MiniGenie: Emotion Detection from Video using CNN Architecture

1st Aashnan Rahman
Department of CSE
Islamic University of Technology(IUT)
Gazipur, Bangladesh
aashnanrahman@iut-dhaka.edu

2nd Sumit Alam Khan
Department of CSE
Islamic University of Technology(IUT)
Gazipur, Bangladesh
sumitalam@iut-dhaka.edu

3rd Sherajul Arifin
Department of CSE
Islamic University of Technology(IUT)
Gazipur, Bangladesh
sherajularifin@iut-dhaka.edu

***Abstract*—Emotions play a pivotal role in human communication and interaction. Being able to precisely detect and understand emotions has a significant effect on several fields that includes psychology, HCI, affective computing, and so on. Emotion detection, a branch of sentiment analysis, refers to the method of identifying and comprehending human emotions from facial expressions, speech, and other behavioral cues. Such a task can be quite challenging, as it requires specific context to be properly understood and identified. Machine Learning is highly capable of improving the accuracy of emotion detection by leaps and bounds. Emotion detection is a rapidly developing field, and here in this paper, we try to propose a method to accurately detect human emotion from video analyzing facing expressions and voice tone with the view of ensuring the psychological well-being of an individual.**

***Index Terms*—sentiment analysis, machine learning**

I. Introduction

Emotion detection is one of the buzzwords in the modern era. The technology has the potential to significantly transform several industries and make human-computer interactions better in a variety of areas. By analyzing human emotions, technology can become more empathetic, personalized, and responsive to human needs and preferences. Emotion detection has a diverse array of applications such as marketing and advertising, education, public safety, and so on.

Our project deals with emotion detection using sophisticated Machine Learning algorithms and taking data from the user. In this paper, we a method has been proposed to detect human emotion from video feeds by extracting features from both facial expressions and audio tones. Our project can detect seven cardinal human emotions such as anger, happiness, sadness, neutrality, and many more.

By combining image-based emotion detection and voice tone analysis with machine learning techniques, this research aims to help develop more reliable emotion recognition systems. The above-discussed systems have the potential to improve an extensive variety of applications in different domains, allowing us to better understand and utilize human sentiments.

In our project we have worked with two broad categories of Emotion Recognition, they are briefly discussed below -

A. *Emotion Detection (ED)*

Emotion detection pertains to the procedure of recognizing and examining human emotions utilizing a range of approaches, typically involving technology and artificial intelligence. The objective is to comprehend and interpret emotional conditions by observing facial expressions, vocal intonation, body movements, and other pertinent indicators.

*1) **Facial Emotion Recognition(FER)**:* This technology identifies and analyzes the emotions expressed by a person's face. Computer vision and ML approaches are used here for the fine-tuning of the model. The typical procedure of facial emotion detection entails the capture of images or videos of an individual's face, followed by the analysis of diverse facial characteristics, including the position of eyebrows, eye movements, mouth shape, and overall muscle activity in the face. These characteristics are subsequently compared and aligned with predetermined patterns or models that represent various emotional states.

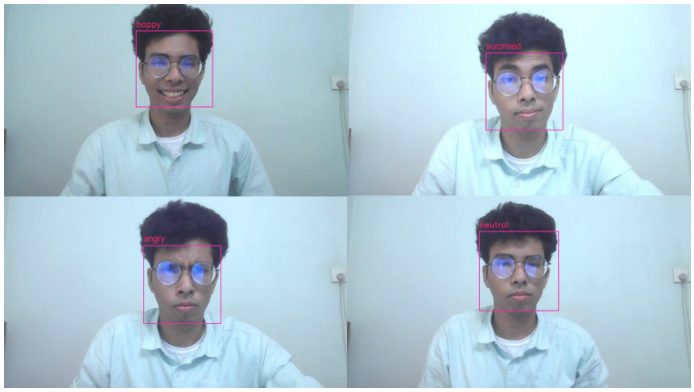

Fig. 1. Facial Emotion Detection Using CNN

*2) **Speech Emotion Recognition (SER)** :* Speech emotion detection is a field of research and technology that focuses on analyzing and identifying human emotions based on speech signals. It involves using algorithms and machine learning techniques to detect and interpret emotional states expressed through speech patterns, prosody, and acoustic features.

The process of speech emotion detection typically begins with the collection of speech data, which can be in the form of recordings or real-time input. Preprocessing is conducted on the input audio signals to extract meaningful features such

as pitch, intensity, speech rate, spectral characteristics, and pauses. These features are analyzed and compared against predefined models or patterns that represent different emotional states.

Both FER and SER technologies are a crucial part of sentiment analysis for our project.

In the following segments we have discussed the Literature Review, Methodologies including Data Acquisition, Data Preprocessing and Proposed Method. Moreover we have included the Experimental Setup, Result Calculation, Result Analysis and Comparison. Finally we concluded our paper with the Challenges that we encountered during this project and the Future scopes of this work.

II. Literature Review

Extensive research has been undertaken in the field of Emotion Detection over the recent years to investigate and analyze human emotions, aiming to enhance different facets of human existence.

Ozka Ezerceli and M.Taner Eskil implemented FER on FER2013 and CK+ datasets using CNN architecture and supportive techniques [1].

In a Kaggle competition, the participants were tasked with working on the FER2013 dataset. This dataset was specifically introduced for the competition, and it posed a challenge for traditional approaches as they struggled to achieve satisfactory accuracy rates. The leading triumvirate in the competition utilized Convolutional Neural Networks (CNN) in conjunction with image metamorphoses to tackle this formidable task. The ultimate victor, Y. Tang, harnessed the primal formulation of Support Vector Machines (SVM) as a loss function during the training process, while also leveraging the L2-SVM loss function.

This approach yielded remarkable outcomes during its inception, attaining a remarkable accuracy of 71.2% to claim the pinnacle in the competition. [2].

In addition to that FER and DeepFace libraries are two libraries that contain built-in models to detect FER.

The FER library incorporates methods and package structures that have been adapted or derived from the implementation of MTCNN by Iv án de Paz Centeno and the facial expression recognition repository by Octavio Arriaga. [3]

In contrast, the DeepFace library presents a streamlined framework for Face Recognition and Facial Attribute Analysis (including Age, Gender, Emotion, and Race) in Python. This hybrid framework amalgamates cutting-edge models, such as VGG-Face, Google FaceNet, OpenFace, Facebook DeepFace, DeepID, ArcFace, Dlib, and SFace, to augment its face recognition capabilities. By integrating and harnessing these state-of-the-art models, the framework attains an elevated level of accuracy. Experimental results reveal that while human beings achieve a recognition accuracy of 97.53% [4]on facial recognition tasks (based on the Labeled Faces in the Wild (LFW) database), the aforementioned models have not only surpassed but also exceeded this level of accuracy.

Significant contributions can also be observed in the domain of Speech Emotion Recognition. Ayad Alsobhani, Hanaa M A ALabboodi, and Haider Mahdi have achieved remarkable accuracy in their implementation of Speech Recognition utilizing Convolutional Deep Neural Networks. [5]. Among multiple implementations of SER, Graves, Alex, and Mohamed also tried to implement speech recognition using RNN and LSTM with only 17.7% error [6].

Another known method of speech emotion recognition is the use of MLP. The MLP classifier refers to the Multilayer Perceptron classifier. The MLP classifier represents a prominent form of artificial neural network extensively employed for classification purposes in the field of machine learning. This classifier encompasses a layered structure comprising interconnected nodes, or neurons, arranged in a feedforward configuration. Such a method was adopted in a work by NN Poojary and his associates among others which is quite relevant to our project [7].

III. Methodology

A. *DATA ACQUISITION*

1) ***Image Data****:* Detection of emotion using facial expressions requires an adequate amount of front-facing face data with significant quality. That is why, FER-2013 dataset is used for the image section.

a) ***Facial Emotion Recognition(FER)-2013*** *[8] :* This dataset is a public dataset containing face images. The resolution of the images is 48×48 pixels and all the images are grayscale. In this context, the facial images were automatically aligned to ensure that the face appears relatively centered within each image. This alignment process aims to make the space occupied by the faces consistent across the entire image dataset. All of the images are classified and labeled into 7 emotion categories and denoted by an integer i.e. 0-angry, 1-sad, 2-disgust, 3-fear, 4-happy, 5-neutral, and, 6-surprise. The dataset is partitioned, allocating 88% of the data for training purposes, encompassing a substantial collection of 28,709 example images. The remaining portion, comprising 3,589 example images, is reserved for testing. Licensed under Open Data Commons, the usability of this dataset is 7.50 according to Kaggle. The dataset was originally created by searching for each of the emotions and their synonyms using Google and collecting the images.

2) ***Audio Data****:* For the audio section, four different datasets are used which were recorded by diversified actors. Later on, these datasets are merged together so that they can be properly utilized.

a) ***Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*** *[9]:* RAVDESS dataset is a huge dataset of 7,356 files containing 16-bit audio data. The portion of the RAVDESS dataset used in this project consists of 1440 files in which 24 actors have participated and each of them has 60 trials. The dataset compiles recordings of speech

from a group of 24 proficient actors, equally divided between 12 females and 12 males. As a result, two linguistically identical statements are uttered. The speech emotion dataset encompasses a neutral North American accent, encompassing a diverse range of seven distinct emotional expressions: serenity, elation, melancholy, rage, trepidation, astonishment, and revulsion. Each of these expressions is captured at two distinct levels of emotional intensity: standard and heightened. The sampling frequency all over the set was 48000 Hz.

b) ***Toronto emotional speech set (TESS)*** *[10]:* Gathered by the University of Toronto, the TESS dataset contains audio clips of two women(aged 26 and 64) who have expressed seven unique emotions. The participants spoke 200 target words within the carrier phrase "Say the word _" that was included in the dataset. The dataset consists of the following seven expressions: anger, disgust, happiness, fear, surprise, sadness, and neutral. If the whole dataset is considered, there are in total 2800 audio files or data points in the TESS dataset. The 7 cardinal emotions expressed by the two actresses are organized within their own folders. For the format of the audio files, WAV is used.

c) ***Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D)*** *[11]:* The CREMA-D dataset, an audio-visual dataset used for emotion recognition, consists of both auditory and vocal cues in sentences spoken by 91 actors where 48 of them were male, and 43 were female. In the dataset, a total of 7,442 [12] original clips are gathered. The age range of the participating actors is 20 to 74 and they come from a variety of races and ethnicities which include African American, Caucasian, Asian, Hispanic, and other unspecified races. In the dataset, actors spoke from 12 selected sentences, and six different emotions which were anger, disgust, fear, happiness, neutrality, and sadness.. The emotional levels for each emotion are categorized as Low, Medium, High, and Unspecified. Since the dataset was large, it was crowd-sourced by 2443 participants. The participants provided ratings for emotions and emotion levels in three distinct segments. Initially, the participants provided ratings for the audiovisual presentation as a whole, followed by separate ratings for the video alone and the audio alone. Each participant of the CREMA-D dataset rated 90 unique clips (30 audio-visual, 30 visual, and 30 audio). Licensed under Open Data Commons Attribution License, CREMA-D has a usability value of 8.75.

d) ***Surrey Audio-Visual Expressed Emotion (SAVEE)*** *[13]:* The SAVEE dataset is another emotion recognition dataset for audio data. The recording was conducted with the participation of four individuals who are native English speakers and all of them were male actors aged between 27 to 31. This dataset has 7 cardinal expressions of emotions which can be categorized as anger, disgust, fear, sadness, happiness, surprise, and neutral. In total 480 Utterances in British English were recorded and combined in SAVEE dataset. The dataset for this project had a Kaggle usability rating of 8.75. It comprised 15 standard TIMIT sentences per emotion, where there were three sentences commonly used, two sentences specific to expressing emotions, and ten generic sentences.. Throughout the dataset, a phonetical balance was maintained. The credit for accumulating the whole dataset goes to the University of Surrey.

B. DATA PREPROCESSING

1) ***Image Data****:*

- **Resize:** All the images of the dataset were resized into 48×48 pixels.
- **One Hot Encoding:** In this step, the text labels of each image were converted to a unique binary value.
- **Store:** The data are stored in a dataframe.
- **Split:** In the evaluation phase, the dataset is partitioned into distinct segments to facilitate training, testing, and validation, employing advanced lexicons.

2) ***Audio Data****:*

- **Data Augmentation:** Data augmentation was implemented to generate novel synthetic data instances through the introduction of minor perturbations to our initial training set. For the creation of syntactic audio data, we infused noise, applied temporal shifts, and manipulated pitch and speed. The objective was to enhance the model's robustness against such perturbations, elevating its capacity to generalize and effectively accommodate variations.
  - Noise: Random noise was added to the dataset.
  - Stretch: Time stretching was applied to the input audio data sequence based on the provided rate. This approach helps to modify the duration of the audio data while keeping its pitch intact.
  - Shift: Random shift in the time domain was performed.
  - Pitch: Pitch shifting was generated for the given input audio data.
- **Feature Extraction:** It's the process through which raw input data gets transformed into new numerical features. The original data can be preserved while making modifications to these numerical features. In most cases, after feature extraction, this processed data performs better.
  - Zero Crossing Rate (ZCR): ZCR serves as a measure of the polarity transition within an audio signal, denoting the frequency at which the audio waveform traverses the zero axis.
  - Mel Frequency Cepstral Coefficients (MFCC): These coefficients find application in the realms of audio, speech, and music recognition, wherein the spectral attributes are converted into discerning coefficients.
  - Root Mean Square (RMS): RMS, a prevalent data representation, embodies the square root of the average of the squared amplitudes of an audio signal,

imparting valuable insights into its overarching energy magnitude.
  - Mel Spectogram: In the context of audio analysis and classification, the Mel Spectrogram is employed as a visual representation illustrating the frequency spectrum of an audio signal across time.
  - Chroma STFT: Using Short Term Fourier Transformation, Chroma STFT represents the 12 different pitch classes in music and captures the harmonic content of an audio signal.
- **One Hot Encoding:** Audio data labels are converted to binary values.
- **Normalizing:** The data set is normalized using their mean.

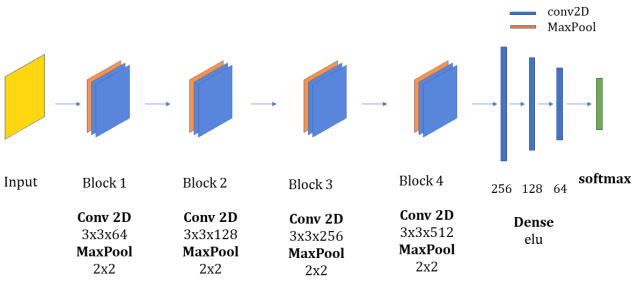

Fig. 2. CNN architecture for FER

C. *PROPOSED METHOD*

For our project, we have tried to come up with a solution that will make a bridge between two different models, FER and SER. At first, to extract meaningful and relevant features and capture the complex patterns inherent in both visual and auditory cues, we have utilized Convolutional Neural Networks (CNN) as our primary algorithm in both the FER and SER sections. The reasoning behind choosing CNN was its ability to effectively learn hierarchical representations from raw input data which enables robust feature extraction and discriminative modeling of emotional expressions.

*1) **CNN based FER Model**:* For FER (Facial Emotion Recognition) we created a CNN (Convolution Neural Network) model of 8 blocks. Each block is composed of a few layers.

The first 4 blocks are identical. They are composed of two 2D convolution layers followed by a 2D-max pooling layer. The size of the convolution layers was 3x3 having the same padding and eLU activation function, and the size of the Max Pooling filters was 2x2. The total number of filters in the first four blocks was 64, 128, 256, and 512. In each block, after every convolution layer Batch Normalization was applied and max pool layer was followed by dropout. Batch Normalization was added for faster convergence and stability of our model whereas a 20% Dropout layer was added to address high variance and to enhance the performance of our model.

Then the model was flattened and 3 dense layers were added of size 256, 128, and 64 with a non-linear activation function 'eLU'. Again these layers were also followed by Batch Normalization and 50% Dropout. Lastly, the softmax activation function was applied and in the last layer, the number of nodes or neurons was 7 each representing one type of emotion.

The total number of training and validation samples was 28,273 and 3,5273, and 64 was set as batch size. To fine-tune the model, we trained for 50 epochs, optimizing the cross-entropy loss using Root Mean Square Propagation (RMSProp). RMSProp adapts the learning rate of individual parameters by evaluating the historical gradients' magnitudes, employing an exponentially diminishing running mean of the squared gradients specific to each parameter. Checkpoints and early stopping were the two callbacks of our model. About 6 million parameters had to be learned in this model.

*2) **Deep-CNN based SER Model**:* For SER (Facial Emotion Recognition) we created a deep CNN (Convolution Neural Network) model consisting of 6 blocks.

Here too the first 4 blocks are identical. They consist of sequential 1D convolution layers, which are subsequently followed by a Max pooling layer. The size of the convolution layers was 5x5 having the same padding and stride value 1, ReLU activation function was applied, and the size of Max Pooling filters were 5x5 with stride 2 and the same padding. The number of filters in the first four blocks was 256, 256, 128, and 64. Each of the first four blocks had a 20% dropout value. Then the model was flattened and a dense layer having 32 units was applied with the non-linear activation function ReLU and a 30% dropout value. Finally, the softmax activation function was applied and in the last layer, the number of nodes or neurons were 8 each representing one type of emotion.

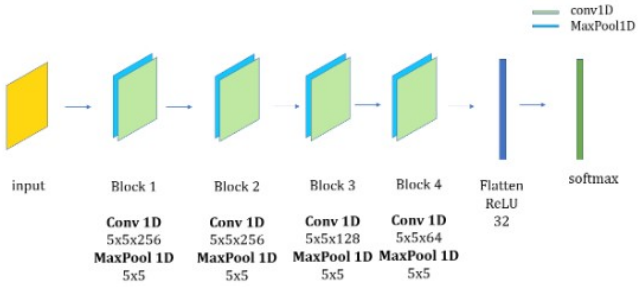

Fig. 3. Deep-CNN architecture for SER

The total number of training and validation samples was 27,364 and 9,122, and 64 was set as batch size. To fine-tune the model, we trained for 50 epochs, optimizing the cross-entropy loss using Adam (Adaptive Moment Estimation). The Adam optimizer adjusts the learning rates of each parameter

based on both the gradient and momentum of the gradients. It adapts the learning rates individually for each parameter, enabling them to be updated at varying rates according to their past gradients. Additionally, it demonstrates resilience to noisy gradients and sparse data. Reduce Learning Rate On Plateau was used as a callback with a factor of 0.4, a patience value of 2, and a learning rate of 0.0000001. RLRP is a technique that enables the model to dynamically modify the learning rate while training, leading to improved convergence and overcoming plateaus. It achieves this by reducing the learning rate when the model's progress becomes stagnant. By doing so, RLRP helps the model escape local minima and discover better optima within the loss landscape.

IV. Experimental Setup

The significant development in this project of ours was made possible with the involvement of a series of steps followed under a supervised environment. Our model is a supervised Machine Learning model that learned from the given labeled datasets. While working with FER, facial images from the FER-2013 dataset were preprocessed, ensuring consistent resolution and grayscale format. In a similar manner, for SER, we processed the audio files from the RAVDESS, TESS, SAVEE and CREMA datasets, maintaining a standardized sampling frequency of 48,000 Hz. The data was augmented and different techniques were used for the feature extraction part i.e. ZCR, MFCC, RMS etc.

The experimental setup involves conducting the training and evaluation processes on the Google Colab platform. There, a T4 GPU was utilized for accelerated computation. In this way, this powerful hardware resource facilitated efficient model training and inference, enabling us to effectively handle large-scale datasets and complex CNN architectures.

As mentioned earlier, during the training phase, a rigorous data augmentation strategy was applied to enhance model generalization. Later on, a loss function was implemented and two separate optimizers i.e. Adam and RMSprop were used to optimize the FER and SER models.

For each dataset, a portion of the data was separated for the testing phase. Those were utilized for assessing the effectiveness of our model. We measured the performance using one of the key evaluation metrics which is accuracy. In addition to all these, cross-validation techniques were employed to validate the robustness of our model.

V. Algorithm and Result Calculation

We split the video into frames. To reduce latency, one frame every 10 frames is used for prediction. For each prediction frame, the prediction is stored. Finally, we check which particular emotion occurs the most, and that is what we deem our final emotion.

The algorithm for audio is rather simple. We just convert the audio to mp4 format and feed it to the model, from where we get the prediction

Finally, to combine the 2 emotions together, we add the corresponding emotions and return the maximum.

Algorithm 1 FER prediction (no. of frames)

procedure FERPREDICTION(n)
  $count\_video\_emotion \leftarrow$ empty list
  for $frame\_no \leftarrow 1$ to n **do**
    if $frame\_no \mod 10 == 0$ **then**
      $emotion \leftarrow$ PREDICTFER($frame\_no$)
      $count\_video\_emotion$.add($emotion$)
  $video\_emotion \leftarrow$ argmax($count\_video\_emotion$)
  return $video\_emotion$

Algorithm 2 SER prediction (audio file)

procedure SERPREDICTION(audio_file)
  $count\_audio\_emotion \leftarrow$ PREDICTSER(audio_file)
  $audio\_emotion \leftarrow$ argmax($count\_audio\_emotion$)
  return $audio\_emotion$

Algorithm 3 Combined prediction

procedure COMBINEDPREDICTION(count_video_emotion, count_audio_emotion)
  emotion_probability $\leftarrow$ empty list
  for emotion_no $\leftarrow 1$ to 7 **do**
    probability $\leftarrow$ (count_video_emotion[emotion_no]+ count_audio_emotion[emotion_no])$/2$
    emotion_probability.add(probability)
  final_emotion $\leftarrow$ argmax(emotion_probability)
  return final_emotion

The process of combining the results obtained from the two algorithms and finalizing the output according to algorithm 3 is illustrated in Figure 4.

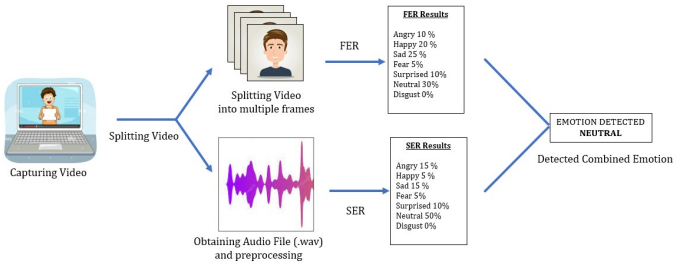

Fig. 4. Emotion Detection Steps

VI. Result and Accuracy Analysis

The accuracy values for training and validation of the CNN-based models utilized in Facial and Speech emotion detection are provided below.

TABLE I
TRAIN AND VALIDATION ACCURACY OF FER AND SER MODELS

	FER	SER
Train Accuracy	90.21%	75.88%
Validation Accuracy	64.85%	62.27%

We can see a graphical comparison of the above-stated comparison in the figure 5 and figure 6 -

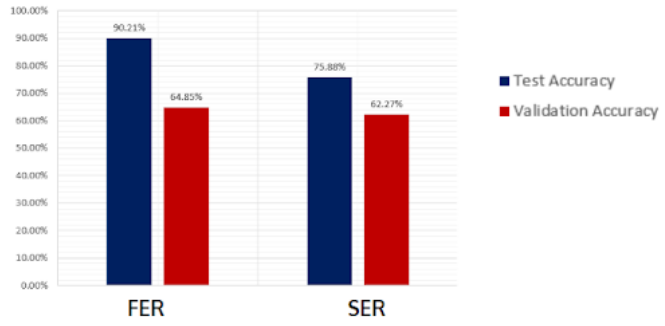

Fig. 5. Graphical comparison of FER and SER model's accuracy

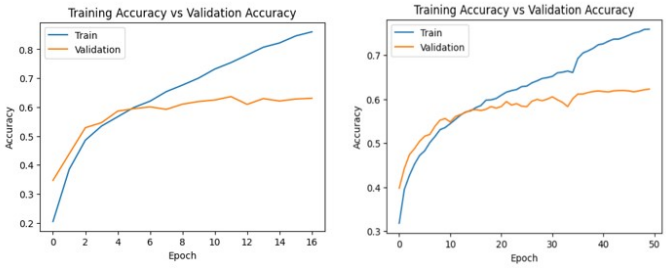

Fig. 6. Training vs Validation Accuracy per epoch for FER (left) and SER (right)

In the above graph we can we the CNN model for FER doesn't require all of the 50 epochs as it stops the training near 20 epochs due to early stopping. The validation accuracy of the CNN models used for FER and SER were 64.85% and 62.27% respectively. The maximum claimed accuracy for FER in the FER-2013 database is about 78% but considering various factors such as the number of emotions, resource constraints, and so on. Moreover, the biasness of the dataset is the main reason for lower accuracy in the dataset. But, keeping the factors in mind, our model performs comparatively well to detect emotions correctly to a good extent.

VII. Result comparison and Analysis

We tested various models to identify a suitable Facial Emotion Recognition (FER) model for our project. Among them, both the CNN and MTCNN-based FER libraries demonstrated similar performance, achieving an accuracy of approximately 65%. However, the DeepFace Library, despite claiming 97.5% accuracy on the LFW Dataset, yielded unsatisfactory results on the FER dataset. Consequently, we decided to opt for the CNN-based architecture. Unfortunately, the FER library does not provide any information about the specific architecture it implements. A comparison of the testes FER models are as follows -

Once more, we conducted experiments using different models to identify an appropriate Speech Emotion Recognition (SER) model for our project. Out of the models tested,

TABLE II
Accuracy of implemented FER models

FER Model	Accuracy
CNN	64.85%
FER Library	65.00%
DeepFace Library	21.47%

our implemented deep-CNN outperformed the MLP classifier and the CNN and LSTM model. Our deep-CNN architecture demonstrated a 62% accuracy in accurately detecting emotions from sounds. In contrast, the MLP classifier exhibited poor performance. Despite the expectation that CNN and LSTM would yield better results, they only achieved around 20% to 25% validation accuracy. Consequently, we selected the deep-CNN architecture as the preferred choice for our SER model. The comparison of the testes SER models are as follows -

TABLE III
Accuracy of implemented SER models

FER Model	Accuracy
Deep-CNN	62.27%
MLP Classifier	50.93%
CNN and LSTM	66.91%

Apart from the models that we have tried to implement, a study shows that the maximum accuracy that can be achieved in a FER 2013 dataset is about 78% [14], and taking various matters into consideration it can be said that the performance of our FER model is decent enough.

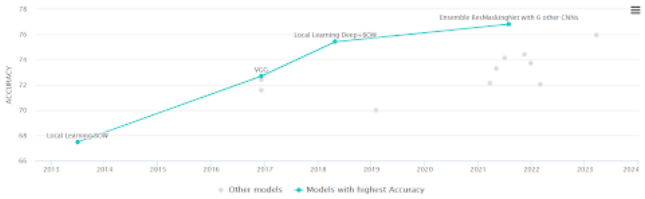

Fig. 7. Performances of Models on FER 2013 Dataset

Again, a study shows that the maximum accuracy of SER models in the RAVDESS dataset is about 78% [15] as well using SVM, and taking the limitations and other matters into consideration, the performance of our model was satisfactory enough.

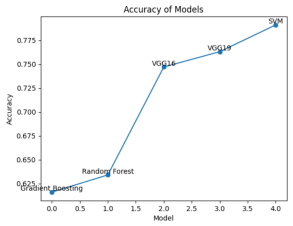

Fig. 8. Performances of Models on RAVDESS Dataset

Hence, our model's performance is deemed satisfactory when compared to existing models and methods of Emotion Detection.

VIII. Website

We developed a website to test out the project. Flask was chosen as our framework as it is very simple to develop and takes a small time to deploy. We created multiple routes for different pages so that the user can easily navigate through them and the frontend can communicate with the backend. The pages were built mostly using vanilla HTML and CSS and to make it dynamic there was some Java Script here and there as well. The trained models were saved as pickel files so the backend could use them as a tool without having to train it again.

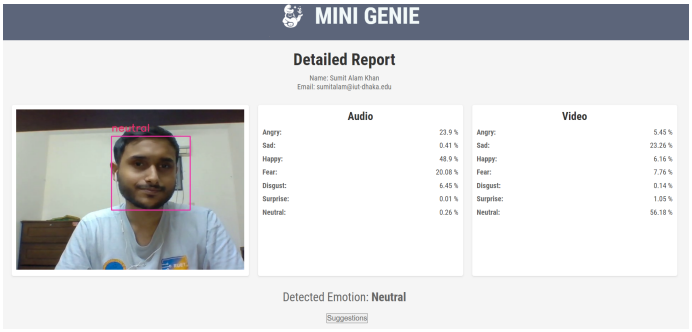

Fig. 9. Screenshot of the user interface of Mini Genie

IX. Code

The code to our project's GitHub Repository is here. Moreover, the colab notebook of our project is this and the FER and SER colab files are attached to the document as well.

X. Limitations and Challenges

During the course of this project, we encountered various challenges and limitations. Those are as follows -

A. *Noise and inadequate lighting*

When there is noise in the audio, or there isn't enough light in the background, the model loses accuracy

B. *Prediction speed*

Currently, the prediction for each frame takes a lot of time. That makes prediction for videos with a bigger duration almost impossible to work with

C. *Cohesion between audio and video prediction*

Human emotions are complex. His face and his voice might not denote the same emotion. This poses a problem for our model, as we have used just a simple mean function to combine audio and video prediction, which is not very accurate

XI. Future Work

There are limitless opportunities to improve emotion detection technology. Here, we present some of the existing problems of emotion detection from facial expression recognition and speech emotion recognition.

A. *Real-Time*

Our objective is to enhance the project by achieving real-time results without the need for batch processing. We aim to obtain output while capturing the video, utilizing faster calculations for more efficient processing.

B. *Noise Robustness*

Another noteworthy upgrade can be to make it work in spite of noises in the background. Even if there is inadequate lighting in the video, or background noises in the audio, the model should be able to work just fine.

C. *Algorithmic Development*

Not to mention, finding a better algorithm to combine the audio and video predictions would increase the effectiveness of this project by a considerable amount.

It is strongly believed that these efforts will significantly improve the performance of our work.

XII. Conclusion

This research study centered for detecting emotions from video data using CNN. Through extensive experimentation and analysis, we have effectively demonstrated the efficacy and capacity of CNN models to effecitvely recognize and categorize emotions depicted in input videos.

Our findings indicate that CNN architectures, purposefully designed for tasks involving visual and audio recognition, possess the capability to adeptly capture and learn intricate features present in facial expressions and speech-based emotions. This leads to enhanced performance in emotion detection. The integration of convolutional layers, pooling layers, and fully connected layers within CNN models facilitates hierarchical feature extraction and robust representation learning from video data.

Through training and fine-tuning our CNN models on diverse and extensive datasets annotated with emotions, we achieved notable accuracy, achieving nearly 65% accuracy in detecting seven types of emotions within a short timeframe. Additionally, we optimized the computational efficiency of our CNN model, enabling real-time emotion detection through a web application we developed, utilizing the webcams of personal computers.

Nevertheless, we acknowledge that there are still challenges to be addressed in the field of emotion detection from video using CNN. Variations in lighting conditions, occlusions, and the presence of multiple individuals in the video frames can impact the accuracy of emotion recognition systems. Further research is needed to develop more robust models that can handle these real-world complexities.

To summarize, our research makes a valuable contribution to the advancement of emotion detection from video through the application of CNN models. Our work underscores their effectiveness, accuracy, and real-time capabilities. We aspire that our findings will inspire future research and applications in this field, ultimately fostering improved understanding of emotions, improving the interaction between humans and

computers, and fostering the advancement of emotion-sensitive systems in various industries.

XIII. Acknowledgement

We are extremely grateful to all those who had their valuable involvement in the successful completion of our research work. Our heartfelt appreciation goes to our teachers, whose guidance, expertise, and unwavering support have been indispensable throughout this journey. Their profound knowledge and mentorship have shaped our understanding of the subject matter and paved the way for our research endeavors.

We also extend our thanks to our friends and classmates, whose valuable insights, constructive feedback, and encouragement have been instrumental at every stage of this project. Their enthusiasm and willingness to engage in meaningful discussions have played a vital role in refining our ideas and strengthening the quality of our arguments.

It is important to acknowledge that without the contributions and assistance of all these individuals, this project would not have been possible. We are sincerely grateful for their involvement and take immense pride in recognizing the significant impact they have had on our academic and personal growth.

References

[1] O. Ezerceli and M. T. Eskil, "Convolutional neural network (cnn) algorithm based facial emotion recognition (fer) system for fer-2013 dataset," in *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 2022, pp. 1–6.

[2] Y. Tang, "Deep learning using support vector machines," *CoRR, abs/1306.0239*, vol. 2, no. 1, 2013.

[3] Justin Shenk, "Fer: Facial emotion recognition library," PyPI, 2023. [Online]. Available: https://pypi.org/project/fer/

[4] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[5] A. Alsobhani, H. M. ALabboodi, and H. Mahdi, "Speech recognition using convolution deep neural networks," in *Journal of Physics: Conference Series*, vol. 1973, no. 1. IOP Publishing, 2021, p. 012166.

[6] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.

[7] N. N. Poojary, G. Shivakumar, and B. Akshath Kumar, "Speech emotion recognition using mlp classifier," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, vol. 7, pp. 218–222, 2021.

[8] G. et al., "Facial emotion recognition (fer)," Kaggle, 2013, available: https://www.kaggle.com/datasets/msambare/fer2013.

[9] R. F. Livingstone SR, "Ryerson audio-visual database of emotional speech and song (ravdess)," Kaggle, 2018, available: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio.

[10] M. K. P.-F. Kate Dupuis, "Toronto emotional speech set (tess)," Kaggle, 2010, available: https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess.

[11] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crowd sourced emotional multimodal actors dataset (crema-d)," Kaggle, 2014, available: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio.

[12] ——, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[13] "Surrey audio-visual expressed emotion (savee)," Kaggle, available: https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee.

[14] E. Barsoum, C. Zhang, C. Canton Ferrer, and Z. Zhang, "Deep learning for facial expression recognition: A comprehensive study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2494–2509, 2017. [Online]. Available: https://paperswithcode.com/paper/deep-learning-for-facial-expression-1

[15] S. N. Zisad, M. S. Hossain, and K. Andersson, "Speech emotion recognition in neurological disorders using convolutional neural network," in *Brain Informatics: 13th International Conference, BI 2020, Padua, Italy, September 19, 2020, Proceedings 13*. Springer, 2020, pp. 287–296.