Recognition of Learners’ Cognitive States using Facial Expressions in E-learning Environments

Karu Prasada Rao[1,*], Dr. M.V.P Chandra Sekahara Rao[2]

[1,*]Research Scholar, Dept.of.CSE, ANU, Guntur, [2]Professor, Dept.of.CSE, RVR & JCCE, Guntur

Abstract:

Technological developments in e-learning systems provide new opportunities for students to enhance academic growth and improve access to education. E-learning is on the rise because it has advantages over conventional learning. The need for time is a faster and simpler learning experience. The spread of the coronavirus disease (COVID-19) pandemic has resulted in school closures around the world. More than one billion students are out of the classroom worldwide. As a consequence, education has taken on a new shape, with a major increase in e-learning, whereby teaching is carried out online and on digital platforms. The scope of this research is to identify the facial expressions of the students and then link these expressions to the cognitive states. This paper presents a hybrid-CNN model to recognize a learner's cognitive state using the manually engineered features and features extracted from the convolutional neural network. In addition, the performance of the model is compared with the manual feature extraction method and CNN methods separately. The proposed method trained and tested with the spontaneous database(DAiSEE) created exclusively for the e-learning environment, and with the state-of-the-art datasets such as JAFFE and CK+. The model has achieved 53.4%, 71.4%, and 99.95% respectively.

Keywords: Cognitive states, Convolutional Neural Networks, E-learning, Facial Expressions

1. Introduction

The learning environment today reflects on the vision of faculty and students collaboratively working towards profound, meaningful, high-quality learning. Information and Communication Technology (ICT) achievements in education are leading learning societies to a new level. Online classes, tutorials, and anything available in online or offline digital formats are included in e-learning. E-learning is at the fastest pace, and it has the edge over conventional learning. The expeditious learning experience has a high demand on time.

The advancement of technology connects learners across the globe and makes everyone feel like they are inside the classroom. There are several different e-learning systems (also called Learning Management Systems) and approaches that typically use course materials in a variety of formats, such as videos, slides, PDFs, and Word documents. E-learning has vast potential to keep the students who are already on the track of the curriculum as well as the employees, off the track of conventional learning abreast of the up-to-date knowledge on industry and technological know-how. Hence, many multinational companies are offer training to their employees on an e-learning basis, which will save time and cost. As individuals, we do not all react in the same manner to one form of education – some people learn visually, while others learn to repeat or write. E-learning responds to these particular needs by using different types of materials. There are many resources available to meet the needs of each learner to learn online in a much more efficient way.

The outbreak of corona virus disease (COVID-19) pandemic has resulted in schools shut all across the world. Globally, more than one billion students are out of the classroom. It expected that the shutdown would impact the learning process to a great extent. But it is not the case. Different companies such as Vedantu, Unacademy, and Byju’s have offered free access to live classes to the students to continue their learning from home. Most existing e-learning methods focus on providing learning materials, conducting online quizzes, content delivery mechanisms (MOOC) and video lectures, etc., through LMS tools like Moodle, Coursera, etc. However, these methods considered to be passive because of ignoring face-to-face communication with the learner.

In the traditional classroom, the teacher and students have direct access to each other. The teacher has an eagle's eye and knows the students who are interested and who are disinterested. So that instructor adapts to the situation and tries to adjust his/her presentation. This direct access is the indwelling gap in the current e-learning system. The new system has to build with an inbuilt camera and de novo application that can read the learners' emotions and send the feedback to the instructor so that he/she can adapt to the situation. Unfortunately, many e-learning systems are not considering the learner's facial expressions. Hence, in the early 2000s, expectations of converting conventional education to e-learning could not meet with web technologies in the sense of course materials as shown in **Fig. 1** [9].

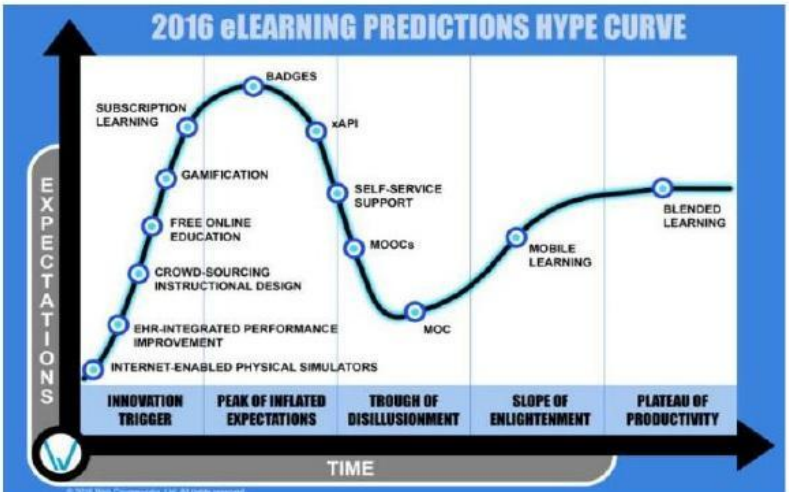


Fig. 1. E-learning Hype Curve

This research specifically focused on the automatic detection of learning status from the facial expressions of the students/learners. The deformation of facial components such as raising the eyebrows, rolling eyes, and opening of the mouth, etc., can map with each emotion. Learner's facial expressions vary according to the level of understanding of the content. Most of the current facial expression recognition (FER) methods are based on six typical emotions, such as happiness, sadness, anger, disgust, fear, and surprise from the posed expression databases like Cohn-Kanade (CK), CK+, Japanese Female Facial Expression (JAFFE). However, these emotions are not correctly indicating the student’s cognitive state.

The proposed framework is to identify the spontaneous facial expressions of students in e-learning environments automatically. **Fig.2** displays the flowchart of the proposed method. Facial expressions identified using deep convolutional neural networks, geometric based methods. Further, results compared with the fusion approach, which outperforms both the techniques. The CNN and handcrafted algorithms trained on the DAiSEE (Dataset for Affective States in E-learning Environments) built on the real-world e-learning environments and tested with state-of-the-art FER datasets such as JAFFE and CK+.

2. Literature Review

An effective emotion recognition system depends on the mapping of facial expressions to a student's cognitive states. The widely used facial expression recognition methods classified as geometry and appearance-based approaches. In a geometric-based approach, by mapping and measuring the deformation of facial components such as eyes, nose, eyebrow, and mouth [8]. In the appearance-based process, features include details on appearance changes in the face, including texture, wrinkles, bulges, and furrows.

S L Happy et al.[1] designed the system to identify students' alertness levels using PERCLOS (PERcentage of CLOSure of eye) and saccadic parameters. Further, facial expressions classified with LBP (Local Binary Pattern) features. C. Tang et al.[2] have developed a fused feature named Uniform Local Gabor Binary Pattern Histogram Sequence (ULGBPHS). Their model is tested with a self-build facial expression database with spontaneous expressions of learners and obtained significant results. Sathik et al.[3] believed that there is a strong connection between student's comprehension levels and the forehead wrinkles. The forehead has a crucial role in identifying facial expressions from the face image. Forehead block separated from the face image, and then the canny edge detection operator has applied to detect emotions using posed databases JAFFE and Yale.

The widely utilized six basic emotions are not probably suitable for e-learning environments; hence customization of more suitable sets of expressions is required. The most suitable database would be to construct our database. Lan Li et al.[4] described the emotional states of the learners as four categories:' confusion',' surprise',' confidence' and 'frustration'. K. P Rao et al. [5] proposed a multimodal emotion recognition approach to understanding students' comprehension levels using facial expressions, audio, and gestures. M. T Yang et al. [6] extracted the features corresponding to facial components: eye, brows, mouth, and horizontal/ vertical motions using the AdaBoost algorithm. Further, HMM (Hidden Markov Model) is applied to get six facial expressions: blink, nod, yawn, shake and talk. Finally, scores for the learning states such as interaction, understanding, and consciousness can be obtained using GMM (Gaussian Mixture Model). A. Gupta et al. [7] designed a new dataset DAiSEE for e-learning environments under wild conditions. Besides, a CNN classifier used to extract the learner's emotional state.

3. Methodology

The proposed system designed to combine features extracted from the convolutional neural network and handcrafted features to analyze students' cognitive states in e-learning environment. This system consists of 4 modules: image pre-processing, face detection and feature extraction, emotion recognition, and post-processing.

3.1. Hybrid Approach

The proposed system combines features extracted (engineered automatically) from CNN with the pose estimator features (engineered manually). Then these combined feature vectors will be trained over an SVM classifier. **Fig.2** shows the schematic diagram of the proposed system. Features extracted using CNN are appended to Handcrafted Features. The algorithm 1.1 describes the steps to implement the above procedure.

3.1.1. Pre-processing:

The original dataset consists of video sequences of the learners. The videos present in the dataset labeled to one of the four cognitive states: engagement, frustration, boredom, and confusion. In pre-processing, each video cut into frames; further, it resized to an image of size 48 x 48 pixels. Followed, by contrast, limited histogram equalization applied to enhance the image quality. The dataset was partitioned into a training set and validation set with a proportion of 80% and 20%.

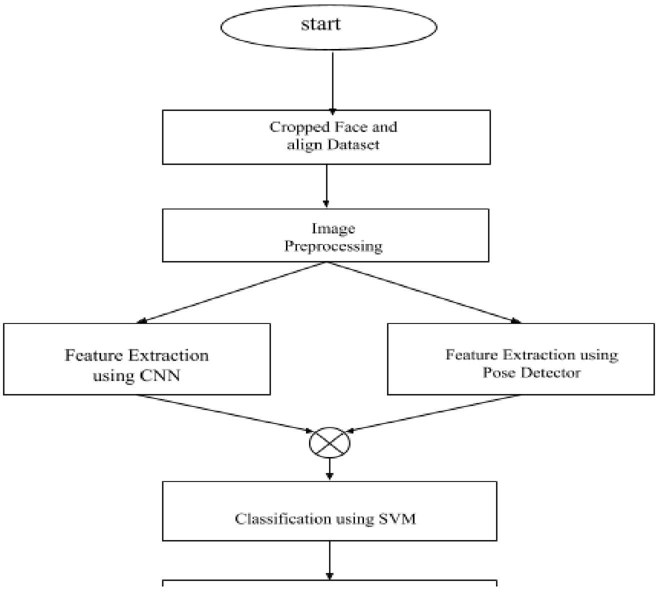
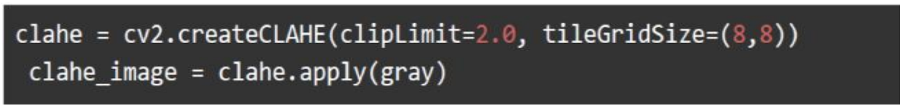

start
Cropped Face and align Dataset
Image Preprocessing
Feature Extraction using CNN
Feature Extraction using Pose Detector
Classification using SVM

Fig.2. Proposed hybrid model

(a) Histogram Equalization

Histogram Equalization is a method for adjusting the intensity of images to increase contrast. This is done by efficiently spacing out the most frequent values of intensity, i.e., by extending the image's intensity range. The idea of this processing is to give the resulting image a linear cumulative distribution function. We performed Contrast Limited Adaptive Histogram Equalization (CLAHE) using OpenCV package. CLAHE will use adaptive histogram equalization rather than the global histogram equalization. It indeed divides the given image into blocks then performs the equalization process.



```
clahe = cv2.createCLAHE(clipLimit=2.0, tileGridSize=(8,8))
 clahe_image = clahe.apply(gray)
```

$$cdf(x) = \sum_{k=-\infty}^{x} P(k), \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

$$S_k = (L-1)cdf(x) \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

3.1.2. Face detection and feature extraction

Face detection followed by features are extracted using two methods:

- Pose estimator using 68 facial landmarks (These are points on the face such as the corners of the mouth, along the eyebrows, on the eyes, etc.,)
- Conventional Neural Network

(a) Feature Extraction using pose estimator

The creation of the feature vector is the most critical part of a pattern classification problem. Essential features extracted from the preprocessed face image. Mainly Features are extracted using Dlib shape predictor or also called as landmark predictors, are used to predict (x,y)- coordinates of the given shape like eyes, eyebrows, nose, lips/mouth, jawline, and so forth. The feature vector formed with 68 landmarks extracted from the face image is shown in **Fig. 3**.

Input: Frames selected from each video sequence corresponding to the cognitive states.

Output: Confusion matrix.

Parameters: key facial points.

Algorithm

1. Build and compile the model with the proposed CNN Architecture as shown in ***Table 1.***
2. Load the model with required parameters, such as training and validation data.
3. Build the new model with the layers preceding to the fully connected layers as it is, and remove all other layers followed by FC.
4. Use this new model to extract features (engineered by CNN) from training set and validation set.
5. Extract the Handcrafted features (68 key facial points) with pose estimator.
6. Compute the distances, and tangent angles from the center point to every other feature point, and the corresponding coordinates are stored.
7. Append hand-crafted features to CNN features.
8. Model is trained with SVM classifier using combined features.
9. Build the confusion matrix to know the model accuracy

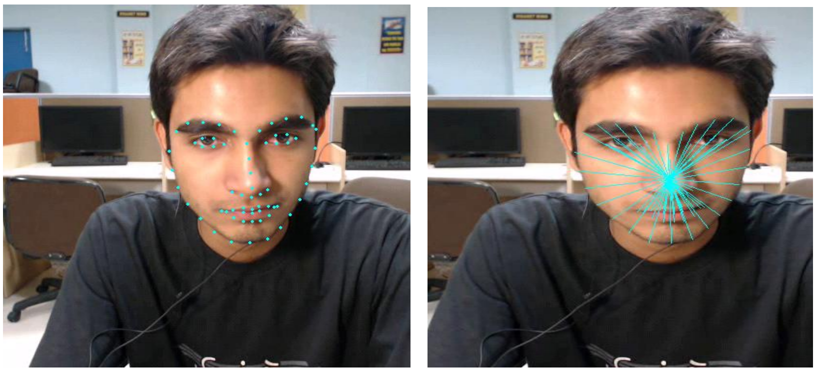


Fig.3. Facial landmarks (left), Distances measured between landmarks (right)

The mean point computed from all the 68 feature points. The distances, and tangent angles from the center point to every other feature point, and the corresponding coordinates are stored. These features collectively create a feature vector of size 272 (68 X 4) for a single image. The dataset contains 3219 images. The feature vector of size 875568 (272 x 3219) then trained with a Support Vector Machine (SVM) classifier [13]. The obtained accuracy of this model is 33.15% and the corresponding confusion matrix shown in **Fig.4.**

(b) Feature extraction using CNN

Convolutional neural networks (CNNs) are beneficial in areas such as image recognition and classification. CNN designed specifically to reorganize two-dimensional images with a high degree of invariance to illumination, translation, scaling, and other forms of distortion. A convolutional neural network consists of three layers, such as convolutional layers, pooling layers, and fully connected layers. The convolutional layer extracts the features using different types of filters, and further, these features sent to pooling layers, which reduces the size of an image. The fully connected layer flattens the features extracted from the convolutional/pooling layers to form a single vector of values representing the scores of a specific feature corresponding to a label.

Automatic detection of the essential features of the given image is a difficult task. The network learns new and complex features in convolutional and subsequent layers. The first layer's feature map detects the raw-information, such as horizontal and vertical edges. The next layers combine these features to classify shapes, and finally, the fully connected layers integrate this knowledge to predict the object.

In this paper, three versions of CNN models are designed and analysed. Initially, model is constructed with 16 layers. The structure of the CNN model is shown in Table 1.

3.1.3. Emotion recognition

CNN consists of two major components: feature extraction and classification. Initially, we used CNN models to extract features and classify them to respective class labels.

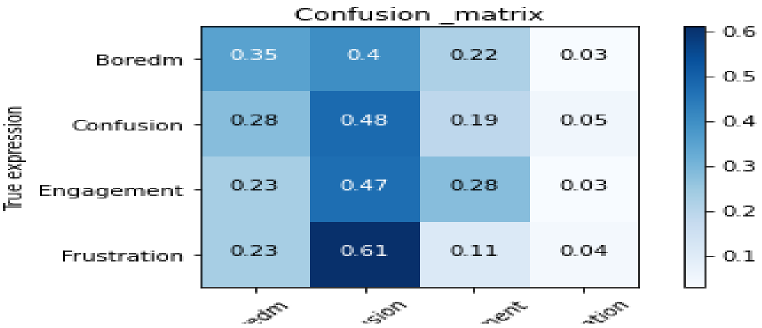


Fig.4. Confusion matrix for emotion recognition using Handcrafted features

Table.1. Proposed CNN architecture

Layer	Output Shape	Param #
conv2d_1 (Conv2D)	(None,46,46,32)	320
conv2d_2 (Conv2D)	(None,46,46,32)	9248
batch_normalization_1	(None,46,46,32)	128
max_pooling2d_1	(None,23,23,32)	0
conv2d_3 (Conv2D)	(None,23,23,64)	18496
batch_normalization_2	(None,23,23,64)	256
conv2d_4 (Conv2D)	(None,23,23,64)	36928
batch_normalization_3	(None,23,23,64)	256
max_pooling2d_3	(None,11,11,64)	0
flatten_1 (Flatten)	(None,7744)	0
dense_1 (Dense)	(None,128)	991360
dropout_1 (Dropout)	(None,128)	0
dense_2 (Dense)	(None,64)	8256
dropout_2 (Dropout)	(None,64)	0
dense_3 (Dense)	(None,4)	260

The architecture fed with an image of size 48 X 48 with a grayscale channel. It consists of 16 layers, as shown in **Table.1.** The fully-connected layer will provide a feature vector with 7744 features. They are sent to the softmax layer to classify them according to the labels. The overall performance of the system is 52.6%. The confusion matrix for each cognitive state has shown in **Fig.5.**

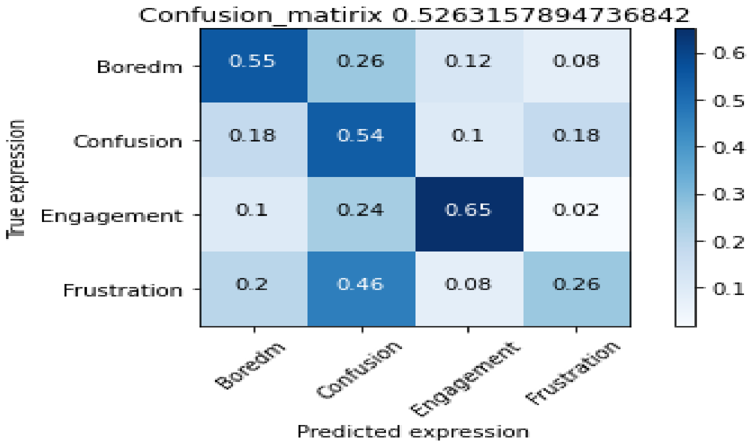


Fig.5. Confusion matrix for FER using CNN

3.1.4. Post processing

The same obtained model will be used to detect the learner's state in a real-time environment. These emotional states will be useful in analyzing learner mood and behavior. The engaged cognitive feeling will result in when the course contents are good, and there is no need to change the course contents. Boredom's cognitive mood will describe that the tutor has to change his course delivery process. The other two feelings (confusion and Frustration) will specify that the course has to be modified.

4. Experimental Results

The performance of the proposed system is evaluated using the DAiSEE dataset developed for the e-learning context and the state-of-the-art datasets such as the extended Cohn-Kanade (CK+) database[10], the Japanese Female Facial Expression (JAFFE) database[12], which are commonly used for facial expression recognition.

4.1. DAiSEE Dataset

Having enough labelled data of learners' facial expressions is a prerequisite in designing automatic facial expression recognition systems. However, the rare and less publicly available of the datasets to recognize learners' cognitive states severely limits the development of e-learning systems. DAiSEE is a publicly available dataset comprising 9068 video recordings captured "in the wild" environment from 112 learners. The learners' cognitive states are classified as engagement, boredom, confusion, and frustration. The dataset also has four different labels to represent the intensity of an emotion, namely - very low, low, high, and very high for each of the affective states.

Fig.6. Sample images from DAiSEE dataset

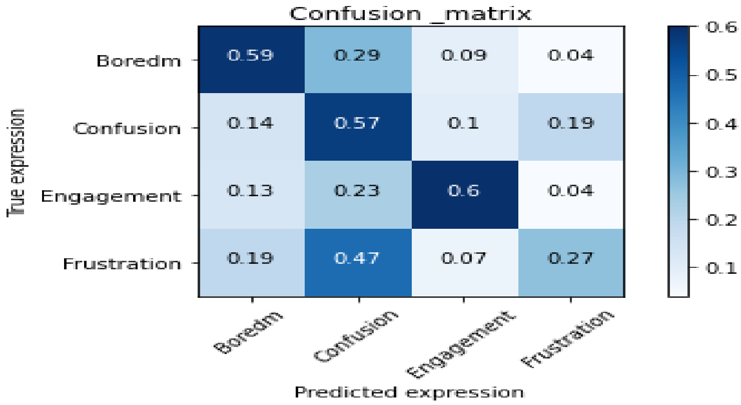


Fig.7. Confusion matrix for FER using hybrid CNN with DAiSEE dataset

4.2. Emotion recognition using hybrid CNN with DAiSEE

The last five layers from the architecture shown in Fig.2 have popped out (at the flatten layer) to extract the features. Further, these features combined with the geometric features (extracted manually using the Dlib shape predictor). The new feature vector is of size 8016 (7744+272) fed to the SVM and create hyperplanes with support vectors to classify the features. This model achieves an accuracy of 53.42%. The recognition rates for various cognitive states are shown in **Fig.7**. This work available in the Git-Hub repository [11].These results are satisfiable, and markable upto now in the e-learning field, since the dataset was collected in the wild (not posed) condition. The confusion matrix is drawn for accuracy.Fromthe results, it is observed that, Most of the Images with frustration cognitive state are predicted as confusion (Although the both emotion states will result in negative feedback). The recognition rate for the emotion/ cognitive state ‘frustration’ is very low.

4.3. Emotion recognition using hybrid CNN with CK+ dataset

The above proposed architecture will result in accuracy up to 100% for CK+ dataset. whereas, the CNN model will result in 98.97 % accuracy. The confusion matrix shown in **Fig.8** is drawn for our hybrid model to depict the false and true predictions.

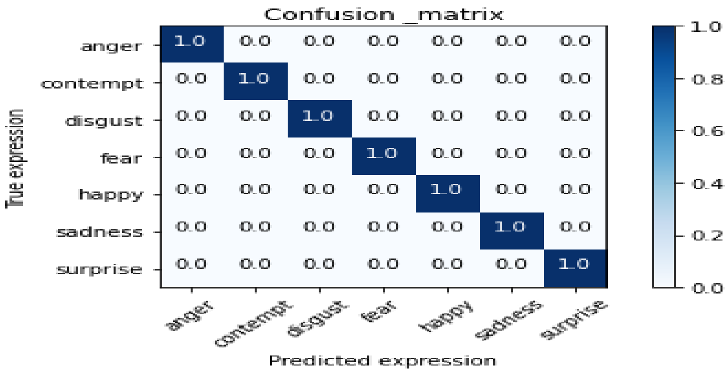


Fig.8. Confusion matrix for CK+

4.4. Emotion recognition using hybrid CNN with JAFFE dataset

The proposed hybrid architecture has got an accuracy of 71.42% for the JAFFE dataset. While, the CNN model has got slightly better performance i.e. 73.8% compared to the proposed system. The confusion matrix (**Fig.9**) is drawn for our hybrid model to show the false and true predictions.

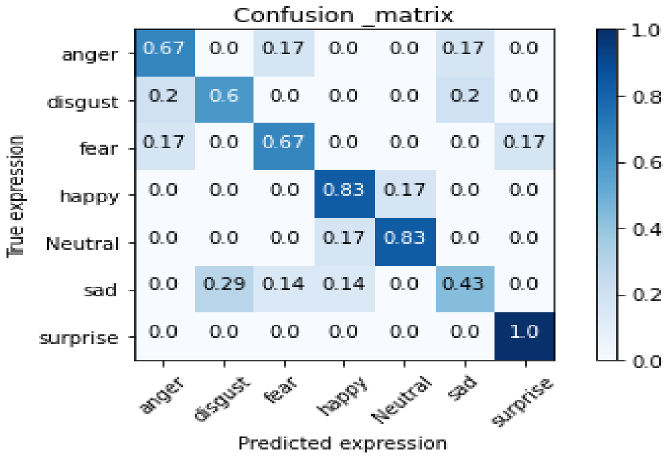


Fig.9. Confusion matrix of CNN with JAFFE

The results obtained from different datasets (DAiSEE, JAFFE and CK+) for CNN (with softmax), CNN(with SVM), Handcrafted and Hybrid models are analysed and illustrated in the **Table 2,**and **Fig. 10.**

Table.2 Performance comparison of various methods

Dataset	Handcrafted features with SVM classifier	CNN (softmax as classifier)	CNN (upto FC) with SVM as classifier	CNN (uptoFC)+ Handcarft _features with SVM classifier
DAiSEE	33.15	52.63	52.63	**53.42**
CK+	51.28	98.97	99.48	**99.95**
JAFFE	21.42	73.8	85.7	**71.4**

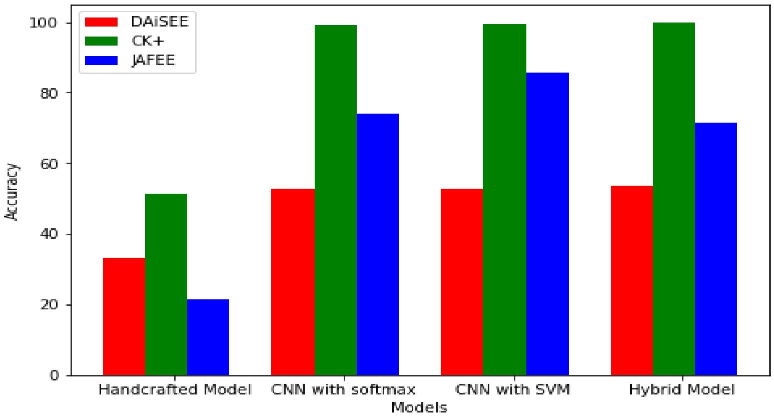


Fig.10. Performance comparison between CNN and handcraft

5. Conclusion

The automatic recognition of students' cognitive levels in e-learning environments is essential to send feedback to the instructor. Hence, instructors can adjust the delivery method according to the learners' cognitive states by speeding up, slowing down, or selecting new teaching methods. Due to the outbreak of the COVID-19 pandemic, millions of students across the world confined to homes. These e-learning systems are allowing learning from their home. The proposed hybrid CNN method got an accuracy of 51.9% on real-time (wild) e-learning dataset which is much better than the state-of-the-art methods.

This work can extend to multimodal information fusion to improve the emotion recognition rates i.e., the earner's audio, and gesture information combined with the visual data. Further, the data collected from the multiple sources such as assessment tests, online Quizzes, feedback, discussion forums can be incorporated to give more personalized teaching.

References:

[1] *Happy, S. L., et al. "Automated alertness and emotion detection for empathic feedback during e-Learning." 2013 IEEE Fifth International Conference on Technology for Education (t4e 2013). IEEE, 2013.*

[2] *Tang, Chuangao, et al. "Automatic Facial Expression Analysis of Students in Teaching Environments." Chinese Conference on Biometric Recognition. Springer, Cham, 2015.*

[3] *Sathik, M. Mohamed, and G. Sofia. "Identification of student comprehension using forehead wrinkles." 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET). IEEE, 2011.*

[4] *Li, Lan, Li Cheng, and Kun-xi Qian. "An e-learning system model based on affective computing." 2008 international conference on cyberworlds. IEEE, 2008.*

[5] *Rao, K. Prasada, MVP Chandra Sekhara Rao, and N. Hemanth Chowdary. "An integrated approach to emotion recognition and gender classification." Journal of Visual Communication and Image Representation 60 (2019): 339-345.*

[6] *Yang, Mau-Tsuen, Yi-Ju Cheng, and Ya-Chun Shih. "Facial expression recognition for learning status analysis." International Conference on Human-Computer Interaction. Springer, Berlin, Heidelberg, 2011.*

[7] *Gupta, Abhay, et al. "Daisee: Towards user engagement recognition in the wild." arXiv preprint arXiv:1609.01885 (2016).*

[8] *K. P. Rao, et.al Assessment of Students' Comprehension using Multi-modal Emotion Recognition in E-learning environments, Journal of Advanced Research in Dynamical and Control Systems, Vol.10, pages: 767-773.,2018.*

[9] *A. Hicken, "2016 Elearning Hype Curve Predictions", 23.12.2015.[Online].Available: http://www.webcourseworks.com/2016-elearning-hype-curve-predictions/. [Accessed 06.07.2017].*

[10] *P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotionspecified expression," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern RecognitionWorkshops. IEEE, 2010, pp. 94–101.*

[11] *chandrasekhar- https://github.com/chandrasekhar36/FER-for-E-Environment*

[12] *Kanade, Takeo, Jeffrey F. Cohn, and Yingli Tian. "Comprehensive database for facial expression analysis." Proceedings Fourth IEEE International Conference on Automatic Face and Gesture recognition (Cat. No. PR00580). IEEE, 2000.*

[13] *] Canedo, Daniel, and António JR Neves. "Facial Expression Recognition Using Computer Vision: A Systematic Review." Applied Sciences 9.21 (2019): 4678.*