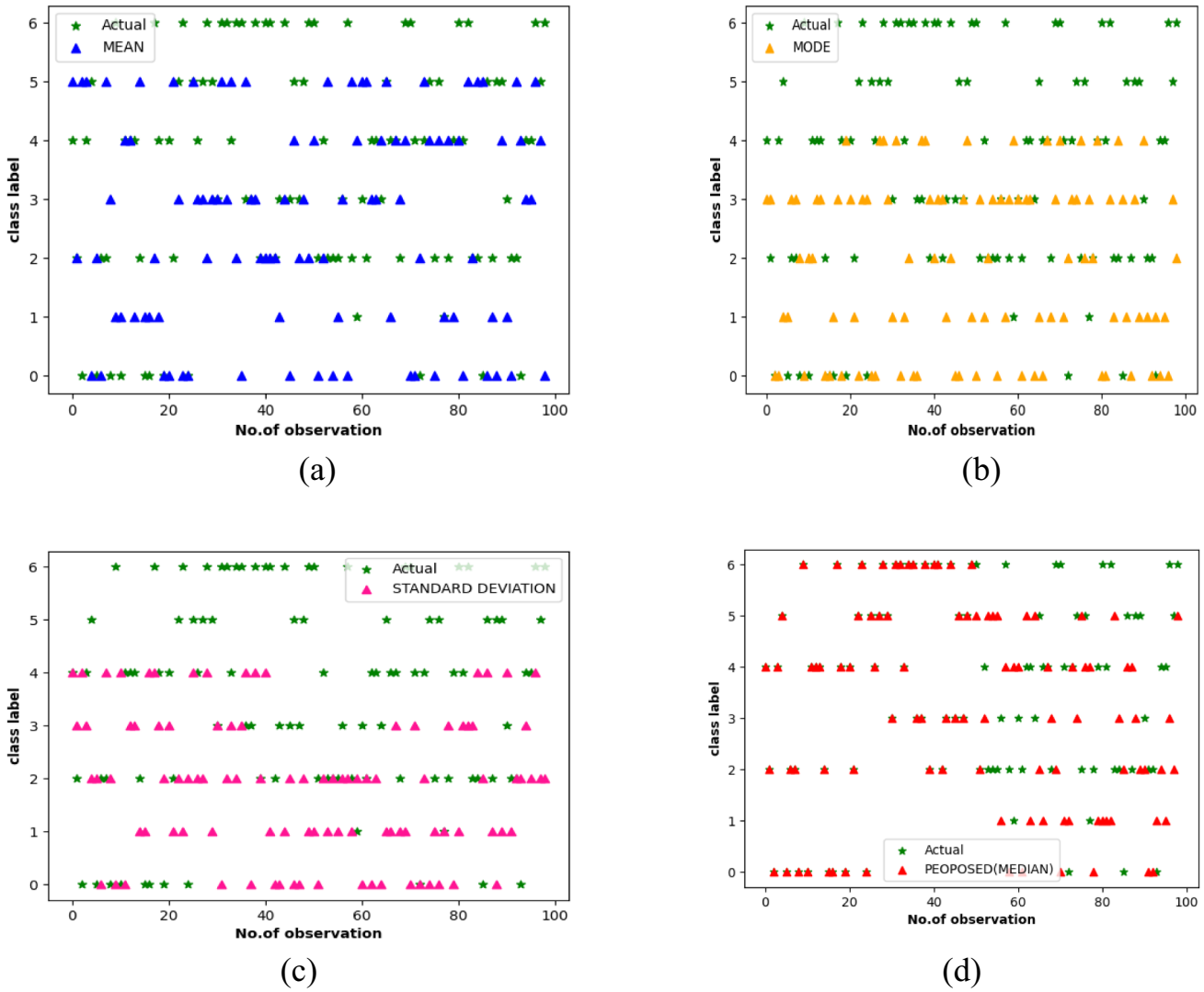

Fig. 11 Analysis of median over other statistical measures **a** mean **b** mode **c** standard deviation **d** median

6.9 Analysis of median over other statistical measures

Figure 11 describes the performance of the median over other statistical measures such as mean, mode, and standard deviation. The middle observation in a dataset arranged in either decreasing or increasing order of magnitude is called the median. As a result, it is an observation that is at the center of the distribution (data). Another name for this is the positional average. Outliers, or extreme values, have no effect on the median. Because there is only one median of a single data set, it is unique and helpful for group comparisons. This makes the median a threshold value for the detection of student engagement. However, other statistical measures such as mean, mode, and standard deviation have some issues. Mean is affected by extreme values (outliers), mode does not consider all the values in the data, and standard deviation doesn't give you the full range of the data. On observing Fig. 11, it is clearly evident that the median has better scores than the other measures. The median has more scores than the actual values, while other statistical measures have lower scores than the actual values. This makes the median a threshold value for detecting the final score for detecting the engagement of students in online classes.

7 Conclusion

This article proposed a new student engagement prediction based on facial emotion for online learners, which includes four basic phases: (a) preprocessing, (b) feature extraction, (c) emotion recognition, and (d) student performance prediction. The preprocessing step was the initial phase in which the Face detection process occurred. Following the

preprocessing phase, the feature extraction step occurs, during which the Improved AAM, SLBT, GBP, and ResNet features are derived. The emotion recognition phase was then applied to these extracted characteristics. The recognition procedure was carried out using a Hybrid Classification technique, which fuses methods such as CNN and IDBN. The student's involvement was identified (enhanced entropy process) based on the emotions recognized, which also displays their performance. When the learning value is 70%, the hybrid method on emotion recognition prediction for online learners achieved higher MCC value for the CK+ dataset; however, the previous methods like CNN, DBN, CNN, LSTM-CNN, SVM, LSTM, and RF correspondingly hold less MCC value. The suggested engagement detection method can enhance students' online learning experiences, including reading, writing, viewing training videos, and taking part in meetings. This also helps to recognize the accurate recognition of facial emotions of students. However, the number of engagement levels is needed for the effectiveness in recognizing during online learning. In future, this work will be extended to include the incorporation of many features, including interactions between learners and tutors as well as facial expressions, eye movement, gestures, and learner learning activities in LMS.

Authors' contributions RBRM conceived the suggested idea and designed the analysis. Along with that, MS helped him write the book and conduct the experiment. Each author contributed to the final manuscript and discussed the findings. The final manuscript was read and approved by all writers.

Funding No specific fund was provided for this article.

Data availability The data underlying this article are available in https://www.kaggle.com/datasets/msambare/fer2013, and https://www.kaggle.com/datasets/shawon10/ckplus.

Declarations

Conflict of interest The authors say they have no conflicts of interest.

Informed consent Not applicable.

Ethical approval Not applicable.

References

Ashwin TS, Guddeti RMR (2020a) R.M.R. impact of inquiry interventions on students in e-learning and classroom environments using affective computing framework. User Model User-Adap Inter 30

Ashwin TS, Guddeti RMR (2020b) Automatic detection of students' affective states in a classroom environment using hybrid convolutional neural networks. Educ Inf Technol 25

Barnouti NH, Al-Dabbagh SSM, Matti WE (2016) Naser MAS Face detection and recognition using viola-jones with PCA-LDA and square Euclidean distance. Int J Adv Comput Sci Appl 7(5)

Benabbes K, Housni K, Hmedna B, Zellou A, Mezouary AE (2023) A new hybrid approach to detect and track learner's engagement in e-Learning. IEEE Access 11:70912–70929. https://doi.org/10.1109/ACCESS.2023.3293827

Buono P, De Carolis B, D'Errico F, Macchiarulo N, Palestra G (2022) Assessing student engagement from facial behaviour in online learning. Multimed Tools Appl 82:12859–12877

Cabada RZ, Rangel HR, Estrada MLB, Lopez HMC (2020) Hyperparameter optimization in CNN for learning-centred emotion recognition for intelligent tutoring systems. Soft Comput 24

Dewan MAA et al (2018) A deep learning approach to detecting engagement of online learners. In: 2018 IEEE SmartWorld, ubiquitous intelligence & computing, advanced & trusted computing, scalable computing & communications, cloud & big data computing, internet of people and smart city innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE

Ding Y, Xing W (2022) Emotion recognition and achievement prediction for foreign language learners under the background of network teaching. Front Psychol. https://doi.org/10.3389/fpsyg.2022.1017570

Ghosh A, Sufian A, Sultana F, Chakrabarti A, De D (2020) Fundamental concepts of convolutional neural network. Chapter January 2020 https://doi.org/10.1007/978-3-030-32644-9_36.

Gupta S, Kumar P, Tekchandani RK (2022) Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimed Tools Appl 8:11365–11394

Hassouneh A, Mutawa AM, Murugappan M (2020) Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods. Inform Med Unlock 20:100372

Hung JC, Lin K-C, Lai N-X (2019) Recognizing learning emotion based on convolutional neural networks and transfer learning. Appl Soft Comput 84:105724

Iqtait M, Mohamad FS, Mamat M (2018) Feature extraction for face recognition via active shape model (ASM) and active appearance model (AAM). In: IOP conferemce series: materials science and engineering vol 332, pp 012032. https://doi.org/10.1088/1757-899X/332/1/012032

Lakshmiprabha NS, Majumder S (2012) Face recognition system invariant to plastic surgery. In: 2012 12th international conference on intelligent systems design and applications (ISDA), Kochi. pp 258–263. https://doi.org/10.1109/ISDA.2012.6416547

Lasri I, Riadsolh A (2023) Elbelkacemi M (2023) Facial emotion recognition of deaf and hard-of-hearing students for engagement detection using deep learning. Educ Inf Technol 28:4069–4092. https://doi.org/10.1007/s10639-022-11370-4

Liao J, Liang Y, Pan J (2021) Deep facial spatiotemporal network for engagement prediction in online learning. Appl Intell 51:6609–6621

Mehta NK, Prasad SS, Saurav S, Saini R, Singh A (2022) Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement. Appl Intell 52:13803–13823

Mhapsekar M, Mhapsekar P, Mhatre A, Sawant V (2015) Implementation of residual network (ResNet) for devanagari handwritten character recognition. In: Advanced computing technologies and applications, algorithms for intelligent systems, https://doi.org/10.1007/978-981-15-3242-9_14

Ngai WK, Xieb H, Zouc D, Chou KL (2022) Emotion recognition based on convolutional neural networks and heterogeneous biosignal data sources. Inf Fusion 77:107–117

Ninaus M, Greipla S, Kiili K, Lindstedt A, Huber S, Klein E, Karnath H-O, Moeller K (2019) Increased emotional engagement

in game-based learning—a machine learning approach on facial emotion detection data. Comput Educ 142:103641
Said Y, Barr M (2021) Human emotion recognition based on facial expressions via deep learning on high-resolution images. Multimed Tools Appl 80:25241–25253
Savchenko AV, Makarov IA (2022) Neural network model for video-based analysis of student's emotions in E-learning. Opt Memory Neural Netw 3:237–244
Schoneveld L, Othmani A, Abdelkawy H (2021) Leveraging recent advances in deep learning for audio-visual emotion recognition. Pattern Recogn Lett 146:1–7
Sivri E, Kalkan S (2013) Global binary patterns: a novel shape descriptor. In: International IAPR conference on machine vision and applications (20–23 May 2013)
Vidyadhari S, Mishra DS (2020) Analysis of facial expressions for predicting student's learning level. JCSE Int J Comput Sci Eng Open Access 8
Wang HZ, Wang GB, Li GQ, Peng JC, Liu YT (2016) Deep belief network based deterministic and probabilistic wind speed forecasting approach. Appl Energy 182:80–93
Xu R, Chen J, Han J, Tan L, Xu L (2020) Towards emotion-sensitive learning cognitive state analysis of big data in education: deep learning-based facial expression analysis using ordinal information. Computing 102:765–780
Yuan Q (2022) Research on classroom emotion recognition algorithm based on visual emotion classification. Comput Intell Neurosci. https://doi.org/10.1155/2022/6453499
Zhang Z, Li Z, Liu H, Cao T, Liu S (2020) Data-driven online learning engagement detection via facial expression and mouse behaviour recognition technology. J Educ Comput Res 58
Zhu X, Chen Z (2020) Dual-modality spatiotemporal feature learning for spontaneous facial expression recognition in e-learning using hybrid deep neural network. Vis Comput 36:743–755