An Approach to Develop the Information Providing System by Recognizing Facial Expressions

Kumar P
Department of CSE
Rajalakshmi Engineering College
Chennai, India
kumar@rajalakshmi.edu.in

Senthil Pandi S
Department of CSE
Rajalakshmi Engineering College
Chennai, India
mailtosenthil.ks@gmail.com

Pratheepa R
Department of CSE
REC, Chennai, India
210701192@rajalakshmi.edu.in

Mannuru Shreeya
Department of CSE
REC, Chennai, India
210701148@rajalakshmi.edu.in

***Abstract*— Facial expressions are very important in understanding one's state of mind. We aim to incorporate this notion in our learning assistant application in order to provide a rich user experience. The facial expression of the user is recognized with the help of a model trained on the FER-2013 dataset using a convolutional neural network. A query from the user is matched with an appropriate title from the list of titles released by Wikipedia using a matching algorithm based on the percentage of words that match in both the query and a title. The information under the title that achieves the maximum score is retrieved and provided for the user. Simultaneously, the facial expressions of the user are captured so as to analyze the level of understanding of the content for the user and display suitable responses. The percentage of correct predictions given by the proposed model on 7178 unseen images is 65.11%.**

Keywords— Face Expression, Expression Recognition, Information System, Attention Mechanism.

I. INTRODUCTION

In today's world and era of growing technology, self-learning in the online mode is increasing every day. Various kinds of resources are made accessible to people all across the world. Wikipedia serves as a huge resource for information about a lot of things and it releases a copy of all the titles and the available content in the form of dumps regularly. This project uses the list of titles to match the query of the user with the appropriate title and display the needed information. Facial expressions are a very important index in understanding the mindset of a person. A human teacher takes all efforts to also look at the facial expressions of the students apart from delivering content in order to modify the lecture and make the students understand better. Likewise, recognizing facial expressions while delivering information can prove to be a very important tool in improving the learning process for the user.

The image is received through a webcam that turns on while the information is being displayed. The quantity of data used to train the expression recognizing model is increased by introducing variations in the existing data in order to avoid class imbalance and also make the model generalize well and perform better in unseen situations. The model is trained on the increased data of around 50000 images using a convolutional neural network along with methods deployed to prevent the model from learning to memorize the properties of the data that it is being trained on. The model classifies any image into one of the seven categories: angry, happy, disgust, sad, surprise, fear, and neutral. The system gives less weightage to the predictions made in the first half of the time period spent in capturing the frames from the webcam, and gives more weightage to the predictions made in the second half. The final prediction is decided by looking at which category has accumulated the maximum number of points that is calculated by summing the weights for each prediction. Suitable responses are displayed for the user based on the final prediction.

II. LITERATURE SURVEY

In the paper [1], the enhanced version of the ResNet-50 model incorporates a Convolutional Block Attention Mechanism (CBAM) to improve expression recognition accuracy. The algorithm in uses a pre-trained machine learning model to recognize six different emotions. A laboratory was set up to capture facial expressions during practical sessions, with images captured in batches of 350 per student. 7000 images were collected, corresponding to six categories: commitment, boring, frustration, focused, interested, and neutral. The work in proposes a lightweight emotion recognition model based on DenseNet architecture, which uses highly linked convolution layers and model compression methods. The proposed model in contains three stages: pre-processing of data, extraction of features, and classification. The ResNet network is replaced in the first half of the VGG network, and the classification layer uses the Softmax function to give the prediction. The research in [2] presents a model for recognizing emotions using Multi-task cascaded convolutional networks (MTCNN) detection, trained using the ADAM optimizer. A FER algorithm for human-computer interaction systems is designed in to enable robots understand facial expressions like humans. The pruning strategy that is proposed has good model acceleration, reducing memory usage and improving classification accuracy. The regenerative Generative Adversarial Network (GAN) is used in [3] to recreate obscured areas of students' faces, essential for precise emotional interpretation. The dashboard offers a user-friendly interface for educators, with visualizations associated with different emotions. Teachers can access real-

time feedback and historical data monitoring to adjust their teaching methods. The Convolutional Autoencoder (AE) is the proposed method in [4] for recognizing partial occlusions in images caused by objects like sunglasses and hands. The encoder's transpose version is the decoder, which reconstructs the original image using the latent space. The AE is used to extract features and a Multilayer Perceptron (MLP) that is a linear classifier is trained, which classifies positive and negative emotions. The Fuzzy C-Means clustering technique is used in for facial detection using the Haar-Like algorithm. Feature extraction is performed using a trained CNN model, and the retrieved characteristics are used for emotion classification using a Support Vector Machine (SVM). The paper [5] introduces a web scraping approach called "UzunExt" that improves the efficiency of extracting content from web pages. UzunExt uses string searching techniques to locate the start and end tags of the desired content, achieving faster content extraction than traditional methods. The paper [6] presents a recurrent neural network using long short term memory algorithm to predict the most plausible outcome for a query. The query is loaded into a JSON document, tokenized, and converted into numeric configuration. Keras is used for creating a sequential model for a deep convolutional neural network trained on the FER 2013 dataset. The paper [7] discusses the various facial expression identification methods used in educational research. A lightweight CNN model for face recognition is trained in using video stills of faces. The model's training process includes a separate process for face recognition and fine-tuning for emotion categorization. A study in [8] involving nine video lectures revealed that students' engagement levels are crucial for improving learning outcomes. Participants were instructed to press a key when they noticed an auditory target, with an average loudness of 70 db and white noise of 0.66 db. Facial features were recorded during the experiment, and it was found that deviation from the lecture can be understood by increased nose wrinkling. Facial expressions can predict learners' attention levels, with video camera-captured facial features predicting reaction times (RTs), which are representative of attentional states. The Naive Bayes classification model is utilized in due to its effectiveness. It uses segment embedding and token embedding to process questions, while BERT, a pre-trained model, responds to user requests using incoming information. The scraping process uses BeautifulSoup (BS4) and Selenium Web Driver. The paper presents a three-channel convolutional neural network to increase the accuracy of facial expression identification. Augmentation techniques like random horizontal flipping and brightness adjustment were applied to increase the volume of data. The study in [9] introduces a divide and conquer convolutional neural network (CNN) learning approach to improve facial expression recognition accuracy. The approach involves identifying, focusing, and normalizing face regions, and adjusting the ResNet-18 model to accommodate modifications like lowering filter sizes and extracting more relevant information from face photos. In the paper, a FreNet model is a lightweight deep learning model that uses the Discrete Cosine Transform (DCT) to convert images into the frequency domain. This process focuses on capturing low-frequency components, which contain critical information about the image. The methodology outlined in [10] focuses on multi-hop open-domain question answering (QA) by using both structured and unstructured information from Wikipedia. The study presents a new retrieval technique named HopRetriever, intended to collect scattered reasoning evidence from several sources. It involves defining a "hop", which is the combination of a hyperlink and its corresponding outbound content. The system systematically acquires additional documents by traversing a sequence of "hops" between them, enhancing the evidence compilation at each stage. This enables HopRetriever to identify relevant documents that aid in addressing intricate inquiries. The paper presents a graph convolutional neural network designed for recognizing micro-expressions. It uses Facial Action Units (FAUs) to analyze detailed changes in facial expressions, and the Optical Flow Method (OFM) to estimate pixel motion in facial images.

III. PROPOSED MODEL

The proposed system mainly contains two components: (A) Information Retrieval and (B) Recognizing Facial Expressions. We will discuss these two components in detail elucidating on the steps adopted to build them.

Information Retrieval: Our work aims to handle queries related to providing information about any given topic. Wikipedia acts as a huge resource of information for a wide variety of topics and it also releases the list of titles and all the available content in the form of dumps in a regular manner. The list of titles released can be used to perform a matching algorithm between the user query and each title so as to find what information the user is seeking. The user query might be in natural language indicating the needs and might not always contain the exact order of words as present in the actual title. So, performing an algorithm to match the query with the actual title as present in the information resource is necessary to retrieve the content that the user is asking for. The algorithm traverses through the titles to compute a score for each of them by comparing the user query and each title. The title that attains the highest score is chosen indicating that it matches the user query to the maximum extent and conveying that the user has asked for this specific information. The score is calculated by computing the average of the fraction of words of the title that are present in the query and the fraction of words of the query that are present in the title. The formula is given by:

$$score = (a/b + c/d) / 2$$

where 'score' indicates the value of similarity between the user query and a title, 'a' indicates the count of words in the title that are present in the query, 'b' indicates the count of words in the title, 'c' indicates the count of words in the query that are present in the title, and 'd' indicates the count of words in the query.

The algorithm only traverses through the subset of the list of titles that begin with the starting alphabets of all the words in the user query to reduce the time taken to retrieve the content and improve the efficiency. The content is displayed letter by letter with a time interval so as to create a situation of capturing the facial expressions of the user while reading through the provided information. Displaying the information and recording the facial expressions are implemented as two different threads that execute simultaneously. While the

content is getting printed, the frames are captured from the webcam at a regular interval of 2 seconds and the expression predictions for each frame are stored.

Recognizing Facial Expressions: A deep neural network model was trained on the FER-2013 dataset using a convolutional neural network to identify facial expressions. The FER dataset has a lot of positive aspects on having images that cover facial expressions of people from different angles and containing images for seven different classes: angry, happy, disgust, sad, surprise, fear, and neutral. The images in the dataset are of shape (48,48,3) with height of 48 pixels, width of 48 pixels, and three channels defining the color of a pixel using the values of red, green, and blue. All the three values of a pixel are identical leading to every image in the dataset appearing with only shades of gray and not containing any other colors. One major issue that needs to be addressed before training any classification model is ensuring class balance. The 'happy' class contains around 7200 images whereas the 'disgust' class contains only around 400 images. This huge difference might make the model not completely capture or learn the features of the 'disgust' class and also might make the predictions to be more biased towards the 'happy' class. Hence, some methods are used to increase the amount of data in the unbalanced classes to ensure that the model learns the features of every class by giving equal importance to all the categories of expressions. The methods used to increase data in the unbalanced classes that have less images are:

Zoom Out: All the images in the dataset contain only the faces of humans and do not have any background parts. But the images in real-time would contain background parts which might make the model not to be effectively trained against those situations. Therefore, introducing images with some background parts in the training phase is necessary to train the model well to adapt to various situations. An image is resized to half its size and a random value is chosen in the range of [0, 255]. The random number is decided as the value of all the three channels that form a pixel and a new image is created of the original size of an image in the dataset with all the pixels of the same values. Then, the resized image is placed onto the center of the new image thus forming an image that seems to contain a background apart from the face.

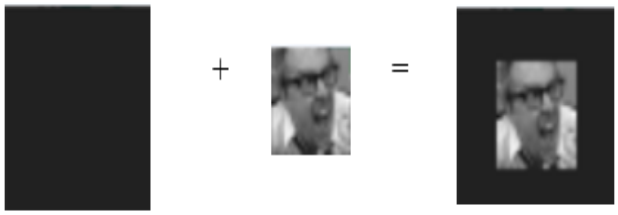

Fig.1 Model Data Augumentation

(ii) Adjusting Brightness: The images that would be delivered for prediction in real-time would be having different brightness levels. Therefore, training images with varying brightness levels is necessary to accommodate all the situations. A random value is chosen in the range of [-127, 127]. This value is added to all the pixels in the image to produce a modified image of a different brightness level. If the values of the pixel go out of the range of [0, 255], they are clipped to the nearest extreme value.

(iii) Adjusting Contrast: Contrast of an image is the difference between the dark regions and the light regions. If the difference is more, the image is said to have a high contrast and the objects would be visible well against the background. If the difference is less, the image is said to have a low contrast. The images given in real-time can have varying contrast. Therefore, including images of varying contrast in the training phase would help the model to handle different situations. A random value is chosen in the range of [0.1, 2.0]. All the pixel values in an image are multiplied by this random value to get different pixel values that may have higher contrast or lower contrast based on the random value chosen. A value greater than 1.0 would increase the contrast as it would widen the difference between pixel values. A value lesser than 1.0 would lead to an image with a low contrast as it would narrow the difference between pixel values. If any value goes out of the range of [0, 255], then it is clipped back to the nearest extreme value. These three techniques not only increase the amount of data to attain class balance, but also introduces variations in the training data that makes the model handle varied situations. After application of these techniques to the images present in the unbalanced classes, each class contains around 7200 images with the total images in the dataset to be 50613.

Fig.2 Model Contrast Adjustment

A model is trained on this entire dataset using a convolutional neural network that would learn the biases, the weights, and the filter values of the network in order to provide a prediction for any new image. The architecture of the neural network contains convolution layers that uses filters to span through the image with a defined kernel size, compute the dot product of the filter values and the pixel values, and use the ReLU activation function to help the model learn the relationships between the input and the output that are not linear, and produce feature maps of multiple channels. The count of channels in a layer are defined by the count of filters used in the previous layer. The training is done in batches in a single epoch. A batch size of 32 is used during the training process. Batch normalization layers are used to make the data in a particular channel across all the images of a batch to have a variance of 1 and a mean of 0. Max-pooling layers are used to capture the most important features in a generated feature map and also reduce the number of features so as to prevent the model from just memorizing the data on which it was trained. If the model learns a large number of features, it will likely tend to learn only the features in the provided data on which it is trained and might not be able to generalize well for unseen situations. Dropout layers are used to drop some percentage of the neurons in some of the layers to prevent the model from relying only on some specific features to decide

the prediction. Residual blocks are used to perform a transformation of a layer using convolution and batch normalization layers and add the initial layer to the transformed layer to provide the final layer. This helps in creating a shortcut and passing the information from the earlier layers to bypass one or more layers and feed the information into later layers. The layer for convolution operation, the layer for normalizing the extracted feature values, residual block layer, the layer to capture the important features, and a dropout layer is considered to be a block and four such blocks are used in succession. It is followed by a global average pooling layer that computes the mean value of all the values in each feature map and generates a one-dimensional vector that then proceeds onto the dense layers. The dense layers also have batch normalization layers and dropout layers used along with them. The final layer is the layer of seven neurons where each neuron predicts the probability of each possible class using the softmax function. The final output is returned as a vector of seven values with each value denoting the probability of a class. The class that has the maximum probability is considered to be the resulting prediction.

The maximum number of epochs is decided to be 60 and an early stopping method is used to stop the process of training when the model starts to overfit the data on which it is being trained. The data is separated into 80% training data and 20% validation data by training on the training data and monitoring the loss values and the accuracy values of the validation data. If the loss on the data used for validation does not decrease after consecutive 10 epochs, the process of training is halted as it means that the model has started to memorize or overfit the data on which it is getting trained and might not be able to give accurate predictions on unseen data. If the loss on the data that is used for validation does not decrease for consecutive 5 epochs, it is understood that the model is approaching a minimum point, and therefore, the learning rate is reduced by a factor of 0.1 so that the further weight updates would be minimal and would not lead to any erratic updates, which would make the model converge at the minimum point. A combination of early stopping and adaptive learning rate methods is used to stop the model from overfitting and also help the model converge at the minimum point. Adam optimizer is used to update the biases, the weights, and the filter values of the network.

This trained model is used to predict the facial expressions of the image captured from the webcam every 2 seconds while displaying the retrieved information. The predictions are stored in the form of an array. The first half of the predictions are given a weightage of 0.5 and the second half of the predictions are given a weightage of 1.0.

The total weightage value for each class is computed so as to find the final prediction. The class that gets the maximum weightage denotes the predominant expression of the user and is considered to be the final prediction of the expression of the user while reading through the content.

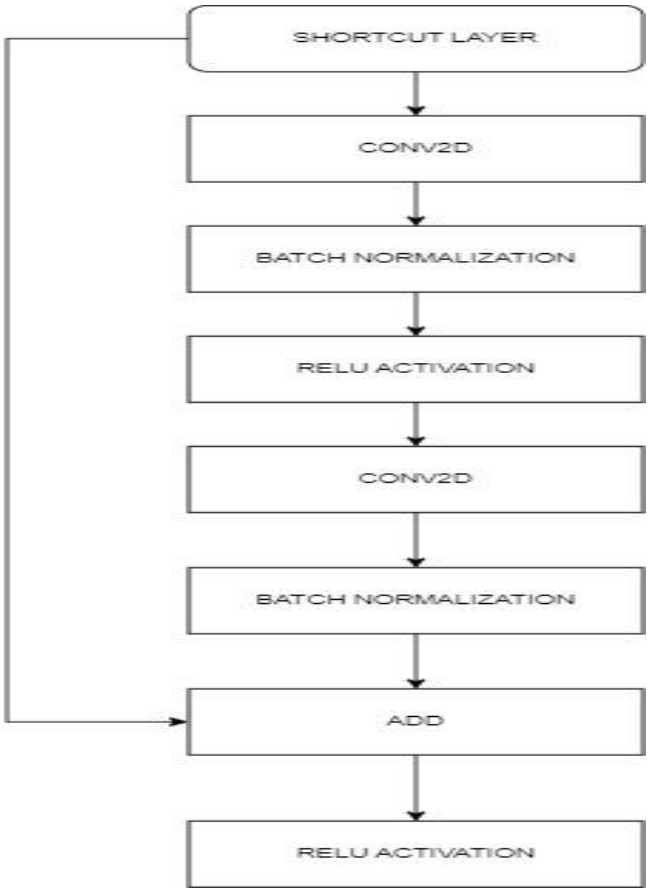

Fig.3 Proposed Model Work Flow

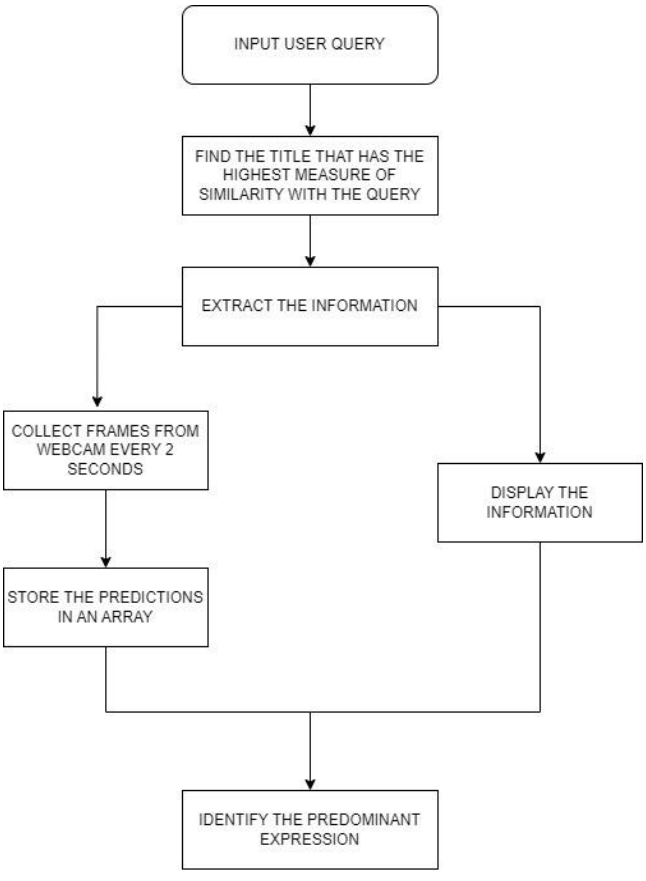

Fig.4. Model Architecture

IV. RESULT

A model was trained using a convolutional neural network before performing data augmentation on the initial 28709 training images from the FER-2013 dataset. The model recorded an accuracy of 44.10% with literally no predictions for the ‘disgust’ class due to the low count of images present in that category. There were only around 400 images present for the ‘disgust’ class against 7200 images for the ‘happy’ class, leading to a great imbalance. Figure 5 represents the confusion matrix of the above model.

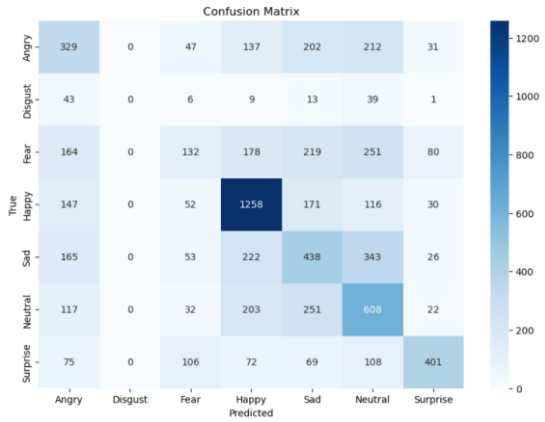

Figure 5. Confusion matrix of the model trained prior to data augmentation

The amount of data was increased by introducing variations to the present data in the form of zoom out operation, adjusting brightness and contrast of the images and creating new images. This led to every class containing around 7200 images leading to class balance. Then, a model was trained on this increased dataset of 50613 images.

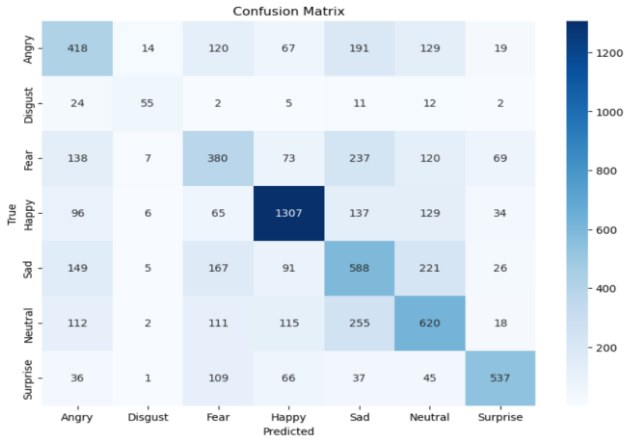

Figure 6. Confusion matrix of the model trained after data augmentation

Figure 6 depicts the confusion matrix of the above model which was trained after increasing the size of the dataset and attaining class balance. Now, there were images that were classified as the ‘disgust’ class and there were improvements in the correct predictions of the other classes as well, leading to the percentage of correct predictions to be 54.40%.

The above model recorded an accuracy of 95% on the training data whereas recorded much lower accuracy on the test data, indicating overfitting. Therefore, dropout layers were introduced in the fully connected layers to tackle the memorization of the data on which the model is getting trained and the accuracy of the further trained model improved to 57.99%. Figure 7 depicts the confusion matrix of the above model which was trained after introducing the dropout regularization technique.

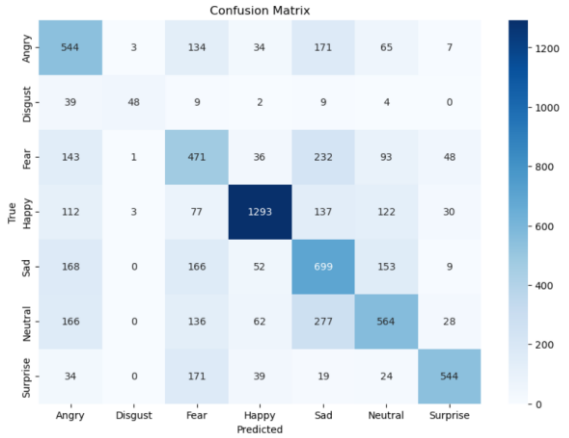

Figure 7. Confusion matrix of the model after introducing dropout layers

Then, residual blocks were introduced to make the information at a layer bypass one or more layers and use it at later layers. Dropout layers were introduced both in the convolutional part and the fully connected part.

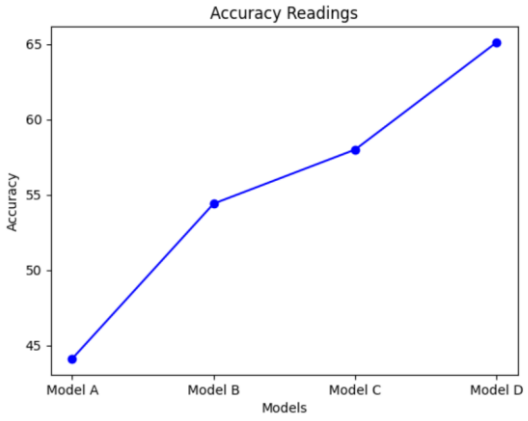

Figure 8. Accuracy readings of different models

The number of filters at each layer in the convolution part were made to increase sequentially which would lead to more and more features being captured in a hierarchical manner. Early stopping method was deployed to further tackle optimization by stopping the process of training when the model started to overfit the data on which it is being trained. The learning rate was reduced whenever the model moves on a plateau region in the loss function which would lead to effective convergence. These were the changes made to attain the proposed model that recorded an accuracy of 65.11%.

Figure 8 shows the different accuracy readings that were recorded in training multiple models which eventually led to the proposed model. Model A denotes the model that was trained before data augmentation. Model B denotes the model that was trained after increasing the size of the dataset to attain class balance. After observing overfitting, model C was trained after including dropout layers in the fully connected part of the model. Model D denotes the proposed model that constitutes residual blocks, dropout layers in both the convolutional part and the fully connected part, early stopping method, and adaptive learning rate method that reduces learning rate in plateau regions to help the model converge better.

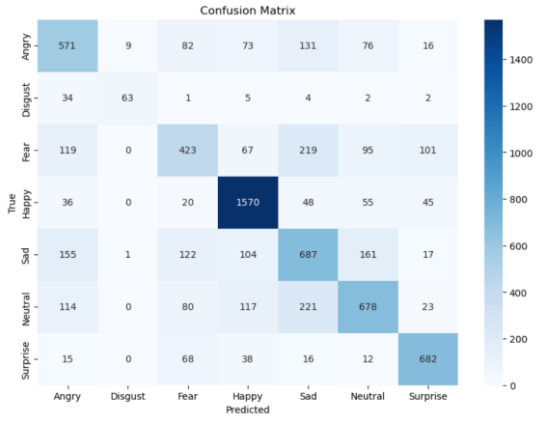

Figure 9. Confusion matrix of the proposed model

The proposed model has recorded an accuracy of 65.11% in recognizing facial expressions from 7178 unseen images. Figure 9 denotes the confusion matrix of the proposed model on the unseen images.

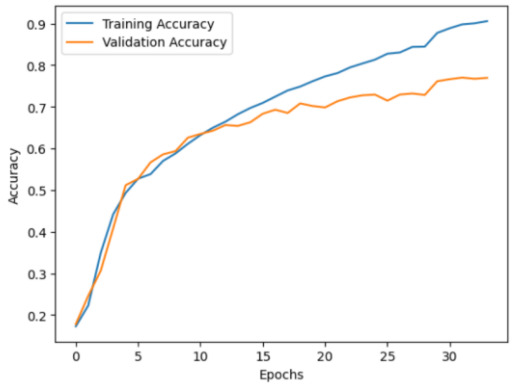

Figure 10. Trends of accuracy on the training data and validation data during training phase

The training dataset is separated into 80% training data and 20% validation data with the training being performed on the training data and the accuracy and loss being monitored on the unseen validation data. Figure 10 shows the trends of the accuracy obtained on the training data and the validation data throughout the training period. Initially both the training accuracy and validation accuracy seem to be of the same values with the validation accuracy even recording a higher accuracy at some stages. As the training progresses, the model starts to extensively learn the training data and therefore, the training accuracy keeps on increasing than the validation accuracy. The training stops with 34 epochs as the loss on the validation data does not decrease for consecutive 10 epochs and the model is saved with the best weights that have been achieved so far to get to the minimum possible loss value.

V. CONCLUSION

The proposed system is an information providing system that not only provides the required information for the user but also recognizes the facial expressions of the user while reading through the provided content. This information about the predominant expression of the user throughout the learning process is used to provide suitable responses and can also be used for the modification of the provided content so as to cater to the needs of the user. This system provides a rich user experience and also aims to provide an effective learning process for the user.

REFERENCES

[1] M. Aly, A. Ghallab and I. S. Fathi, "Enhancing Facial Expression Recognition System in Online Learning Context Using Efficient Deep Learning Model," in IEEE Access, vol. 11, pp. 121419-121433, 2023, doi: 10.1109/ACCESS.2023.3325407.

[2] W. E. Villegas-Ch, J. García-Ortiz and S. Sánchez-Viteri, "Identification of Emotions From Facial Gestures in a Teaching Environment With the Use of Machine Learning Techniques," in IEEE Access, vol.11,pp.38010-38022, 2023, doi: 10.1109/ACCESS. 2023.3267007.

[3] G. Zhao, H. Yang and M. Yu, "Expression Recognition Method Based on a Lightweight Convolutional Neural Network," in IEEE Access, vol. 8, pp.38528-38537, 2020, doi: 10.1109/ACCESS.202 0.2964752.

[4] J. Liu, H. Wang and Y. Feng, "An End-to-End Deep Model With Discriminative Facial Features for Facial Expression Recognition," in IEEE Access, vol. 9, pp.12158-12166, 2021, doi: 10.1109/ACCESS.202 1.3051403.

[5] N. Zhou, R. Liang and W. Shi, "A Lightweight Convolutional Neural Network for Real-Time Facial Expression Detection," in IEEE Access, vol.9, pp. 5573-5584, 2021, doi: 10.1109/ACCESS.2020.30467 15.

[6] J. Pu and X. Nie, "Convolutional Channel Attentional Facial Expression Recognition Network and Its Application in Human–Computer Interaction," in IEEE Access, vol. 11, pp. 129412-129424, 2023, doi: 10.1109/ACCESS.2023.3333381.

[7] P. Ganesan, S. Kumar Jagatheesaperumal, I. Gobhinath, V. Venkatraman, S. N. Gaftandzhieva and R. Z. Doneva, "Deep Learning-Based Interactive Dashboard for Enhancing Online Classroom Experience Through Student Emotion Analysis," in IEEE Access, vol. 12, pp. 91140-91153, 2024, doi: 10.1109/ACCESS.2024.3421282.

[8] M. D'incà, C. Beyan, R. Niewiadomski, S. Barattin and N. Sebe, "Unleashing the Transferability Power of Unsupervised Pre-Training for Emotion Recognition in Masked and Unmasked Facial Images," in IEEE Access, vol. 11, pp. 90876-90890, 2023, doi: 10.1109/ACCESS.2023.3308047.

[9] M. Shi, L. Xu and X. Chen, "A Novel Facial Expression Intelligent Recognition Method Using Improved Convolutional Neural Network," in IEEE Access, vol. 8, pp. 57606-57614, 2020, doi: 10.1109/ACCESS.2020.2982286.

[10] E. Uzun, "A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages," in IEEE Access, vol. 8, pp. 61726-61740, 2020, doi: 10.1109/ACCESS.2020.2984503.