An Emotion Recognition System using Facial Expressions

Abstract—**Emotion refers to the display of a particular type of feeling or thought experienced by an individual as a result of changes in his or her immediate surroundings. Several emotion recognition systems have emerged in recent years as emotions play an important role in helping one understand the mental state of the people around them. The majority of the existing models use physiological signals such as EEG, ECG, GSR, BVP, temperature, etc. in order to detect and classify different emotions. These models are equipped with bulky hardware devices that are worn by the subjects in order to collect the relevant physiological signals and detect emotions. This dependence on hardware reduces the portability and ease of use to a great extent. This paper proposes a portable emotion recognition system designed on Jetson Nano that uses facial expressions for recognition. It recognizes seven different emotions which are - happy, sad, anger, Fear, disgust, surprise and neutral. Convolutional Neural Networks (CNN) was used to detect a face, extract facial expressions and later classify them into corresponding emotions. The model was most efficient in detecting the 'happy' emotion with 83.2% accuracy. This was followed by 'surprise' which had an accuracy of 82.4%. The average testing accuracy of the model was 71.54%.**

Keywords—CNN, Emotion recognition, facial expressions, Jetson Nano, Computer Vision

I. Introduction

Emotion refers to the display of a particular type of feeling or thought experienced by an individual as a result of changes in his or her immediate surroundings. Several emotion recognition systems, which are intelligent models that examine different aspects of a human body like facial expressions, physiological signals such as Electroencephalogram (EEG), Galvanic Skin Resistance (GSR), Electrocardiogram (ECG), Blood Volume Pressure, and temperature, etc., have emerged in recent years. This increase in popularity of the emotion recognition systems can be attributed to the importance of emotions in one's daily life as they play a primary role in understanding the mental state of the people around us. Emotions are also used as a medium to gain knowledge about a person's confidence in situations like criminal interrogations. Detection of emotions of customers and employees can help a company gain useful insights about their work environment hence allowing them to implement adequate changes. Several sentiment analysis based emotion recognition systems have also emerged in order to understand the emotions of people on social media sites like Twitter. The increasing popularity of emoticons while communicating via virtual platforms also suggests that recognition of emotions indeed plays a vital role in an individual's daily life.

Emotions are primarily linked to a person's facial expressions and examining these expressions to determine emotions plays a major role in immaculate interactions of human to human and machines. The state-of-the-art emotion recognition systems rely on physiological signals such as EEG, ECG, GSR, BVP, temperature, etc. to detect and classify emotions. A system that detected angry emotion [1] used four physiological signals i.e ECG, GSR, blood volume pulse and pulse to detect the emotion. The SVM algorithm was trained on the data obtained from the sensor. Another similar model used EMG, BVP, and GSR [2] to detect emotions. A pilot emotion recognition system [3] used EEG signals to track the change in a pilot's emotion during the course of the flight. The signals were captured using Epoch flex and EEG caps worn by the subjects. Positive, negative and neutral emotions were detected using ECG and Photoplethysmogram signals [4] which were later used to train a CNN-based model. While these systems may achieve high accuracy in terms of detecting emotions, they required the sensors, which were sometimes bulky, to be physically attached to the subjects reducing the system's portability and the ease of use to a great extent. Additionally, some of these state-of-the-art models required the subjects to sit in a particular position for several hours [1] in order to collect relevant sensor data. Special audio-video rooms were also needed in order to make the collected data accurate and noise-free. These conditions cannot be satisfied in many of the applications that require the detection of emotions in real-world scenarios. In order to successfully deal with the said challenges, this paper proposes a Computer Vision-based solution to detect and recognize emotions in real-time.

A model that classified facial expressions using CNN [5] and FER-2013 dataset recognized the seven basic human emotions, i.e., fear, anger, sad, surprise, happy, disgust and neutral. It used the haar cascade method to detect faces from images captured via webcam. The model showed an accuracy of 66% which could be increased by using more images in the training data. Another facial-expression-based detection system [6] extracted image frames from real-time videos which were followed by detection of a human face and recognition of emotions. It used the Multi-Layer Perceptron-based Neural Network in order to perform the said recognition. Some of the outputs categorized a single image as showing multiple emotions while in some cases, it failed to detect any emotion in the input image. Models like detection of distress [7] based on facial expressions that determine the distress level of a person are particularly helpful in fields of medicine and marketing, etc. A facial expression-based emotion recognition system [8] used a combination of Active Appearence Model (AAM) and Deep Belief Network (DBN) to detect emotions. It extracted features from image sequences and used Bayesian Network to map relationships between facial expressions and emotions and converted these mappings to a dynamic sequence. The model showed an accuracy of 96%. However, AAM was found to be heavily dependent on the number of facial landmarks and required an advanced structure in most of the cases which made the overall model complex. LDA-based facial emotion recognition system [9] used facial landmarks of grayscale images to detect emotions. The facial landmark features made the model highly accurate. Instead of recognizing prototypical emotions, [10] focused on the changes in facial muscles that are responsible for changing the display of emotions. This approach was found to work well even while dealing with side profile images.

II. Literature Survey

This Section discusses Emotion Recognition based on Physiological signals and Computer vision-based Systems for human emotion recognition. Physiological signals were used by many researchers. One of them, EEG based emotion recognition [11] has become significant for improving Human-to-Computer Interaction (HCI) systems' intelligence. In general, it used two major types of systems to recognize emotions: deep machine learning and shallow machine learning. Different types of deep learning, such as the Convolutional Neural Network (CNN), the Deep Neural Network (DNN), and the Deep Belief Network (DBN) were used. Recurrent Neural Networks (RNNs), Bimodal Deep AutoEncoders (BDAEs), Voting Ensembles (VEns), etc, were used for classification. A shallow learning-based classifier system used SVM classifier, Decision Tree classifier, Artificial Neural Networks (ANN), KNN classifier, Random Forests, Probabilistic Neural Networks and Multi-level Optimized.

A. Physiological signal based models

Deep transfer learning (DTL) models were able to detect choices from Electroencephalogram signals. Power spectral density (PSD) and of the EEG were extracted and valence the power spectral density. Each trial displayed 2,367 unique features of EEG signal [12]. The classifiers were evaluated based on different metrics (accuracy, recall, and precision) and through various validation procedures (k-fold cross-validation, holdout, LOOCV). Cohn–Kanade (CK+) and Japanese Female Facial Expression (JAFFE) datasets were used in this approach. It was trained using the Distance metric learning (DML) and 3-D CNN algorithms, and the test accuracy was (87%–96%), for Sparse-deep simultaneous recurrent network (DSRN) +softmax it was 90.82%. One advantage of S-DSRN was how it used the concept of weight sharing in the hidden layers. The major challenge was how the model was not able to accurately detect emotions when occlusions occurred, in-person human interactions and when the emotion was associated with micro-expressions [13]. A wearable device that detected emotions using distal facial Electromyogram signals to detect face motion in an unobtrusive manner as a replacement to using Computer Vision was suggested [14]. The EMG technique of video recognition was more robust against occlusions and changes in bright light had higher temporal resolution and didn't depend on movement. The wearable device was tested for its ability to recognize micro-smiles. Despite the speed and subtlety of the spontaneous expressions, the results demonstrated the potential of EMG. A wearable device was used to train or monitor the subject's smile while he or she is mobile [15]. Using the proposed interface, Quality of life (QoL) and wellbeing were measured over long periods of time by observing the facial expressions of people during activities in daily life. This system combined a head-mounted video camera with a gaze tracking system that detected facial movements. This allowed users to point and select using the gaze while using facial gestures. Users with limited hand mobility were considered while developing a wireless system [16] so that the device was not needed to be taken off when the user was moving away from the computer. The algorithms were used for determining eye and head orientations and also for detecting movements based on capacitance measurements along with those for mapping gaze direction to on-screen coordinates.

B. Computer-vision based models

Shepard Convolutional Neural Networks (SHCNN's) architecture classified both static and microexpressions simultaneously by employing only three layers and did not require large training datasets. Shepard Convolutional Neural Networks (CNN) architecture introduced a shrinkage factor derived from existing saliency maps and the vanishing gradient problem. datasets used were CASME, CASME II, FER2013, FERPlus, and Spontaneous Micro-Facial Movement (SAMM). The method was better on the datasets - FERPlus, CASME, CASME II, Facial Expression Recognition (FER2013), and SAMM than methods offering source code (or pseudocode) [17], FER2013 emotion recognition average accuracy without augmentation 64% and without regularization 68% and overall accuracy was 68%. Another emotion recognition model used the intensity-based EIDB-13 dataset [18], which contained 10393 facial images of 13 expressions. CNN, InceptionResNetV2CCBAM algorithms were used. The detected emotions were seven universal expressions. The accuracy obtained was 78%. Automated feature learning via deep convolutional neural networks model (DCNNs) using discriminant temporal pyramid matching (DTPM) for recognizing speech emotion [19] was proposed. DCNNs were used for feature extraction on segment level analyzing three discrete log-Mel-spectrograms, which corresponded to the representation in RGB. Berlin dataset EMO-DB, the RML, eNTERFACE05 and BAUM-1s audio-visual datasets gave accuracies of 87.31%, 69.70%, 76.56%, and 44.61% respectively. The algorithm used was the Discriminant Temporal Pyramid Matching (DTPM) and the emotions that were detected were seven universal expressions.

The Computer vision-based System for human emotion recognition was the Active learning (AL) CNN framework, active appearance model [20] and the Compound Dataset defined twenty-one different emotions that included the six basic and the rest of them being compound. They have used the optimization of sample selection technique (OSS). The training was done on 78% of the total labelled data. The model achieved an accuracy of 82.5% and 98.8% after optimization and incremental active phases respectively, which was better than the comparative deep neural networks. Autistic Meltdown Detector (AMD) [21] based model used Deep Spatio-temporal geometric keypoints and Meltdown Crisis dataset to detect emotions. Feed Forward (FF), Cascade Feed Forward (CFF), Recurrent Neural Network (RNN), and Long Short Term Memory (LSTM) were the algorithms used. The model was able to detect complex emotions such as - angrily disgusted, Sadly surprised, sadly fearful, etc. The accuracies of FF, CFF, RNN and LSTM were 80.5%, 81.0%, 84.2%, 67.4% respectively. Deep convolutional neural networks (CNN) algorithm for facial expressions from holistic features was used. The datasets used were FER2013, JAFFE, CK+, KDEF and RAF [22]. This proposed architecture examined the 25 different advanced algorithms like HOG-TOP, Major CNN, Shallow CNN on LBP images and gray-scale images. It also considers the partial VGG16 pre-trained model, weighted mixing of the double channel, and three partial networks. Appearance-based CNN trained using LBP images, Combination of geometric and appearance keypoints, facial features detection by 3D Inception-ResNet, Autoencoders and SON, Spatio-temporal features LSTM, Ensemble DNNs, binary classification using DCNN, ICANN, 2B(N+M) Softmax, SDSRN, Gabor transform on Color features, DCMACNNs, and Ensemble MLCNNs. The seven universal emotions were recognized. FER2013 showed an accuracy of 78%, JAFFE and CK+ was 98%, KDEF showed 96% accuracy, and RAF was 83%.

Natural Language Processing (NLP)-based sentiment analysis techniques, Deep Long Short-Term Memory (DLSTM) [23] were used to detect emotions. The dataset used was Sentiment140 and Emotional Tweets dataset wherein 11,110 tweets depicted the positive emotions , and 5,674 depicted negative emotions such as sadness, disgust, anger and fear. The remaining tweets were a mixture of positive and negative emotions. The training accuracy of the model was 96% and validation accuracy was 81%. The accuracy of the tweets was found to be 90%. This architecture was unable to grasp the context, especially in a sarcastic tweet. Also, any image containing the text couldn’t be processed by the model.

Edge computing device based NVIDIA Jetson TX2, Raspberry Pi was used for emotion recognition. Convolutional Experts Constrained Local Model (CECLM) used the CK+ dataset and RAF-DB dataset, which contained emotions like neutral, surprised, Joy, anger, sadness, fear and disgust. Light CNN algorithm was used. The Action Unit (AU), which was defined by the Facial Action Coding System (FACS), categorized emotions on the basis of facial muscle movements. DATASET CK+ achieved a 93.85% accuracy, RAF-DB was able to achieve a test accuracy of 81.05%. In the edge computing process, the main advantage was less data exchange with the server, making it computational effective making it cheaper. The accuracy was minimal when compared with the server, normal computers are capable of giving 30% more accuracy. A Dynamic approach to recognise facial expression based on a spatiotemporal deep learning model [24]-[25] used DISFA, MMI and CK+ datasets. Their accuracies were 71%, 88.9% and 97% respectively. The algorithms used were 3D-Convolutional Neural Networks (3D-CNNs) and Convolutional-Long-ShortTerm-Memory (ConvLSTM). The proposed model detected emotions of Neutral, Onset, Apex and Offset phases. The strength of this technique was how it was able to reveal patterns and features efficiently of the entire sequence without being affected by time variations and spatial disturbances. Adding to the strengths the network was able to retain itself despite database change.

A 3D-hybrid deep model based Happy emotion Recognition architecture used distant features (HappyER-DDF) [26]. The datasets used were AM-FEDC, AFEW, and MELD datasets and their accuracies were 95.97%, 94.89% and 91.14% respectively. The facial keypoints were processed by a hybrid deep neural network. It contained a combination of 3D Inception-ResNet, LSTM and CNN. Happy and non-happy expressions were detected by this model. The issue of CNN being sensitive to input image shapes was resolved including spatial transformer network block technique. A weighted mixture deep neural network (WMDN) [27], was proposed for the FER tasks-based VGG16 model. Datasets used were CK+, JAFFE,” and Oulu-CASIA. The seven emotions were detected . CK+, JAFFE,” and Oulu-CASIA, dataset, and the recognition accuracy was 96.68%, 92.2%, and 92.3%, more accuracy could have been achieved but the lack of training samples bottlenecked the performance. A unique strategy for detecting pain using DCNN and CNN algorithms [28] was proposed. The UNBC dataset was used, and the target emotions were painful, becoming painful, and extreme pain. The precision achieved for classifying non pain emotions was 99.75%, when they were training the model with a small batch of images they achieved an accuracy of 92.93% but on tweaking the batch size they achieved an accuracy of 95.15%. This method finds its usefulness in hospitals where the emotions of patients can be understood when they don't express it. The increase in accuracy for classifying pain emotions was achieved by using larger batch size. Additionally, the model was able to perform well with a fewer number epochs. Face expression recognition was performed which was based on S-DSRN for robustness. Sparse-deep simultaneous recurrent network (DSRN) used dropout learning to obtain feature sparsity as opposed to manually specifying the penalty terms to account for sparse representations of data [29].

College students with depression experienced the rise because of the growing competition at universities, so researchers proposed mining the text of Sina Weibo data of college students by storing what topics they were reading to detect depression. Feature extraction was done with DNN. Discretely integrated SVM were utilized to classify the input data in order to identify depression. DISVM algorithm approach was able to improve the accuracy of depression diagnosis to some extent and made recognition models more stable. According to simulation experiments, the depression recognition algorithm detected potential patients who might be suffering from depression among college students using Sina Weibo data [30]. Another model used three general datasets: the 300W-LP dataset, the AFLW2000 dataset, and the NIMH-ChEFS dataset. Results were comparable to those from advanced algorithms such as the AADL technique, i.e. anisotropic angle distribution learning, JFA algorithm (joint head pose estimation and face alignment) and RAFA-Net [31].

Face recognition-based emotion expression system [32]. Started by identifying the peak expression frame by using a double local binary pattern. The DLBP’s small dimensional size made it possible to process the data fast and significantly reduced the detection time. A logarithm-Laplace (LL) domain was also proposed for handling variations in illumination in LBP, resulting in a more powerful feature available for detecting faces. A Taylor Feature Pattern (TFP) was derived from the Taylor Feature Map based on the LBP and Taylor expansion. In terms of extracting facial expressions, the TFP method appeared superior. A detailed face representation system based on expression decoding was proposed [33]. It used AAM for facial subregion modeling using a system that utilized facial features, facial coding and expression recognition. It then located and tracked facial features. Though the conditions for 3D pose, lighting, and other features were flexible, the results suggested a potential failure for use in real-life situations. The WearCam technology , which provided key features like unobtrusiveness and portability enabled users to recognize the dynamics of optical behavior when they engaged in complex human interactions, such as playing with children. WearCam's ability to detect extreme attention with eyes still and shifting attention by movement of eyes in unconstrained settings could be examined in future studies of ASD. The wearable conveyed emotions like happiness and bliss. The term primary subject was alternatively being utilized in the context of this system, with visually impaired individuals as the essential subject. Furthermore, when a third person smiled or was content with regards to a conversation, the wearable took note of the electrical activity in their brain and predicted the happy quotient using machine learning. In addition, the primary subject's wearable was employed with a buzzer which was triggered when the third person smiled or showed contentment [34]-[35].

The systems discussed above are currently being used to detect and recognise emotions. They use advanced deep learning based architectures which make them computationally complex and heavily dependent on huge and accurate datasets for training the said deep learning algorithms. The system proposed in this paper used a CNN architecture with three hidden layers thus reducing the overall complexity, as the end goal of this system was its integration with a portable and lightweight embedded system. The model also performed well in real-time and provided accurate results thus making it a decision assistive system.

III. METHODOLOGY

This paper proposes a framework for detecting the emotions of humans using facial expressions. The seven different emotions that were recognised were (i) angry, (ii) disgust, (iii) fear, (iv) happy, (v) sadness, (vi) surprise and (vii) neutral. The process of emotion detection involved three steps, the first being face detection, second was the feature extraction and the last was facial expression classification. All of these processes were performed on CNN (Convolutional Neural Network) algorithm because of its superior performance on image datasets and live video feeds. The block diagram of the proposed approach is shown in Fig.1. The proposed system took one frame of image from the live video feed of the camera. This frame of image was processed by the Jetson Nano by passing it through a Convolutional Neural Network model (CNN) that was ported on the board. The CNN model identified the emotion of that particular frame by passing it through various filters and displayed the output on the screen of the emotion detected. The CNN model was synthesised on Google Colab using GPU acceleration. The figure given below represents the block diagram of the proposed system.

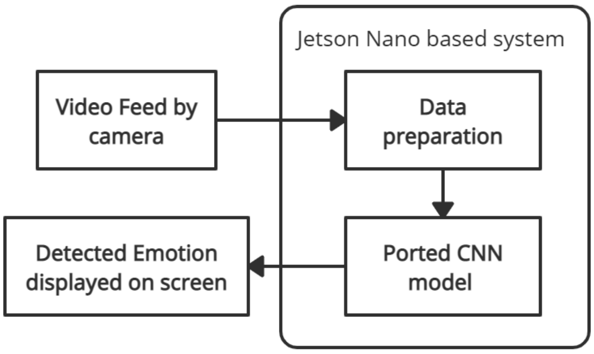

Fig.1. Proposed system architecture.

A. Dataset and Preparation

The dataset used was the FER-2013 dataset which was taken from Kaggle. In the dataset, there were a total of 35,887 images of 7 different emotions. All the images in the dataset were in grayscale with an orientation of 48 pixels x 48 pixels. Distribution of images in the dataset is presented in Table 1.

TABLE I: DATASET

Emotions	**Total images (Training + Validation)**
Angry	4953
Fear	5121
Happy	8987
Neutral	6198
Sad	6077
Surprise	4002
Disgust	547

Sample images from the dataset are shown in the figure below. The image size was (48,48) pixels and the images were grayscale.

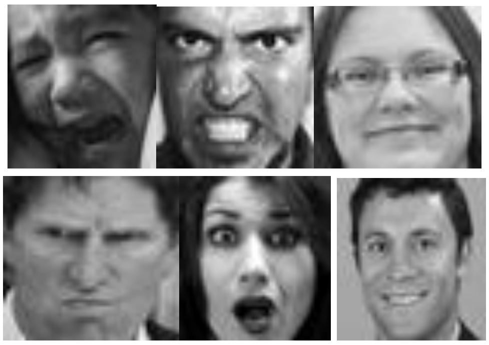

Fig.2. Sample dataset images.

The image set was processed and made suitable for the model to train on. A training image dataset was generated using the ImageDataGenerator class from the tensorflow library which augmented the images in real-time by applying random transformations on each image as it was passed to the model. Further, batches of 64 images were created which were fed to the model for training. The training data set of the model consisted of 28,709 images which belonged to the 7 emotions. The same process was used for creating the testing dataset for which a total of 7178 images were generated belonging to the 7 classes.

The data preparation process is depicted in Algorithm 1:

Algorithm 1: Image batch generation.

Input: Dataset images

Output: Batch of 64 transformed images

```
Initialization:
Define number of images in one batch (64)
Set height and width to resize the image (48,48).
Generate images with different transformations(ImageDataGenerator (Rescale=1./255)
Add these images to the batch
    Set color mode for the images (grayscale)
    Define the class of the images (categorical).
Pass this batch to the model for training/testing.
return batch of 64 images
```

B. Feature Selection and Classification of images

CNN (Convolutional Neural Network) model was used for classifying the emotions because of its higher accuracy when working with image datasets. There were many important parameters that needed to be specified before the model could be trained. First were the filters that the convolutional layer would learn, filters acted as spatial pattern detectors for example, edges. With this, it created a feature map that showed the presence of detected features in the image. The first stage of the CNN layer was creating a convolutional window for defining the parts of the images whose features need to be extracted. More layers were added after the input layer to the network for better feature extraction and network learning. Each layer was followed by a pooling method to get rid of unwanted data from the convolutional vector. This process is depicted in Fig.3.

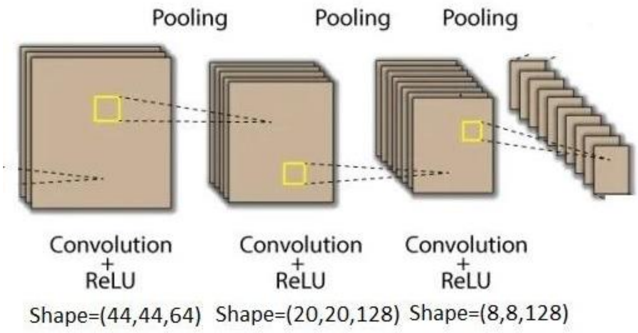

Fig.3. Feature extraction layers.

The feature map generated was a three dimensional vector, which was reduced to a single dimension using the flatten method available in Tensorflow for better absorption of data by the neurons in the network. The flattened layer was followed by a last layer of 7 neurons that was defined because we had 7 emotions to detect. The final output layer is depicted in Fig.4.

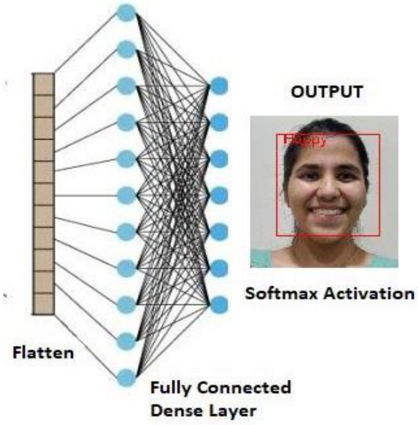

Fig.4. Output Connected layer.

The model started with the first layer having 32 filters, the second layer having 64 filters, third and the fourth layers having 128 filters each. The number of filters increased as the layers kept increasing because the layers which were early in the network learned fewer convolution filters than the ones which were deeper in the network. Each layer was followed by a pooling method in the form of max-pooling that was used to downsample the feature map created by the filter. Which meant reducing the spatial dimensions of the data.

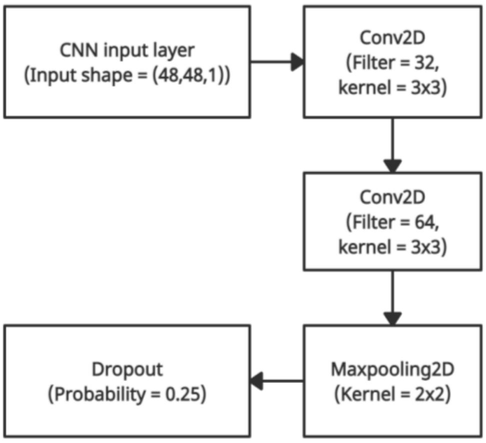

Fig.5. Input layer of CNN.

Having defined the filter size, the kernel size needed to be specified. Kernel size was a very important parameter in the model because it specified the height and width of the convolutional window, which defined how much part of the image will be scanned. The output dimension that was generated after running all the layers of the CNN was a 3 dimensional vector. So to bring the multidimensional tensor into a single dimension, the flatten function of Keras was used, this also made sure that every neuron in the network got the data efficiently. In order to increase the output shape of the CNN network we added a neural network layer with 1024 neurons to increase the dimensionality of the data. The prevention of overfitting was done by using the dropout function of Keras that was used after every layer. The final output layer was created with 7 neurons because we had 7 emotions to classify.

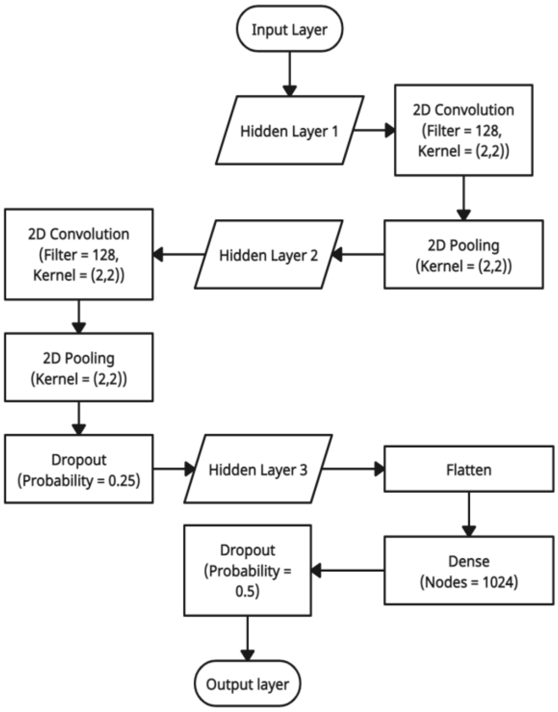

Fig.6. Hidden layer architecture.

Each hidden layer kept on increasing the dimension of the data which improved the learning of the network. The next section discusses about the results and functional evaluation of the facial emotion recognition model.

IV. Results and Functional Evaluation

The dataset was divided in a 80%-20% ratio for training and validation. Tha validation dataset was used while fitting the model to evaluate its training parameters and also to modify the learning rate accordingly. The validation process was carried out on 7178 images which made up 20% of the dataset. The testing consisted of three different parts, i.e. validation, testing on images and real-time testing. This process can be seen in Fig.7.

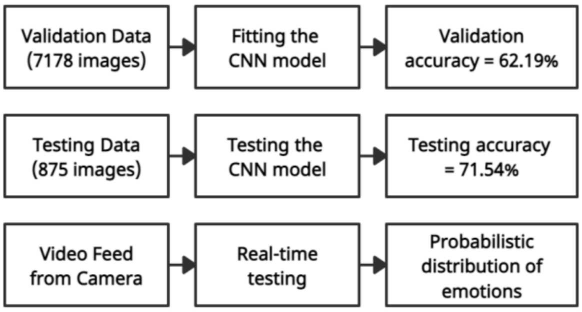

Fig.7. Testing process

The testing was carried out on 875 images. The test dataset contained 125 images each for the seven emotions, i.e. angry, disgust, fear, happy, neutral, sad, surprise. The highest accuracy has been observed for the happy emotion. It classified the images with an accuracy of 83.2%. This was followed by the emotion 'surprise', which classified with an accuracy of 82.4%. The lowest accuracy is while predicting the disgust and fear emotions. This was because of the indistinguishable expression of the people that cannot be classified clearly. Most of the fear test images were classified as surprised. The table below describes the total number of correct and incorrect classifications of the seven emotions.

TABLE II: Classification Table of Testing Images

Emotion	***Accuracy***
Angry	74.4%
Disgust	64.8%
Fear	49.6%
Happy	83.2%
Neutral	72%
Sad	74.4%
Surprise	82.4%

The model was evaluated using the metrics accuracy and loss was evaluated by calculating the categorical cross-entropy. This is represented by equation (i)

$$CE = -\sum_{neurons=1}^{classes} y_{true} * ln(y_{pred}) \text{ (i)}$$

In this case, the number of classes was 7, and y was the label for each emotion ranging from 0 to 6. The average testing accuracy was 71.54%. The categorical cross-entropy was 1.22.

The next part of the testing was carried out using a real-time video feed. The process followed for the real-time testing is described in algorithm 2.

Algorithm 2: Algorithm for real-time testing

```
Input: Video feed
Output: Frames with detected emotion
    Initialization
Read frames from video
Define haar cascade frontal face classifier
Convert frames to grayscale
Detect faces
    Loop Process
for (x,y,w,h) in faces do
        Display rectangle on detected face
        Get region of interest i.e. face
        Predict emotion on cropped image
        Display emotion on detected face
end for
return frame with detected emotion
```

The output of the above testing is shown in the Fig.8. A total of 50 real-time experiments were done. In these experiments, the emotions - happy, surprise and neutral were detected with full confidence every time. Rest of the emotions, i.e. anger, sadness, disgust and fear were classified as roughly 70% of the experiments.

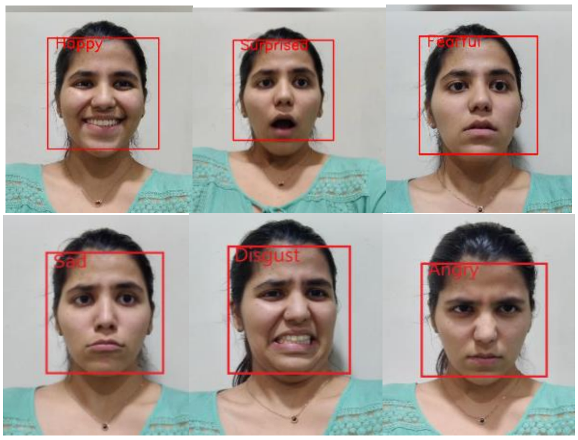

Fig.8. Output of real-time emotion detection

The more easily distinguishable emotions were happy, surprised and neutral. The emotions of sadness and fear were particularly difficult because of the variance in the intensity of expressions of different people. The probability of every emotion was added to the output to provide a better solution. This enabled the end-user to get a better understanding of what percentage of emotion the model was detecting. The probabilistic distributions of the detected emotions given by the model can be seen in Fig.9.

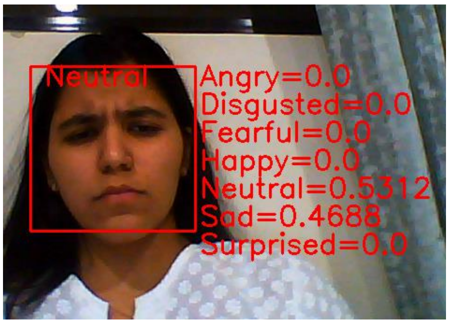

Fig.9. Probabilistic distribution of emotions

The training images contained all different types of emotions with male, female faces of many age groups. This facilitated the accurate prediction of emotions in real-world situations.

V. Conclusion

This paper provided an overview of an efficient computer vision system for the detection of emotions based on human expressions. The proposed system used convolutional neural network techniques to make accurate predictions on the input image. It was a camera-based system that recorded a live video feed to give input to the model. The model then processed the video frames and gave the detected emotion as an output. The proposed model classified seven basic emotions, i.e. angry, disgust, fear, happy, neutral, sad, surprise with an accuracy of 71.54%. The system first detected the face using the Haar-cascade detection algorithm and then predicted the emotion on the captured region of interest.

This system was deployed very conveniently on embedded systems like the Jetson Nano board along with a camera. The distinctive feature of this system was the minimal usage of hardware. This made the system very convenient and portable.

The utility of the system in the real world is very high because emotion detection plays an important role in the social interactions of people. It also enables various machines based on artificial intelligence to process this social information and interact with humans easily. The accuracy of the model can be increased by tuning the layer parameters and testing the images with different face detecting algorithms. One additional layer of RNN can be further added to the CNN model to improve the performance as R-CNN is popularly used for image description and captioning.

References

[1] C. Chang, Y. Lin and J. Zheng, "Physiological Angry Emotion Detection Using Support Vector Regression," 2012 15th International Conference on Network-Based Information Systems, 2012, pp. 592-596, doi: 10.1109/NBiS.2012.78.

[2] C. Joesph, A. Rajeswari, B. Premalatha and C. Balapriya, "Implementation of physiological signal based emotion recognition algorithm," 2020 IEEE 36th International Conference on Data Engineering (ICDE), 2020, pp. 2075-2079, doi: 10.1109/ICDE48307.2020.9153878.

[3] F. Wei, D. Wu and D. Chen, "An investigation of pilot emotion change detection based on multimodal physiological signals," 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT, 2020, pp. 1029-1034, doi: 10.1109/ICCASIT50869.2020.9368711.

[4] C. -J. Yang, N. Fahier, W. -C. Li and W. -C. Fang, "A Convolution Neural Network Based Emotion Recognition System using Multimodal Physiological Signals," 2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan), 2020, pp. 1-2, doi: 10.1109/ICCE-Taiwan49838.2020.9258341.

[5] R. Jaiswal, "Facial Expression Classification Using Convolutional Neural Networking and Its Applications," 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), 2020, pp. 437-442, doi: 10.1109/ICIIS51140.2020.9342664.

[6] D. Dagar, A. Hudait, H. K. Tripathy and M. N. Das, "Automatic emotion detection model from facial expression," 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 2016, pp. 77-85, doi: 10.1109/ICACCCT.2016.7831605.

[7] P. Nair and S. V., "Facial Expression Analysis for Distress Detection," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1652-1655, doi: 10.1109/ICECA.2018.8474761.

[8] [K. Ko and K. Sim, "Development of a Facial Emotion Recognition Method Based on Combining AAM with DBN," 2010 International Conference on Cyberworlds, 2010, pp. 87-91, doi: 10.1109/CW.2010.65.

[9] L. Sun, J. Dai and X. Shen, "Facial emotion recognition based on LDA and Facial Landmark Detection," 2021 2nd International Conference on Artificial Intelligence and Education (ICAIE), 2021, pp. 64-67, doi: 10.1109/ICAIE53562.2021.00020.

[10] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 36, no. 2, pp. 433-449, April 2006, doi: 10.1109/TSMCB.2005.859075.

[11] M. R. Islam et al., "Emotion Recognition From EEG Signal Focusing on Deep Learning and Shallow Learning Techniques," in IEEE Access, vol. 9, pp. 94601-94624, 2021, doi: 10.1109/ACCESS.2021.3091487.

[12] M. S. Aldayel, M. Ykhlef and A. N. Al-Nafjan, "Electroencephalogram-Based Preference Prediction Using Deep Transfer Learning," in IEEE Access, vol. 8, pp. 176818-176829, 2020, doi: 10.1109/ACCESS.2020.3027429.

[13] H. A. Gonzalez, S. Muzaffar, J. Yoo and I. M. Elfadel, "BioCNN: A Hardware Inference Engine for EEG-Based Emotion Detection," in IEEE Access, vol. 8, pp. 140896-140914, 2020, doi: 10.1109/ACCESS.2020.3012900.

[14] M. Perusquía-Hernández, M. Hirokawa and K. Suzuki, "A Wearable Device for Fast and Subtle Spontaneous Smile Recognition," in IEEE Transactions on Affective Computing, vol. 8, no. 4, pp. 522-533, 1 Oct.-Dec. 2017, doi: 10.1109/TAFFC.2017.2755040.

[15] A. Gruebler and K. Suzuki, "Design of a Wearable Device for Reading Positive Expressions from Facial EMG Signals," in IEEE Transactions on Affective Computing, vol. 5, no. 3, pp. 227-237, 1 July-Sept. 2014, doi: 10.1109/TAFFC.2014.2313557.

[16] V. Rantanen et al., "A Wearable, Wireless Gaze Tracker with Integrated Selection Command Source for Human-Computer Interaction," in IEEE Transactions on Information Technology in Biomedicine, vol. 15, no. 5, pp. 795-801, Sept. 2011, doi: 10.1109/TITB.2011.2158321.

[17] S. Miao, H. Xu, Z. Han and Y. Zhu, "Recognizing Facial Expressions Using a Shallow Convolutional Neural Network," in IEEE Access, vol. 7, pp. 78000-78011, 2019, doi: 10.1109/ACCESS.2019.2921220.

[18] K. Zheng, D. Yang, J. Liu and J. Cui, "Recognition of Teachers' Facial Expression Intensity Based on Convolutional Neural Network and Attention Mechanism," in IEEE Access, vol. 8, pp. 226437-226444, 2020, doi: 10.1109/ACCESS.2020.3046225

[19] S. Zhang, S. Zhang, T. Huang and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," in IEEE Transactions on Multimedia, vol. 20, no. 6, pp. 1576-1590, June 2018, doi: 10.1109/TMM.2017.2766843.

[20] S. Thuseethan, S. Rajasegarar and J. Yearwood, "Complex Emotion Profiling: An Incremental Active Learning Based Approach With Sparse Annotations," in IEEE Access, vol. 8, pp. 147711-147727, 2020, doi: 10.1109/ACCESS.2020.3015917.

[21] S. K. Jarraya, M. Masmoudi and M. Hammami, "Compound Emotion Recognition of Autistic Children During Meltdown Crisis Based on Deep Spatio-Temporal Analysis of Facial Geometric Features," in IEEE Access, vol. 8, pp. 69311-69326, 2020, doi: 10.1109/ACCESS.2020.2986654.

[22] K. Mohan, A. Seal, O. Krejcar and A. Yazidi, "Facial Expression Recognition Using Local Gravitational Force Descriptor-Based Deep Convolution Neural Networks," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-12, 2021, Art no. 5003512, doi: 10.1109/TIM.2020.3031835.

[23] A. S. Imran, S. M. Daudpota, Z. Kastrati and R. Batra, "Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets," in IEEE Access, vol. 8, pp. 181074-181090, 2020, doi: 10.1109/ACCESS.2020.3027350.

[24] J. Yang, T. Qian, F. Zhang and S. U. Khan, "Real-Time Facial Expression Recognition Based on Edge Computing," in IEEE Access, vol. 9, pp. 76178-76190, 2021, doi: 10.1109/ACCESS.2021.3082641.

[25] D. Al Chanti and A. Caplier, "Deep Learning for Spatio-Temporal Modeling of Dynamic Spontaneous Emotions," in IEEE Transactions on Affective Computing, vol. 12, no. 2, pp. 363-376, 1 April-June 2021, doi: 10.1109/TAFFC.2018.2873600.

[26] N. Samadiani, G. Huang, Y. Hu and X. Li, "Happy Emotion Recognition From Unconstrained Videos Using 3D Hybrid Deep Features," in IEEE Access, vol. 9, pp. 35524-35538, 2021, doi: 10.1109/ACCESS.2021.3061744.

[27] B. Yang, J. Cao, R. Ni and Y. Zhang, "Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based on Double-Channel Facial Images," in IEEE Access, vol. 6, pp. 4630-4640, 2018, doi: 10.1109/ACCESS.2017.2784096.

[28] K. Pikulkaew, W. Boonchieng, E. Boonchieng and V. Chouvatut, "2D Facial Expression and Movement of Motion for Pain Identification With Deep Learning Methods," in IEEE Access, vol. 9, pp. 109903-109914, 2021, doi: 10.1109/ACCESS.2021.3101396.

[29] M. Alam, L. S. Vidyaratne and K. M. Iftekharuddin, "Sparse Simultaneous Recurrent Deep Learning for Robust Facial Expression Recognition," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 10, pp. 4905-4916, Oct. 2018, doi: 10.1109/TNNLS.2017.2776248.

[30] Y. Ding, X. Chen, Q. Fu and S. Zhong, "A Depression Recognition Method for College Students Using Deep Integrated Support Vector

Algorithm," in IEEE Access, vol. 8, pp. 75616-75629, 2020, doi: 10.1109/ACCESS.2020.2987523.

[31] T. Singh, M. Mohadikar, S. Gite, S. Patil, B. Pradhan and A. Alamri, "Attention Span Prediction Using Head-Pose Estimation With Deep Neural Networks," in IEEE Access, vol. 9, pp. 142632-142643, 2021, doi: 10.1109/ACCESS.2021.3120098.

[32] Y. Ding, Q. Zhao, B. Li and X. Yuan, "Facial Expression Recognition From Image Sequence Based on LBP and Taylor Expansion," in IEEE Access, vol. 5, pp. 19409-19419, 2017, doi: 10.1109/ACCESS.2017.2737821.

[33] I. Bacivarov, P. Corcoran and M. Ionita, "Smart cameras: 2D affine models for determining subject facial expressions," in IEEE Transactions on Consumer Electronics, vol. 56, no. 2, pp. 289-297, May 2010, doi: 10.1109/TCE.2010.5505930.

[34] S. Das, "A novel Emotion Recognition Model for the Visually Impaired," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1-6, doi: 10.1109/I2CT45611.2019.9033801.

[35] S. Magrelli et al., "A Wearable Camera Detects Gaze Peculiarities during Social Interactions in Young Children with Pervasive Developmental Disorders," in IEEE Transactions on Autonomous Mental Development, vol. 6, no. 4, pp. 274-285, Dec. 2014, doi: 10.1109/TAMD.2014.2327812.