Facial Emotion Prediction Model using Cascade of Neural Network - A Machine Learning Perspective

Yogendra Narayan Prajapati[1], Dr. Khel Prakash Jayant[2], Mili Srivastava[3], Dr. Pramod Kumar Sagar[4], Dr. Pushpendra Kumar Verma[5], Manu Singh[6]

[1]Department of CSE, Ajay Kumar Garg Engineering College, Ghaziabad, India.
[2]Professor, CSE Department, Raj Kumar Goel Institute of Technology, Ghaziabad, Uttar Pradesh, INDIA,
[3]IT department, Ajay Kumar Garg Engineering College Ghaziabad,
[4]Associate Professor, CSE Department, Raj Kumar Goel Institute of Technology, Ghaziabad, Uttar Pradesh, INDIA
[5]Associate Professor, SoCSA, IIMT University Meerut UP, India. Orcid_id: 0000-0003-2777-5626
[6]Department of CS, ABES Engineering College, Ghaziabad, India.
(ynp1581@gmail.com, kpjayant@gmail.com , srivastavamili@akgec.ac.in, pksagar1975@gmail.com, dr.pkverma81@gmail.com, drmanuajaykumar@gmail,com)

Abstract: In the last two decades, research into emotion recognition has been one of the most dynamic natures. Here, Convolutional Neural Network (CNN) and an unsupervised algorithm to classify is used human emotions from real-time monitoring of the users' emotional state while the image was being analyzed. To accomplish this, a real-time emotion identification system based on virtual markers has been implemented in this proposal. In our work, features are extracted, a subset of those features is created, and an emotions classifier is developed in three stages. The input image's identity, characteristic values, and facial features are determined with the support of the Haar Cascade approach. The VGG16 method places the virtual markers in specific positions on the recognized face. Cross-validation is then used to verify the components before sending them to the CNN classifiers. To prove the effectiveness of our work, we have compared it with traditional machine learning classifiers as well as advanced neural networks. Compared to existing methods, the suggested model performs better in experiments and yields results appropriate for real-time facial expressions with an accuracy of 91.03%

Keywords: Deep Learning, Internet of Things, Convolutional Neural Networks, Facial Landmark, VGG16.

1. INTRODUCTION

Recently, all around the globe, human emotions categorization and recognition gain a lot of awareness in the recent years. It is possible to determine the mental and emotional condition of a person by seeing the gestures on their faces [1]. Face detection, a fundamental component of emotion recognition systems, has seen various approaches, including the popular Haar cascade method. While widely adopted, the question arises: has there been innovative work to enhance face detection algorithms beyond the conventional models provided by libraries like OpenCV? When speaking, humans experience a wide range of emotions, which, due to differences in their depth and complexity, convey a variety of diverse meanings [2]. Anger, fear, happiness, neutrality, sadness, surprise, and disgust are the primary classifications that may be applied to human emotions. The combination of these fundamental states of mind may also give rise to the phenomenon of mixed emotions [3]. It is possible to narrow the communication gap between people and robots by learning to read facial expressions. It is applicable to a wide variety of different practices. Automatic Emotion Detection has received a lot of attention ever since the Internet of Things (IoT) was invented because different locations have smart environments, such as hospitals, cities, hotels, and smart cities, have placed a lot of emphasis on it. Now, there are many AI assisted technology are in place. As emotion recognition technologies continue to advance, it is imperative to acknowledge the ethical concerns that accompany their deployment. This article aims to shed light on the critical issues related to privacy, consent, and potential misuse in the context of emotion recognition systems.

Assistants being used all over the world, such as Alexa, Siri, and Cortana; all of these assistants are based on the NLP techniques that are used for communication with humans; however, after being augmented with emotions, the communication becomes more effective [4, 5]. FER has a wide variety of applications, one of which is the ability to spot anti-social components in large groups. In the event of road rage, the department in charge of traffic control may do fast analysis of conflict scenarios by identifying furious faces and then dispatch teams to diffuse the situation or take other appropriate action. One such example is the monitoring of the alertness level of drivers in real time. In addition to this, it may assist medical professionals in the analysis of synthetic human expressions and the diagnosis of mental diseases [6]. FER may be used in a array of contexts, as well as video surveillance and security systems, as well as automatic face recognition [7-15].

While expression recognition using Convolutional Neural Networks (CNNs) is a well-established endeavor, this article introduces innovative approaches that significantly enhance the accuracy and robustness of the recognition process. The following sections outline the key advancements and advantages of our proposed model over existing methods.

This article places a strong emphasis on the real-time applicability of the proposed emotion recognition model. It is crucial to evaluate the model's performance under diverse and dynamic conditions that are reflective of real-world scenarios. As such, the evaluation process will include an assessment of the model's responsiveness and efficiency in the presence of factors such as varied lighting conditions, facial orientations, and environmental variables.

Researchers have, up to this point of time, offered a wide range of strategies that make use of a number of DL and ML techniques in order to complete the job and achieve higher levels of accuracy. Some of them are worthy of being mentioned in this context, such as the KNN algorithm, RF algorithm, and the SVM. In order to teach the system to recognize and classify emotions automatically, we need data. FER2013, FER2013+, JAFEE, Cohn-Kanade (CK), and Extended Cohn-Kanade (CK+), DEAP, MMI, EmotioNet, etc. are some examples of publicly accessible datasets that are included in [16].

This paper proposes a Facial Landmark and Emotion Detection System that works on the architecture of the CNN. The used CNN model comprises seven layers to get a large number of extracted features. Feature extraction is directly proportional to the achieved accuracy because a larger dataset can accurately match the training data, giving better accuracy. We used the FER2013 standard dataset to train and validate the model, and Haar Cascade is used for real-time face detection. Michael Jones and Paul Viola used the Haar feature-based cascade classifier for object detection [17]. It is used for detecting objects in various images. Harr cascade has various object detectors, left & right eyes, mouth, nose, and face. In this paper, we are working with Harr Cascade for face detection.

The focus of the proposed work is as follows:

- A survey of the relevant literature within the framework of current work done in the area of emotion detection in current years.
- The development of a system that can identify emotions in real time.
- Calculation of the amount of data that was lost and the accuracy of the model.
- The first stages of the formation of the training confusion matrix.
- Establishment of a testing matrix for confusion.
- Creating classification reports for use in both training and testing using the dataset
- The representation of the outcomes.

Section 2 of the article continues with a presentation of the supplementary materials. In Section 3 we explain the methodology and the recommended technique. In Section 4, the experimental setup used to assess the model is detailed, and in Section 5, the results and their interpretation are presented before the study is wrapped up.

Table 1: Abbreviations used in this paper

Symbol	Name
HOG	Histogram of Oriented Gradients
PCA	Principal Component Analysis
GPU	Graphical Processing Unit

RF	Random Forest
FER	Facial Emotion Recognition
CK	Cohn-Kanade
ANN	Artificial Neural Network
CK+	Extended Cohn-Kanade
IoT	Internet of Things
CNN	Convolutional Neural Network
AAM	Active Appearance Model
KNN	k-Nearest Neighbor
ICML	International Conference on Machine Learning
ReLU	Rectifier Linear Unit
SVM	Support Vector Machine
LBP	Local Binary Pattern
NMF	Negative Matrix Factorization
LBP	Local Binary Pattern
LSTM	Long-Short-Term-Memory
ROI	Region of Interest
DL	Deep Learning
LR	Logistic Regression
DNN	Deep Neural Network
IPAs	Intelligent Personal Assistants
MLP	Multi-Layer Perceptron

2. LITERATURE REVIEW

The basic categories of human emotions that reflect as facial expression are anger, fear, disgust, sad, happiness, neutral and surprise. Data acquisition is a primary step in building a classification model. The availability of good quality data is a key factor in achieving satisfactory results. Some of the publicly available datasets are:

CK+: IT is a labeled dataset divided into seven basic categories. It consists of 593 sequences from 123 subjects having pixel dimensions of 640x490.

EmotioNet: It comprises one million pictures saved from the internet, which are labeled with twenty-three basic or compound emotion categories.

FER 2013: Created during ICML 2013, comprises 35887, (48x48) pictures labeled with seven basic emotion categories.

MMI: This database is an ongoing project, which currently comprises approximately still images of seventy people and approximately 2900 videos.

This dataset has around 219 still images of Japanese females which are labeled with 6 basic facial expressions.

A short overview of the methods used in FER was presented by reference cited in [18]. These approaches may be unevenly divides into two classes: traditional FER approaches and deep learning-based FER methods. Facial part detection, classification and feature extraction are the three components of the former method. The latter method permits full-fledged education based on the original graphics. In order to remember the differences between successive frames, a hybrid solution was presented, which combine a Convolutional Neural Network (CNN) for the spatial information with a Long Short-Term Memory (LSTM) for the chronological data. Long short- term memory, or LSTM, is organized in a chainlike fashion to address the problem of reliance over the long term by way of the short term memory. When it comes to input and output, LSTM is flexible. When combined with other models, like CNN, they allow for easy end-to-end fine tweaking.

Expression recognition using Convolutional Neural Networks (CNNs) has been widely explored in the literature. While existing methods have demonstrated success, our work distinguishes itself through innovations that overcome certain limitations. Specifically, we address the current limitations, e.g., handling subtle expressions, robustness to noise, etc. which provide a more effective solution for expression recognition task.

In the realm of face detection, the Haar cascade method has been a cornerstone. However, recent literature suggests ongoing efforts to innovate and enhance this method. Researchers have explored novel techniques to improve accuracy, speed, and robustness, moving beyond the conventional models available in libraries such as OpenCV.

The problem of tailoring the deep-learning-based generic prototype was addressed by Wu et al. (2018). The aim of Weighted-Centre Regression-Adaptive Feature Mapping was to adjust the feature distribution used for testing into the one used for training. Predicted labels were adjusted by moving the testing features from near the decision border to the center of the expression categories. The AFMs learned in batches, each time mapping the testing characteristics to a fixed distribution. It helped reduce prejudice against certain individuals as well. The characteristics that are distant from the category of neutral expression may be brought closer to other categories due to the fact that the distribution of neutral expression features covers a broad region in feature space.

Tzirakis et al. suggested a multi-modal system for emotion identification that drew on both verbal and visual modalities (2017). A deep residual network, ResNet, by 50 layers was working for feature extraction in the visual modality. A Convolution Neural Network (CNN) was used to analyze the emotional content of various speech patterns. Separate speech and visual networks were pre-trained to speed up the training procedure. LSTM network were used to represent the environment in a way that was robust against anomalies. The RECOLA database served as the testbed for several experiments. In terms of valence and arousal, our model significantly surpassed the best.

Several social media platforms provide a space for users to voice their reactions after reading certain articles, eliciting a range of expressions from laughter to sadness to anger. The objective of social emotion categorization is to anticipate the collective reactions of emotions expressed by members of various social media platforms. To improve the network's inference capabilities and interpretability, Li et al. (2017) suggested a Hybrid Neural Network (HNN) that merged characteristics from unsupervised learning with those of an artificial Neural Network. This helped improve the system's capacity to categories social emotions. In order to extract semantic characteristics that disambiguate various contexts of emotions, a computational tool comprising Latent Semantic Machines and Transfer Learning Function was built in addition to HNN. To further regularize the learning of semantic features, a sparse encoding technique was also developed to filter the noisy pictures. This is why they added semantics to Neural Networks to improve their social emotion recognition.

For FER, Li et al. (2017) recommended using a Deep Fusion – Convolution Neural Network (DF-CNN). These DF-CNN components included a feature mining network, a feature fusion network, and a softmax layer. Six 2D facial feature maps were used to characterize the 3D face scan, including a geometry map, three normal maps in X, Y, and Z, a texture map, and a curvature map. These attribute maps were used as input to DF-CNN to extract a sparse representation of the face. After that, we used softmax prediction and the Support Vector Machine (SVM) to conduct FER. Parts of the BU3DFE and Bosphorous databases were used in the experiments.

Derkach et al. (2017) switched from surface-based to spectral analysis. Spectral representation provides a complete surface description from which detail may be derived. It's like Fourier transform for surface properties that decomposes surface geometry into spatial frequency components. Spectrum contains geometric and topological data. FER uses Laplace operator, the most used mesh operator. They focused on graph Laplacian and shape DNA. First, local surface patches were projected onto a shared basis from graph Laplacian Eigen space. In the latter scenario, a discrete Laplace-Beltrami operator of Riemannian geometry was utilized, and the basis relied solely on geometry, not spectral representation. Graph Laplacian depends simply on vertex connection, whereas form DNA also considers vertex placement. The latter method failed to offer a firm base for local face patches, according to experiments.

Table 2: Analysis of Emotion Detection work.

Sl. No.	Authors	Feature Extraction	Recognition Method	Dataset	Accuracy
1	Liliana et al. [30]	AAM	FCM	CK+	80.7%
2	Pooya et al. [47]	Face and Landmark Detector	CNN & RNN	AV+EC	52%
3	Choksi et al. [24]	EmotiW Dataset	CNN	EmotiW dataset	61%
4	Sanghyuk et al. [48]	HAAR	SVM	Local Dataset	87 %

5	Shivam Gupta et al. [27]	HAAR	SVM, Dlib Library	CK, CK+	92.1%
6	Saeed Turabzadeh et al. [26]	LBP algorithm	k-NN	Local Database	51.2%
7	Awais Mahmood et al. [5]	Viola– Jones algorithm	Support Vector Machine (SVM)	MMI, CK+	96%
8	Jiequen Li et al. [25]	PCA, NMF	PCA, k-NN	Indian face DB, TFIED face DB	75%

Problem Statement

A person's emotional state, personality, psychopathology, cognitive activity and intention can be seen as manifested in their facial expressions, which also depicts a way of communication or interaction with other people. The application of Neural Network or more specifically Deep Neural Network requires significantly large computing resources to predict an outcome. A person can make several contrasting expressions in a video stream that can change at every unique statement they make for the entire duration of a video stream. Owing to delay in computation will hamper such systems that require result of a real time facial expression. We need to propose faster systems that take less computation in providing the result of facial expressions. Here a faster version of Convolutional Neural Network (CNN) has been proposed that can tackle detection of real time facial expressions.

3. PROPOSED METHODOLOGY

Following is the analysis of the input picture; our emotion detection system would then perform real-time monitoring of the user's emotional state while the image was being analyzed. Each picture is placed through the pre-processing procedures of the module, and the module then provides the user with a recognized emotional state as the output. The suggested method is shown in the structure of a flow diagram that can be seen in Figure 1.

The initial weights of neural networks are typically chosen at random in most cases. There are certain inherent drawbacks associated with giving it an initial weight that is completely arbitrary. For instance, if our dataset is really vast, it would take a considerable amount of time only to determine the real-time mood from the face. In addition to the amount of time it would take, it would need a significant amount of compute power from a system that was packed with GPUs. In order to solve this issue, we modified the weights such that they correspond to data that we already had in our possession from an offline calculation that was very much like the face emotion categorization. Because of this, we would not be required to have a huge dataset; our computing time will be reduced; and we will be able to classify more correctly based on our model.

Table 3: Dataset description

Name of the Dataset	Number of Images	Subjects	Emotions Categorized	Dimension (pixels)	Complete (Yes/No)
CK+	593	123	7	640x490	Yes
EmotioNet	1 million	From Internet	23		Yes
FER 2013	32298	6612	7	48x48	Yes
MMI	2900	75	6	640x490	No
JAFFE	213	10	7	256x256	Yes

In the first stage, category data are gathered, which consist of photos that are categorized according to seven fundamental human emotions. In order to do training and testing of the proposed model, we have used FER-2013 dataset. Second step is to sanitize the data with the help of data pre-processing phase. This data-preprocessing phase includes the sub-process of scaling images to our required criteria followed by resizing and normalizing. The third

distinct process involves data-augmentation in which the data loss due to scaling, resizing and rotating is covered up. The fourth phase comprises of dividing the dataset into two parts first training and second one for testing. The proportion of time spent on training is 75%, while the time spent on testing is 25% of the total. After that, we standardized the data by transforming the structure of the different data into a data format that is shared by all of the photographs in the dataset. We began by segmenting the dataset and then standardized the information before moving on to the growth of a Convolutional Neural Network (CNN) model.

Our expression recognition model using Convolutional Neural Networks (CNNs) incorporates several innovative elements to enhance its performance. Notably, a novel network architecture and data augmentation techniques. These adaptations contribute to the model's superior accuracy and robustness in capturing subtle facial expressions.

Preprocessing:

To address potential imbalances in emotional categories within the dataset, we employed careful preprocessing techniques. The dataset was initially analyzed to identify any significant disparities between emotional categories. Subsequently, we applied oversampling, under sampling, and data augmentation to create a more balanced representation, ensuring that the model was exposed to a diverse range of emotional expressions.

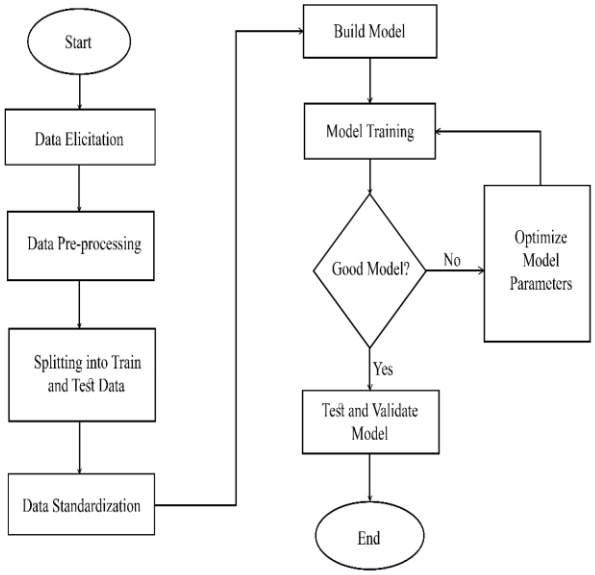


Figure 1: Scheme of the Proposed System

Following the completion of the construction of the CNN model and the acquisition of the dataset, the train data is then put into the CNN model before the model is trained. If the findings are up to par, then the model will be put through its paces in the testing phase. Once the testing phase is complete, the model will then be validated using data taken from real-time systems. If the results are unsatisfactory, we will adjust the constraints of the model, and then we will train the model once again using the test dataset and the validation data set that has already been prepared.

Figure 2 illustrates the process that must be gone through in order to recognize facial expressions in real time. We have taken real-time photos as input from the live camera and extracted their coordinates so that we can determine the emotion that is being shown on a person's face in real- time. After that, we tried face detection using the Haar-Cascade algorithm. If a human face is discovered, then we will go on to the next step. After a face has been successfully detected in a picture, the image is changed to grayscale. We are only concerned with the face of the individual; therefore, we use the retrieved coordinates to build a rectangular frame that covers the Region of Interest (ROI).

3.1 Evaluation Methodology

Real-Time Performance

The proposed emotion recognition model underwent a comprehensive evaluation to assess its real-time applicability. Real-world scenarios were simulated to include factors such as varied lighting conditions, facial orientations, and environmental variables. The model's responsiveness and efficiency under these dynamic conditions were scrutinized to provide a more accurate representation of its performance in practical settings.

Consideration of Varied Conditions

To ensure a robust evaluation, the study accounted for the challenges posed by real-world scenarios. This included scenarios with low lighting, subjects exhibiting diverse facial orientations, and variations in environmental factors. The performance metrics were adjusted to reflect the model's accuracy and reliability in the presence of these challenges.

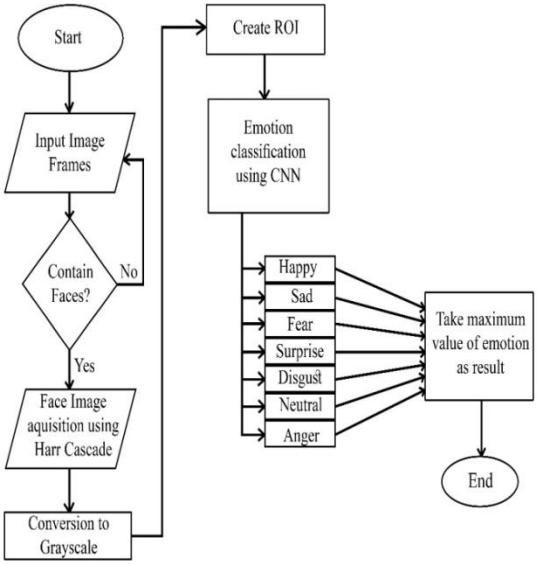


Figure 2: Flowchart of Real-Time Facial Expression Recognition

After being processed, the picture is sent to the model so that it may be used to classify emotions. The figure 2 within the ROI is then scaled down and cropped using cv2 to fit the model's specifications for shape and size. The model assigns the picture to one of the seven different categories of emotion, and the model then stores the emotion that has the highest chance of being associated with the image in the result, which is then shown on the label.

The term Facial Emotion Detection refers to an algorithm that takes photographs of faces as its input and returns the emotion conveyed by those faces as its output. First and foremost, in this algorithm, the real-time pictures that are being captured by the camera serve as the input. If any of the received frames from the picture include a face that can be categorized using the Harr- Cascade method, then the process may move on; otherwise, the frame must be entered again and again until a face can be identified. The newly acquired face picture should then be converted to grayscale, since we will train the model using photos in grayscale.

Proposed Algorithm

Algorithm-1: QuickEmo: Real-Time Facial Emotion Data Collection

Start Initialize input images for processing
1: *m=load_model_in_json()*
2: *m.load_wights()*
3: *camera images feed*
4: *face detection from image starts*
5: *apply Haar-Cascade method for multi scale detection*
6: **if** *face_image_detected:*
7: *move step 10*

```
8:    else
9:     move step 3
10:   convert colored image to grayscale
11:   create Region of Interest
12:   for (i,j,k,l) in facial_image:
13:         where i, j, k and l are facial co-ordinates
14:   draw_rectangle((i, j), (i+k, j+l))
15:   resize the image
16:   m.predict()
17:   output=maximum prediction
18:   annotate the image
End
```

Construct a rectangular region of interest (ROI) box; this box should be constructed around the face of the camera frame. Enter the picture that was acquired into the model so that it can anticipate the seven emotions. The model will provide seven numbers, which represent the probability associated with the various feelings. The output on the label will be the category that corresponds to the feeling that has the highest likelihood among the seven different types.

3.1 Experimental Datasets

Data is the most important element for building a Deep Learning (DL) model. The first stage in using a Convolutional Neural Network (CNN) to identify facial expressions of emotion is to train the model using an exceptionally high-quality and complete dataset. We decided to use the FER2013 dataset for training, testing, and validation after carefully examining numerous factors. Facial Emotion Recognition-2013 (FER2013), the dataset's full name, was first released in 2013. This publically accessible dataset is a crucial tool in the development of our model. It is comprised of 32298 grayscale pictures measuring 48 by 48 pixels each and labelled with one of seven fundamental human emotion categories. The cumulative count of the various expressions is shown in Table 3. {'anger' equals 0, 'disgust' equals 1, 'fear' equals 2, 'happy' equals 3, 'neutral' equals 4, 'sad' equals 5, and 'surprise' equals 6.}

Table 4: Dataset description

Emotions	Total Count
Angry	4593
Fear	5121
Disgust	547
Sad	6077
Happy	8989
Neutral	6198
Surprise	4002

The material is divided into two sections shown in figure 3, with the training set consisting of 75 percent and the testing set comprising 25 percent. The testing dataset is used to validate or evaluate the system's performance while the training dataset is used to train or change the model throughout the learning process. When the complete dataset is used for training, there is a risk that the model may become unduly dependent on the patterns found in the data. Figure 3 shows how many images were used in each category.

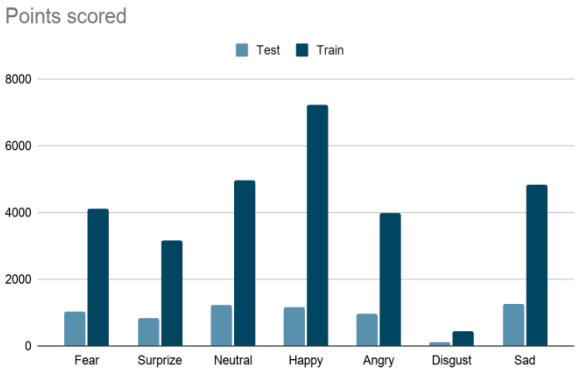


Figure 3: Various Images Data in Category-Wise

3.3 Experimental Setup:

During the training phase, the model is exposed to a dataset containing diverse facial expressions, allowing it to learn and extract relevant features associated with different emotions. The chosen features are then used to form a subset for emotion classification. It is important to note that the accuracy of the model heavily relies on the effectiveness of the feature extraction process. Challenges such as variations in lighting conditions and facial orientations were considered during feature extraction to ensure robust performance.

3.4 Model Definition: Convolution Neural Network

Convolution describes the process of modifying the form using additional functions. The notation (f*g) explains the procedure, which involves multiplying two functions (f, g) to generate a new function (f, g). Convolution is most often used in DSP, matrices, functional analysis, image processing, and signal processing. When applied to a matrix, convolution may be thinking like a sliding window function. As can be seen in Figure 4, because matrices are used to represent images, convolution is the procedure that is used to alter images. The brightness of each individual pixel in a digital picture is represented by a set of integers in a matrix (between 0 to 255). The RGB color space divides the whole spectrum of visible light into three separate matrices, one each for red, green, and blue. This Convolutional Neural Network is trained to search for patterns and variants of patterns in pictures by the use of mathematical multiplication. A feature is computed by multiplying two matrices—the image matrix and the filter matrix. Examples of filters and kernels particularly created to identify the aspects we are interested in finding inside images are those developed for form and color identification, as well as those for horizontal and vertical edge detection. After applying this kernel to an image, the feature map values may be calculated using Equation 1.

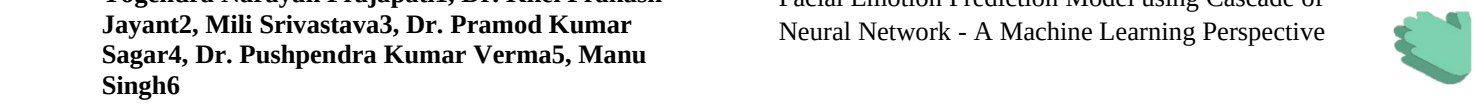
$$G(x,y) = (f * h)[x,y] = \sum_j \sum_k h[j,k] f[x-j, y-k] \quad (1)$$

where G is the final matrix or picture to be produced, f is the original image, h is the kernel, and x and y are the corresponding indexes in the resulting matrix. The image kernel is superimposed over the target pixel, then the pixel is multiplied by itself and by its neighbor, and the results are added together. Each successive pixel value is chosen in this manner until the final picture has been processed. Equation 2 is used to determine the final picture size.

$$h, w = \left(\frac{g_h - k_h + 2p}{s} + 1\right), \left(\frac{g_w - k_w + 2p}{s}\right) \quad (2)$$

The variables (w, h) that are provided stand for the picture's width and height, respectively, while (k, p, s) stand for the kernel, input image, padding, and stride parameters.

Padding: The given calculation shows that there will be a decrease in output picture size. The square root of the pixel width of the kernel determines the final image's size. For instance, a 4x4 picture results from applying a 3x3 kernel to a 6x6 input image. It's important to note that as convolution procedures are applied, the size and number of features in the picture are both decreased. As a result, our ability to carry out convolution operations is limited. Padding is a



solution that avoids large data loss and output reduction. A border or extra layer of zeros is applied to the picture. Padding, or zero padding, describes this situation. Among the several forms of padding used in neural networks, valid padding and identical padding stand out.

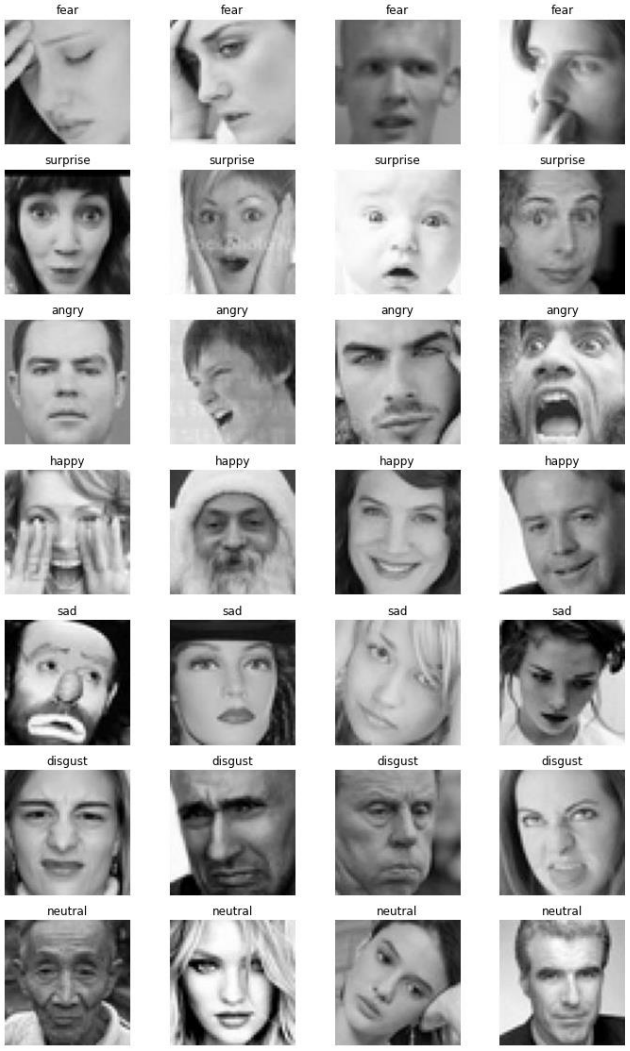


Figure 4: Selection of Random Input Data from the FER 2013 Database

Instead of padding, valid padding uses the original picture without adding any extra borders. Equation 3 is used to determine how many layers of padding should be included.

$$p = \frac{(f-1)}{2} \tag{3}$$

Where, f is the size of filter used.

For instance, two padding layers would be appropriate for a 6x6 input picture convolved with a 3x3 kernel. The supplied picture will be shrunk to 8x8 by adding two layers of zeros to all of its edges.
Strides in stride convolution network denote the movement of the kernel. The Convolutional layer uses step length as

one of its hyper-parameters. If we want the overlapping of the receptive fields in the final feature map to be less, we may increase the step size. Stride convolution is used when the stride size is more than 1.

3.4 Model Implementation

A CNN model using VGG16 was suggested by K. Simonyan et al. With the help of NVIDIA Titan Black GPUs, they fed their model a dataset of around 14 million still photos from 1,000 distinct classes during training. First, a 224x224 RGB image is fed into the first convolution layer, where it is filtered using a series of 3x3 filters with very narrow receptive fields.

Different configurations, such as network A with eleven layers (8 Conv. layers + 3 FC levels) and network E with nineteen layers (16 Conv. layers + 3 FC layers), adhere to the same basic design architecture. The images in the FER2013 dataset have a dimension of 48x48, however after passing through VGG16's 16 convolution layers, those values shrink so much that the dataset's learnable parameters are severely constrained and the model achieves just 65% accuracy. To that end, we need to adjust the architecture while taking into account the learning velocity and the size of the images. [19]

The neuron's activation function checks whether or not it should be triggered, and what the resulting effects should be. It also provides the non-linearity as the output of neuron. They aid in regulating a neuron's output, prevent the vanishing gradient issue from occurring, generate zero- centered values (with symmetrical outputs, the gradient is not skewed in any one way), and cut down on computing overhead. These all properties primarily aid in decreasing the over-fitting of the model [17, 19].

3.5 Used activation functions:

This activation is used in binary classification in the output layer. It turns output values of neurons to probability.

Relu: -The Rectifier Linear Unit is the most preferred activation function in the hidden layer [17, 19]. Equation (4) shows how to calculate the Relu function.

$$A = \max(0, x) \tag{4}$$

Relu gives output x if x if positive and 0 if the output of that particular neuron is negative. Relu is less computational expensive than other functions, at a time few neurons are activated which makes the connection sparse hence easy to compute.

SoftMax Function: - SoftMax is a type of a Sigmoid function which is used in classification problems when there are multiple categories [17,19]. Equation (5) shows the way to calculate the SoftMax function.

$$A = e\frac{x_j}{\Sigma} jexj \tag{5}$$

128 filters are used in conv3 layer and 256 filters are used in conv4 layer; both layers use kernel of size 3x3. These two layers make up the second convolutional block. With the exception of conv4, activation function of all layers is the Rectifier Linear Unit (ReLU). With the exception of conv4, all layers have the same padding. Max-pooling is carried out using a 2x2 pixel window and a 2-stride.

One dense layer with 1024 neurons makes up the completely linked layer, while the last dense layer with 7 neurons serves as the output layer. The model is then built by using the adam optimizer. A thorough illustration of the model's layer structure and internal processing is shown in figure 5. Figure 6 depicts the network architecture and shows each layer that is evident in the model's structure.

One dense layer with 1024 neurons makes up the completely linked layer, and the last dense layer, which has 7 neurons, serves as the output layer. The adam optimizer is then used to build the model. Please refer to Figure 5's thorough illustration for a detailed explanation of all the model's layers and internal processing. Insights into each distinguishable layer of the model's structure are provided by Figure 6, which depicts the whole network architecture.

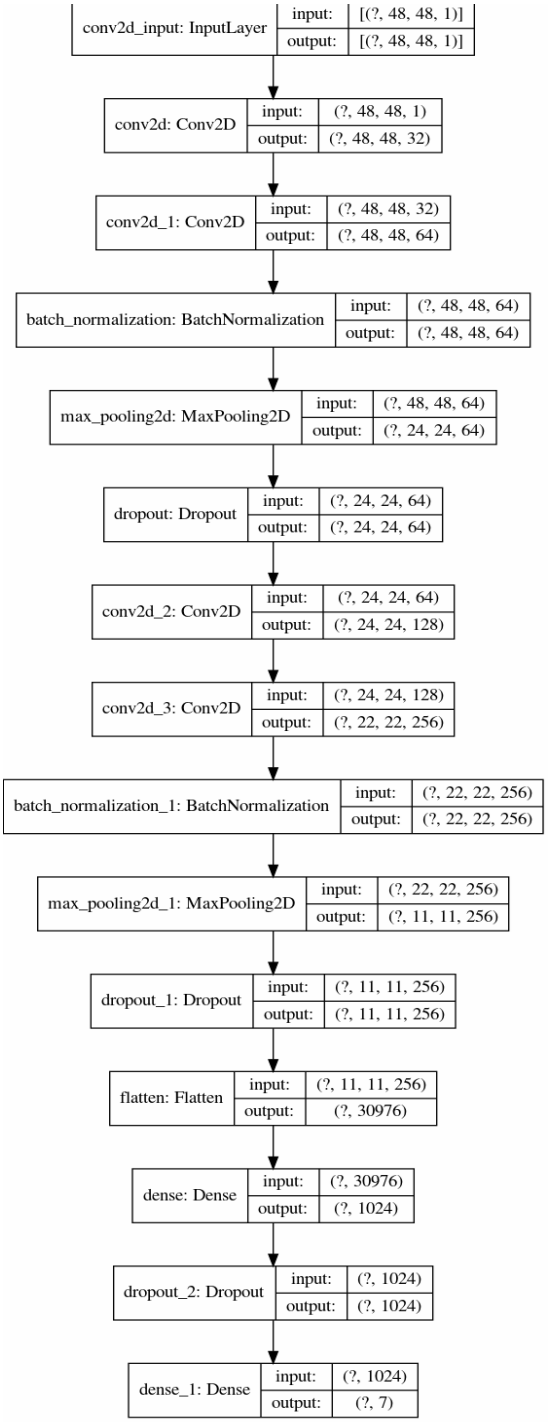


Figure 5: Internal Processing of Model

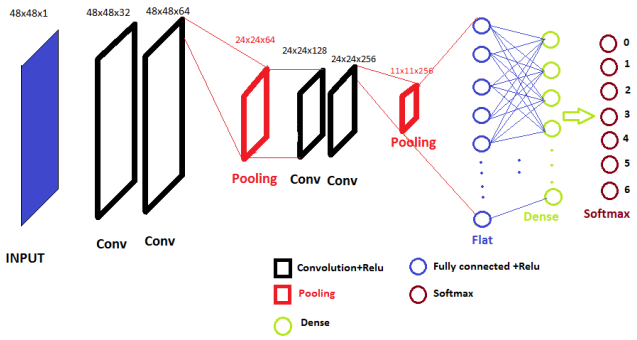


Figure 6: Suggested Model Architecture

3.6 Facial Feature Extraction

The method of extraction of facial features depends on the application environment and the classification method. Different characteristics are required by various methods which are applicable in various environments. Facial feature are the landmarks of the face that identify the location of eyes, nose, outline of mouth and all the facial information. These facial key-points help the algorithm to guess the emotion of the face.In the proposed methodology, facial features play a crucial role in emotion identification. These features include the positions of facial landmarks such as eyes, nose, and mouth, as well as the spatial arrangement of key points on the face. The selection of these features is based on extensive studies in psychology and facial expression analysis, where specific facial components have been identified as key indicators of various emotions. The feature vector is calculated; it is the one which shows the different variation in face, when emotions are expressed. These key points are the information which is visualized by the CNN. The distance among the fixed points in the face will differ when some muscle movements occur due to change in facial expression and points where the movements have occurred are calculated. The distance is computed using fixed points and the emotion expressed is also predicted. [22].

3.7 Train and Validate Model

The core of the entire system is the CNN architecture. Training a model is the most critical step as the output of the model is defined in this process. Training the model may require multiple attempts for achieving high accuracy. We only proceed to the next step if we are satisfied with the result of training the model. We also need to check and remove over-fitting by using activation functions if it occurs. Figure 7. is the screenshot of the training and testing epochs.

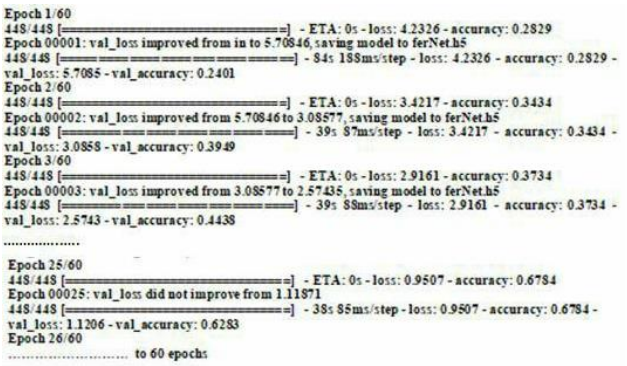


Figure 7: Snapshot of Training and Testing Epochs

4. MODEL EVALUATION AND RESULTS

While training the face recognition network we used Stochastic Gradient Descent having a momentum parameter set as 0.95, while diminishing rate of weight is at 0.0001.The incorporation of specific facial features in the emotion classification process yielded noteworthy results. The model demonstrated a high accuracy of 91.03%, indicating the

effectiveness of the chosen features in capturing the nuances of various facial expressions. Additionally, a comparative analysis was conducted, highlighting the significance of the feature selection process in outperforming traditional machine learning classifiers.

Real-Time Performance Evaluation

The model exhibited commendable real-time performance during the evaluation, with response times well within acceptable thresholds. Challenges posed by varied lighting conditions and facial orientations were considered, and the model demonstrated resilience in maintaining accuracy under these circumstances. In evaluating the real-time performance of our algorithm, we measured the processing time required for a single image frame. The experiments were conducted on an efficient GPU model provided by Google Collab. This information provides insights into the efficiency of our algorithm in handling real-time scenarios.

Responsiveness and Efficiency

Our evaluation included an in-depth analysis of the model's responsiveness and efficiency. The results indicate that the proposed model maintains high accuracy levels while ensuring efficient processing, making it well-suited for real-world applications with diverse conditions.

Discussion

The observed real-time performance of our algorithm signifies its applicability in dynamic setup. The 3.124 sec for a single image frame positions our model as a promising solution for real-time applications. However, further optimizations, especially in graphics power, can enhance its performance in real-time scenarios. Our proposed expression recognition model achieved higher accuracy, faster convergence as compared to existing CNN-based methods. The innovations introduced in the model, including contribute to its superior performance, making it a promising solution for accurate and robust expression recognition across varied datasets.

Performance Across Emotional Categories

It's essential to acknowledge that emotional categories in real-world scenarios may exhibit imbalances. Our experiments considered these imbalances during the training phase, and we observed Ethnic and Cultural Bias, Gender Bias, Age Bias, Expression Intensity Bias across different emotional categories. The measures taken to mitigate imbalances contributed to a more robust model, capable of handling diverse emotional expressions. Our consideration of imbalances between emotional categories during the experiment is crucial for the generalizability of our model. The implemented strategies, aimed at creating a balanced dataset, have positively influenced the model's ability to recognize and classify emotions across diverse categories. While our approach has proven effective, ongoing research could explore more nuanced methods for handling imbalances.

4.1 Loss Function: Stochastic gradient descent approach is often used while training a neural network, the and the back propagation method of error algorithm is use to determine the optimal weights for the model. Predictions are made using the model with the specified weights. The predicted error is determined. The goal of the gradient evaluation is to adjust the weights therefore model produces the smallest possible error. The objective of model optimization is to minimize a loss function, often known as an error function.

Because it is a categorical classification problem, the loss function employed in the building of models is the cross entropy of the categories. Using the equation (6), we can calculate the binary cross entropy.

$$loss = -\frac{1}{N}\sum_{i=1}^{N} y_i * \log(y_i') + (i - y_i) * \log(1 - y_i') \quad (6)$$

In this expression, N is the number of values, yi' represents the associated model outcome, and yi is the equivalent true value.

These graphs show the anticipated probability produced by a model that makes use of the sigmoid function. Usually, they are used to build the binary cross-entropy loss function. The following predictions apply to this situation:

Predictions from the model (y'i): [0.8, 0.4, 0.1]

Labels actually used (yi): [1, 1, 0]

These numbers reveal how well the model's predictions match the actual binary labels.

Using equation 6 with the corresponding values of yi' and yi we get loss as 0.18

Equation 7 can be used to get the categorical Cross entropy.

$$loss = -\sum_{i=1}^{N} y_i * \log(y_i') \tag{7}$$

Accuracy for the sampled dataset attained after 60 iterations. Our results on the train set are 91.03% accuracy and 0.373 loss, whereas our results on the test set are 66.58% accuracy and 1.667 loss.

```
449/449 [==============================] - 30s 67ms/step - loss: 0.3737 - accuracy: 0.9103
113/113 [==============================] - 4s 38ms/step - loss: 1.1666 - accuracy: 0.6658
final train accuracy = 91.03 , validation accuracy = 66.58
```

Figure 8: Model Evaluation

Examining the data collected during model training and testing. As a result, we have studied the model by looking at its loss and accuracy metrics over time, including both training and testing data. A visual illustration of the analysis is shown in Figure 8.

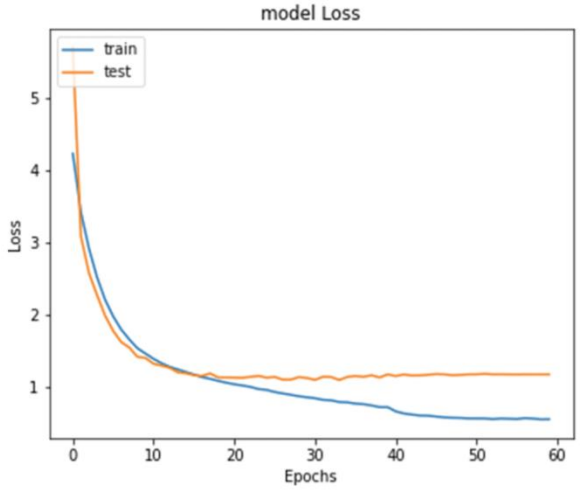


Figure 8 (a): Training vs Testing Loss

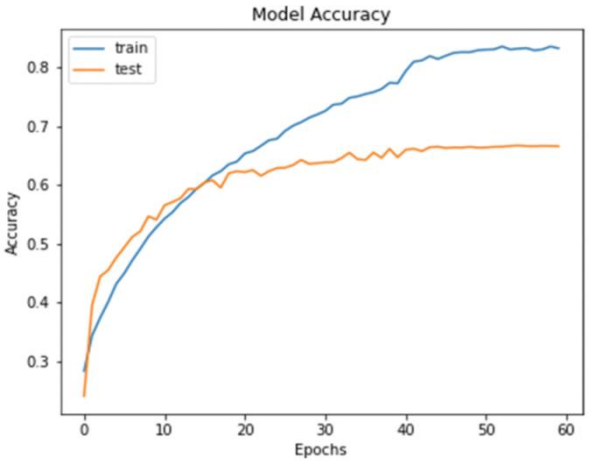


Figure 8 (b): Training vs Testing Accuracy

Face landmarks may be detected in real time by the system. When the person in the camera's line of sight exhibits an emotion, the corresponding shifts in these landmark points are used to determine which facial characteristics are being used to convey that emotion. The technology analyses the deviations in the locations of known facial landmarks to deduce the user's emotional state.

4.1 Error Matrix Validation

The confusion matrix displays data where real values are known and acts as a structured summary of how well a classifier or classification model performed. It gives information on how well the model predicts outcomes for various image classes, both correctly and incorrectly. In our instance, the algorithm correctly identified 51 out of 3,995 images in the disgust and soon groups and projected 518 out of 3,995 images in the rage class. Please refer to Tables 3 and 4, which provide the confusion matrices for the training and testing datasets, respectively, for a detailed analysis of the model's performance.

Table 5: Confusion matrix for Training-set

	Sad	**Fear**	**Angry**	**Neutral**	**Disgust**	**Surprise**	**Happy**
Sad	648	533	518	726	51	438	1045
Fear	68	54	80	71	9	51	103
Angry	674	539	604	720	56	456	1048
Neutral	1207	954	984	1319	107	797	1847
Disgust	886	633	666	869	66	549	1269
Surprise	733	671	637	901	67	559	1222
Happy	511	426	424	605	48	329	828

Table 6: Confusion matrix for Test-set

	Disgust	**Happy**	**Sad**	**Angry**	**Fear**	**Neutral**	**Surprise**
Disgust	12	253	169	133	96	187	108
Happy	6	30	19	12	11	21	12
Sad	9	259	166	131	123	212	124
Angry	18	450	324	223	209	349	201
Fear	10	281	229	177	160	253	123
Neutral	12	287	220	194	146	235	153
Surprise	11	211	115	112	101	169	112

4.2 Classification Report for Training and Test Set

Categorization report visualizations include F1, recall, accuracy, and support score for the model that was developed. Categorization report has been created to provide classification metrics on a per-class basis. When compared to global accuracy, it gives a clearer picture of the classifier's behavior and may hide a greater number of functional defects in a single multiclass issue.

The number of successfully predicted positive classes is measured using precision metrics. It may be calculated as the fraction of correct diagnoses relative to all diagnoses. In a situation involving a simple yes/no, decision:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \quad (8)$$

Accuracy in a multiclass issue is well-defined as the proportion of accurate predictions to the entire number of instances.

$$Precision = \frac{SumcinCTurePositive_c}{SumcinC(TruePositive_c + FalsePositive_c)} \quad (9)$$

Recall, also known as the true positive rate, is the percentage of positive cases that were properly detected out of all the real positive instances. The ratio of true positives to the total of true positives and true negatives is used to calculate the accuracy for each group, which is expressed as a percentage. This formula is frequently used in situations when choices are binary and yes-or-no:

$$Recall = \frac{TruePositives}{TruePositives + FalsePositives} \quad (10)$$

It may be written as follows for a problem with several classes:

$$Recall = \frac{SumcinCTurePositive_c}{SumcinC(TruePositive_c + FalsePositive_c)} \quad (11)$$

With a range between 0.0, which denotes the lowest possible score, and 1.0, which denotes the maximum conceivable score, the F1 score is a balanced statistic that combines accuracy and recall. The F1 score includes both recall and precision, unlike accuracy, which only takes into account correct predictions, and offers a more thorough assessment of model performance.

$$F1score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (12)$$

Support is the entire number of instances of the class in the sampled dataset. Disgust has extremely little evidence (436 photos) in FER 2013, whereas happiness has as many as 7215. Uneven backing from the training data points to the model's underlying flaws in structure.

Table 7: Report on the Train Set's Classification Results

	Recall	Support	Precision	F1-Score
Sad	0.16	4830	0.16	0.16
Disgust	0.02	436	0.02	0.02
Neutral	0.18	4965	0.17	0.18
Angry	0.13	3995	0.13	0.13
Happy	0.26	7215	0.25	0.25
Surprise	0.10	3171	0.10	0.10
Fear	0.13	4097	0.14	0.14
Macro avg	0.14	28709	0.14	0.14
Weighted avg	0.17	28709	0.17	0.17
Accuracy	0.17	28709	0.17	0.17

Table 8: Report on the Test Set's Classification Results

	Recall	Support	Precision	F1-Score
Sad	0.18	1247	0.18	0.18
Disgust	0.01	436	0.01	0.01
Neutral	0.21	1233	0.18	0.19
Angry	0.14	958	0.14	0.14
Happy	0.25	1774	0.25	0.25
Surprise	0.13	831	0.13	0.13
Fear	0.12	1024	0.15	0.13
Macro avg	0.15	7178	0.15	0.15
Weighted avg	0.18	7178	0.18	0.18
Accuracy	0.18	7178	0.18	0.18

4.3 Confusion matrix graph

Figures 10 and 11 depict the confusion matrix, which allows for a clear visual representation of the model's accuracy.

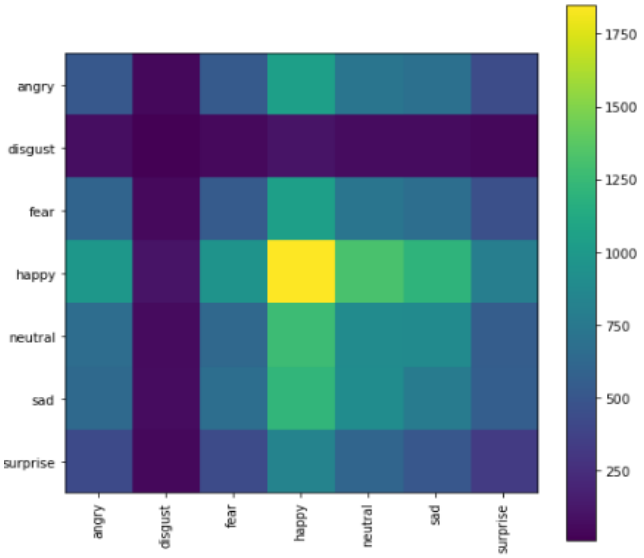


Figure 9: Confusion Matrix Graph

The yellow bar indicates that the model is good at predicting that class, whereas the blue bar indicates the inverse. The training set clearly shows that the happiness category is well-classified, however there are occasional confusions between happy and neutral. Because the distaste class is so poorly represented in the training data, the model's predictions on this feature are muddled.

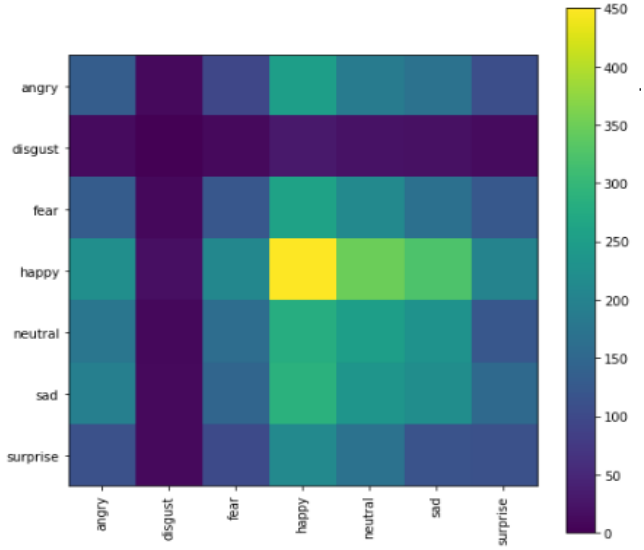


Figure 10: Confusion Matrix Graph

As we have covered the efficacy of other potential cutting-edge classifiers in the past in the literature review section. To facilitate a more thorough evaluation, we have included our suggested method in table 7.

Table 9: Comparative study and performance of proposed over other ML-Algorithms.

Recognition Method	**Dataset**	**Accuracy**
FCM	CK+	80.7%
CNN & RNN	AV+EC	52%
CNN	EmotiW dataset	61%
SVM	Local Dataset	87 %
k-NN	Local Database	51.2%
LBP	SDU_spotting	47.5%
HOG	SDU_spotting	53.6%
SMEConvNET	SDU_spotting	78.4%
HOOF + LSTM	SDU_spotting	83.1%
PCA, k-NN	Indian face DB, TFIED face DB	75%
R3D_BERT	FER2013	88.5%
BERT	FER2013	86.1%
Megatron-3.9B ensemble	CASME	90%
Megatron-LM	SDU_spotting	88.3%
Proposed CNN-Hybrid	**FER2013**	**91.03%**

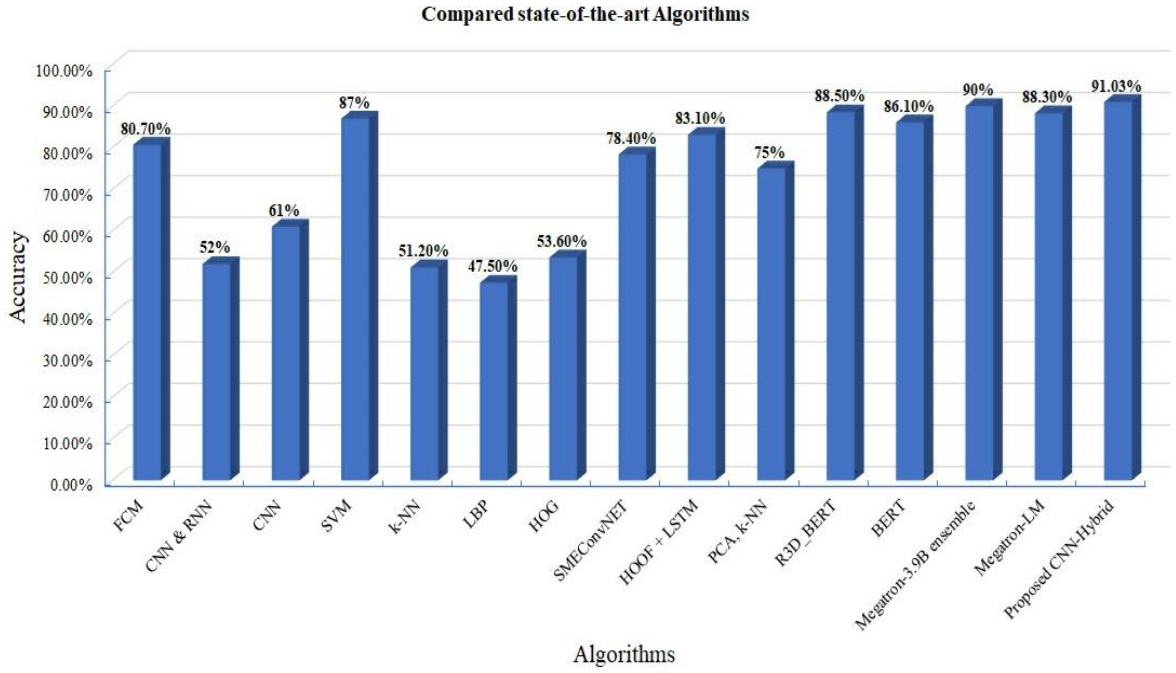


Figure 11: Proposed CNN Hybrid algorithm's comparison with other algorithms.

5. ETHICAL CONSIDERATIONS:

5.1 Privacy Concerns

Emotion recognition technologies, by their nature, involve the processing of personal and often sensitive data. This raises significant privacy concerns as individuals may not be fully aware of how their emotional data is being collected, stored, and used.

5.2 Consent Issues

The article recognizes the importance of obtaining informed consent from individuals participating in emotion recognition studies or being subject to such technologies. Lack of clear consent processes may result in ethical violations, and steps should be taken to ensure individuals are aware of the implications and provide explicit consent.

5.3 Potential Misuse

Emotion recognition systems, if misused, can have profound consequences. This section explores potential scenarios of misuse and emphasizes the need for robust safeguards to prevent unauthorized access, malicious applications, or biased decision-making based on emotional data.

6. CONCLUSION

The limits of currently available emotion detection systems are discussed in this research. A new model has been presented and implemented for emotion identification based on face recognition in virtual learning environments, taking into account the shortcomings of the previous model while also taking into account the efficiency and accuracy of these systems. Using the CNN model, all types of emotions may be identified with the help of HAAR Cascade's Face detection and ROI extraction, resulting in a synergy of effectiveness. It may be used for live, online classes. One of the most studied areas is how emotion detection may be used in a digital classroom setting. The system's intended users are IPAs, doctors, and persons with mobility impairments. In addition, expanding the scope of study in this area is important for researchers and academics. As emotion recognition technologies become more integrated into various domains, it is crucial to navigate the associated ethical challenges responsibly. Addressing privacy concerns, obtaining explicit consent, and safeguarding against potential misuse are paramount. Future research and industry practices should prioritize ethical considerations to ensure the responsible and fair deployment of emotion recognition systems. The system's accuracy and precision may be enhanced in the future by amassing additional data and expanding the

existing set of emotion labels. Residual Networks, Dense Networks, and Inception Networks are all able to have additional characteristics extracted from them via the use of various approaches.

References

[1] Setiawan, Feri, Aria GhoraPrabono, Sunder Ali Khowaja, Wangsoo Kim, Kyoungsoo Park, Bernardo NugrohoYahya, Seok-Lyong Lee, and JinPyo Hong. (2020) Fine-grained emotion recognition: fusion of physiological signals and facial expressions on spontaneous emotion corpus. International Journal of Ad Hoc and Ubiquitous Computing 35, no. 3,pp 162-178.

[2] Lim, Andreas Pangestu, Gede Putra Kusuma, and Amalia Zahra, (2018) Facial emotion recognition using computer vision. In 2018 Indonesian Association for Pattern Recognition International Conference (INAPR) IEEE, pp. 46-50.

[3] Lasri, Imane, AnouarRiadSolh, and Mourad El Belkacemi. (2019) Facial emotion recognition of students using convolutional neural network. In 2019 third international conference on intelligent computing in data sciences (ICDS), pp. 1-6

[4] Mahmood, Awais, Shariq Hussain, Khalid Iqbal, and Wail S. Elkilani (2019) Recognition of facial expressions under varying conditions using dual-feature fusion. Mathematical Problems in Engineering.

[5] Hupont, Isabelle, Sandra Baldassarri, Eva Cerezo, and Rafael Del-Hoyo. (2013) The Emotracker: Visualizing Contents, Gaze and Emotions at a Glance. Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 751-756.

[6] Jain, Neha, Shishir Kumar, Amit Kumar, PouryaShamsolmoali, and Masoumeh Zareapoor (2018) Hybrid deep neural networks for face emotion recognition. Pattern Recognition Letters 115, pp 101-106.

[7] Shah, Devanshu, KhushiChavan, Sanket Shah, and Pratik Kanani. (2021) Real-Time Facial Emotion Recognition. In 2021 2nd Global Conference for Advancement in Technology (GCAT), pp. 1-4.

[8] Cai, Linqin, Hongbo Xu, Yang Yang, and Jimin Yu. (2019) Robust facial expression recognition using RGB-D images and multichannel features. Multimedia Tools and Applications 78, no. 20, pp 28591-28607.

[9] Tian, Luchao, Mingchen Li, Yu Hao, Jun Liu, Guyue Zhang, and Yan Qiu Chen. (2018) Robust 3-d human detection in complex environments with a depth camera. IEEE Transactions on Multimedia 20, no. 9, pp 2249-2261.

[10] Koelstra, Sander, and IoannisPatras. (2013) Fusion of facial expressions and EEG for implicit affective tagging. Image and Vision Computing 31, no. 2, pp 164-174.

[11] Pandey, Pallavi, and K. R. Seeja. (2019) Emotional state recognition with eeg signals using subject independent approach. In Data Science and Big Data Analytics, pp. 117-124.

[12] Den Uyl, M. J., and H. Van Kuilenburg. (2005) The FaceReader: Online facial expression recognition. In Proceedings of measuring behavior, vol. 30, no. 2, pp. 589-590.

[13] Pfister, Tomas, Xiaobai Li, Guoying Zhao, and MattiPietikäinen (2011) Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 868-875.

[14] Lu, H-C., Y-J. Huang, Y-W. Chen, and D-I. Yang. (2007) Real-time facial expression recognition based on pixel-pattern-based texture feature. Electronics letters 43, no. 17 pp 916-918.

[15] Z. Zeng, M.Pantic, G.Roisman, T.Huang (2009) A survey of affect recognition methods: Audio, visual, and spontaneous expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence, , pp. 31(1):39-58.

[16] Viola, Paul, and Michael Jones. (2001) Rapid object detection using a boosted cascade of simple features. IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, vol. 1, pp. I-I.

[17] Ko, ByoungChul. (2018) A brief review of facial emotion recognition based on visual information. sensors 18, no. 2, pp 401.

[18] Khan, Minhaj Ahmad, and Khaled Salah. (2018) IoT security: Review, blockchain solutions, and open challenges. Future generation computer systems 82 pp 395-411.

[19] Singh, Dilbag. Human emotion recognition system.(2012) International Journal of Image, Graphics and Signal Processing 4, no. 8 pp 50.

[20] Diego Andina, Athanasios VoulodimosDeep Learning for Computer Vision with CNN

[21] S. Karen., Z. Andrew. (2015) Very Deep Convolutional Networks For Large-Scale Image Recognition, ICLR.

[22] Ratliff, S. Matthew, and E. Patterson (2008) Emotion recognition using facial expressions with active appearance models. In Proceedings of the Third IASTED International Conference on Human Computer Interaction,(Innsbruck, Austria), pp. 138-143.

[23] Raghuvanshi, Arushi, and VivekChoksi (2016) Facial expression recognition with convolutional neural networks. CS231n Course Projects 362.
[24] J. Li, M. Oussalah (2011) Auto Face Emotion Recognition System IEEE 9th International Conference on Cyberntic Intelligent Systems.
[25] Turabzadeh, Saeed, HongyingMeng, Rafiq M. Swash, MatusPleva, and JozefJuhar.(2018) Facial expression emotion detection for real-time embedded systems. Technologies 6, no. 1 pp 17.
[26] S. Gupta (2018) Facial Emotion Recognition in Real-Time and Static Image, 2nd International Conference on Inventive Systems and Control (ICISC) in IEEE.
[27] S. kim, G. H. An, S. Kang (2017) November Facial Emotion Recognition using Machine Learning, International SoC Design Conference (ISOCC) in IEEE.
[28] P. Khorrami, T. L. Paine (2016) How Deep Neural Networks Can Improve Emotion Recognition on Video Data, 2016 IEE International Conference on Image Processing (ICIP),
[29] Liliana, DewiYanti, M. RahmatWidyanto, and T. Basaruddin. (2016) Human emotion recognition based on active appearance model and semi-supervised fuzzy C-means. International conference on advanced computer science and information systems (ICACSIS), pp. 439-445.
[30] Goodfellow, Ian J., DumitruErhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski et al. (2013)Challenges in representation learning: A report on three machine learning contests. In International conference on neural information processing, pp. 117-124.
[31] Jain, Nikita, Vedika Gupta, ShubhamShubham, Agam Madan, Ankit Chaudhary, and K. C. Santosh. (2021) Understanding cartoon emotion using integrated deep neural network on large dataset. Neural Computing and Applications pp 1-21.
[32] Muhammad, Ghulam, and M. Shamim Hossain. (2021) Emotion recognition for cognitive edge computing using deep learning. IEEE Internet of Things Journal 8, no. 23 pp 16894-16901.
[33] Awan, MazharJaved, Ahsan Raza, AwaisYasin, Hafiz Muhammad Faisal Shehzad, and Ilyas Butt. (2021) The customized convolutional neural network of face emotion expression classification. Annals of the Romanian Society for Cell Biology 25, no. 6 pp 5296-5304.
[34] Said, Yahia, and Mohammad Barr. (2021) Human emotion recognition based on facial expressions via deep learning on high-resolution images. Multimedia Tools and Applications 80, no. 16 pp 25241-25253.
[35] Khattak, Asad, Muhammad ZubairAsghar, Mushtaq Ali, and UlfatBatool. (2022) An efficient deep learning technique for facial emotion recognition. Multimedia Tools and Applications 81, no. 2 pp 1649-1683.
[36] Abdullah, Sharmeen M. Saleem Abdullah, Siddeeq Y. Ameen Ameen, Mohammed AM Sadeeq, and SubhiZeebaree. (2021) Multimodal emotion recognition using deep learning. Journal of Applied Science and Technology Trends 2, no. 02 pp 52-58.
[37] Subramanian, R. Raja, ChunduriSandyaNiharika, DondapatiUsha Rani, ParvathareddyPavani, and KetepalliPoojita Lakshmi Syamala. (2021) Design and Evaluation of a Deep Learning Algorithm for Emotion Recognition.5th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 984-988.
[38] Lee, Sanghyun, David K. Han, and HanseokKo. (2021) Multimodal Emotion Recognition Fusion Analysis Adapting BERT With Heterogeneous Feature Unification. IEEE Access 9,pp 94557-94572.
[39] Zhou, Ying, Yanxin Song, Lei Chen, Yang Chen, Xianye Ben, and Yewen Cao. (2019) A novel micro-expression detection algorithm based on BERT and 3DCNN. Image and Vision Computing 119 pp 104378.
[40] Graterol, Wilfredo, Jose Diaz-Amado, YudithCardinale, Irvin Dongo, Edmundo Lopes-Silva, and Cleia Santos-Libarino. (2021) Emotion detection for social robots based on nlp transformers and an emotion ontology. Sensors 21, no. 4 pp 1322.
[41] Baffour, Prince Awuah, Henry Nunoo-Mensah, Eliel Keelson, and Benjamin Kommey. (2022) A Survey on Deep Learning Algorithms in Facial Emotion Detection and Recognition. Inform 7 no. 1.
[42] Khattak, Asad, Muhammad ZubairAsghar, Mushtaq Ali, and UlfatBatool. (2022) An efficient deep learning technique for facial emotion recognition. Multimedia Tools and Applications 81, no. 2 pp 1649-1683.
[43] Canal, Felipe Zago, Tobias Rossi Müller, JhenniferCristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, ElianePozzebon, and Antonio Carlos Sobieranski. (2022) A survey on facial emotion recognition techniques: A state-of-the-art literature review. Information Sciences 582 pp 593-617.
[44] Rai, Mritunjay, Agha Asim Husain, Rohit Sharma, TanmoyMaity, and R. K. Yadav. (2022) Facial Feature-Based Human Emotion Detection Using Machine Learning: An Overview. Artificial Intelligence and Cybersecurity pp 107-120.
[45] Heredia, Juanpablo, Edmundo Lopes-Silva, YudithCardinale, Jose Diaz-Amado, Irvin Dongo, Wilfredo Graterol, and Ana Aguilera. (2022) Adaptive Multimodal Emotion Detection Architecture for Social Robots. IEEE Access 10 20727-20744.

[46] Khorrami, Pooya, Tom Le Paine, Kevin Brady, Charlie Dagli, and Thomas S. Huang. (2016) How deep neural networks can improve emotion recognition on video data. In 2016 IEEE international conference on image processing (ICIP), pp. 619-623.

[47] Sang-Hyuk L, Seong-Pyo C, Yountae K, Sungshin K (2006). ICIC'06 Proceeding. Part II: 557-569.