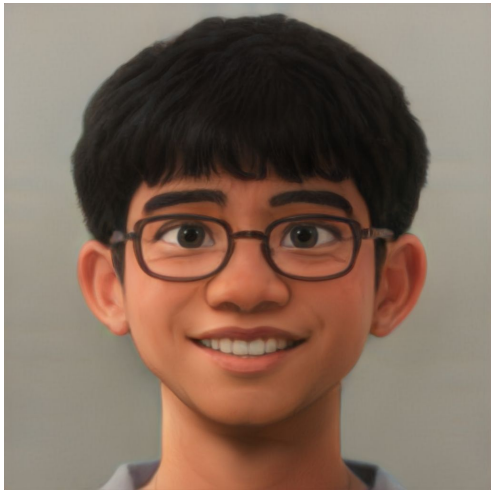
MiniGenie

Emotion Detection from Video using CNN Architecture

PROJECT PRESENTATION

CSE 4622 : Machine Learning Lab

TEAM MEMBERS



Sherajul Arifin
190041201

Aashnan Rahman
190041204

Sumit Alam Khan
190041207

Introduction

Idea behind the project

MiniGenie is a tool that detects human emotions from a **video** by combining **Facial Emotion Recognition (FER)** and **Speech Emotion Recognition (SER)** technologies.

Facial Emotion Recognition (FER)

Identifies and analyzes the emotions expressed by a person's face using computer vision and machine learning techniques.

Speech Emotion Recognition (SER)

Recognizes and analyzes human emotions expressed through speech signals by speech processing, machine learning, and signal analysis.

Both of the technologies are part of sentiment analysis.

Datasets

Image Dataset

FER 2013 (Facial Expression Recognition)

The dataset contains grayscale images of 48x48 pixels. These face mages are classified into seven emotion categories: angry, sad, disgust, fear, happy, neutral, and surprised. The dataset has **28,709** training examples and **3,589** test examples. Images were collected by searching for emotions and their synonyms on Google.

Audio Dataset

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

The RAVDESS dataset is a huge dataset of **7,356** files containing **16-bit** audio data. For our project, we have a used a dataset of **1440** files created by 24 actors. The files are categorized into **seven** expressions.

Steven R. Livingstone, & Frank A. Russo. (2019). ***RAVDESS Emotional speech audio*** [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/256618

Audio Dataset

Toronto emotional speech set (TESS)

TESS dataset contains audio clips of two women where they have expressed **seven** different emotions. In the dataset, there are **200** target words. If the whole dataset is considered, there are in total **2800** audio files or data points in the TESS dataset.

Pichora-Fuller, M. Kathleen; Dupuis, Kate, 2020, "***Toronto emotional speech set (TESS)***", https://doi.org/10.5683/SP2/E8H2MF, Borealis, V1

Audio Dataset

Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D)

CREMA-D is an audio-visual dataset used for emotion recognition and it has a total of **7,442** original clips. The actors are versatile races and ethnicities. **Six** different emotions were expressed in those files. Since the dataset was large, it was crowd-sourced by 2443 participants.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). ***CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. IEEE transactions on affective computing,*** 5(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

Audio Dataset

Surrey Audio-Visual Expressed Emotion (SAVEE)

Four native English speakers recorded the SAVEE dataset. This dataset has **7** cardinal expressions of emotions. Throughout the dataset, phonetical balance was maintained. The credit for accumulating the whole dataset goes to the University of Surrey.

Data Preprocessing

For FER,

1. Data Scaling and Resizing (48x48 pixels)
2. Encoding labels

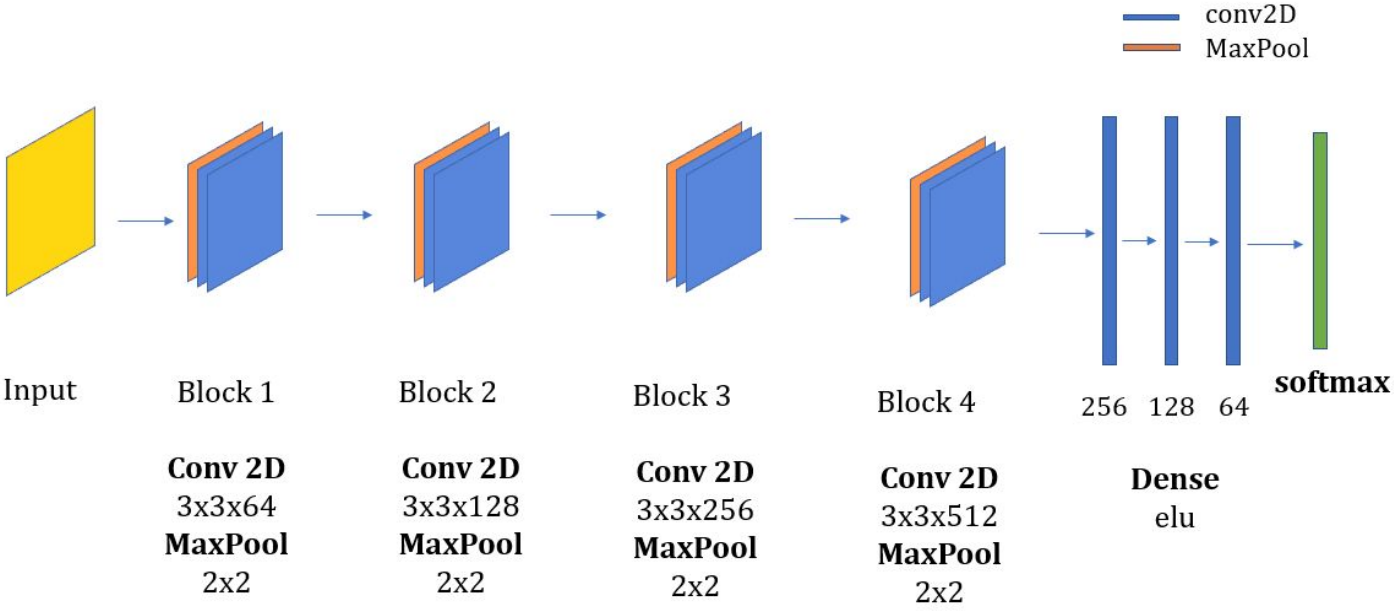
For SER,

1. Data Augmentation (adding small perturbations - noise, stretch, shift, pitch)
2. Feature Extraction (ZCR, MFCC, RMS, Melspectogram)
3. One hot encoding
4. Normalizing

Solution Approach

Hyperparameter of the FER model

Model Architecture (CNN)



Hyperparameter of the FER model

Common Attributes

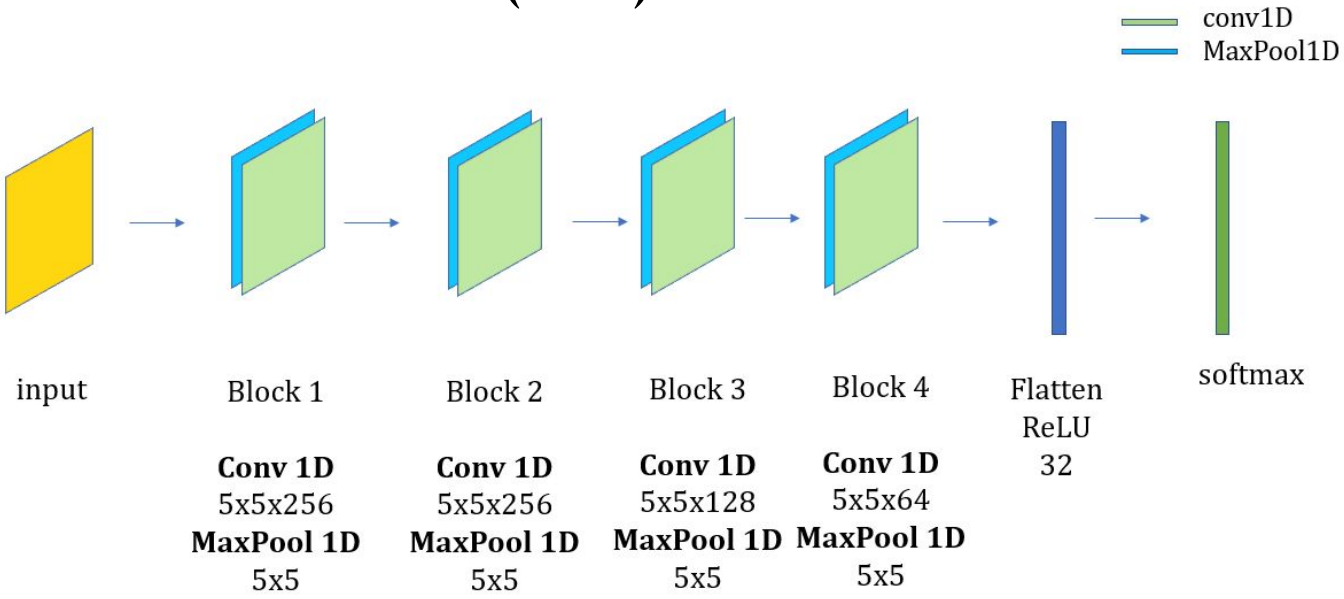
Epoch = 50
Batch size = 64

Loss function = categorical crossentropy
Optimizer = rmsprop
Metric = Accuracy

Callbacks = Checkpoint and Early Stopping

Hyperparameter of the SER model

Model Architecture (CNN)



Hyperparameter of the SER model

Common Attributes

Epoch = 50
Batch size = 64

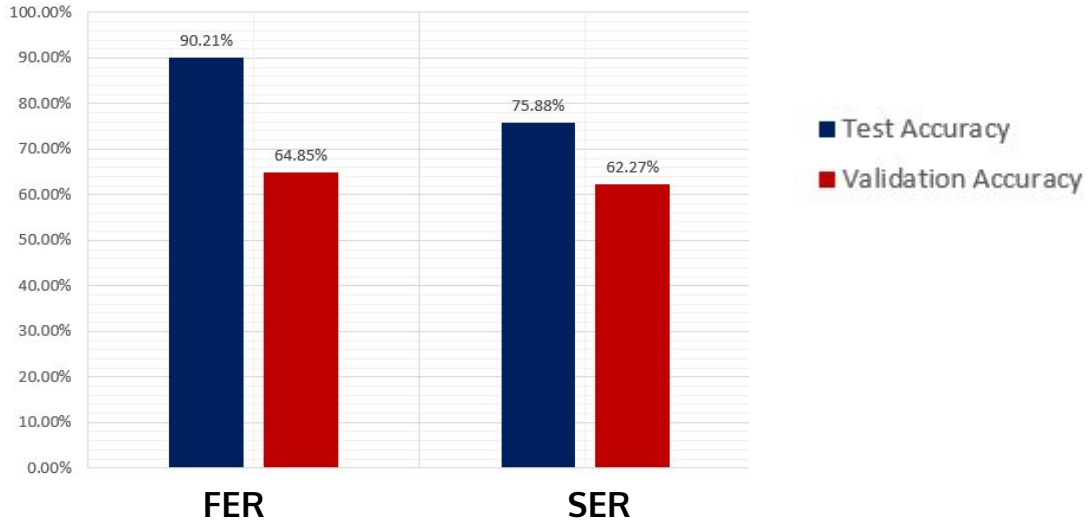
Loss function = categorical crossentropy
Optimizer = adam
Metric = Accuracy

Callbacks = ReduceLROnPlateau

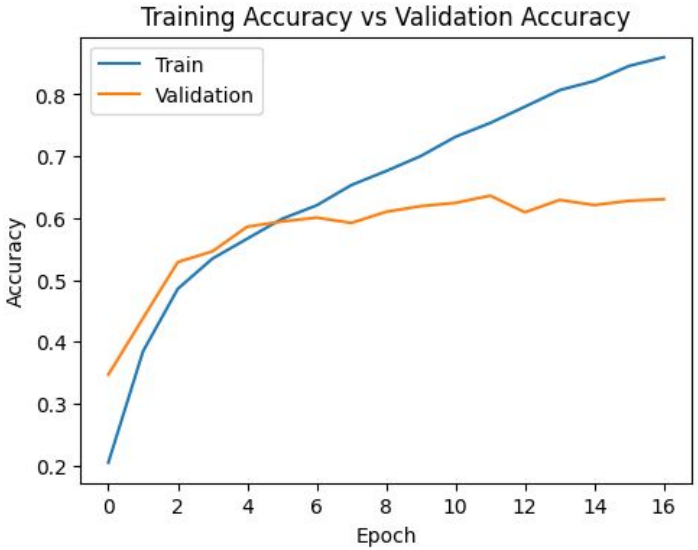
Accuracy Analysis

Model Accuracy

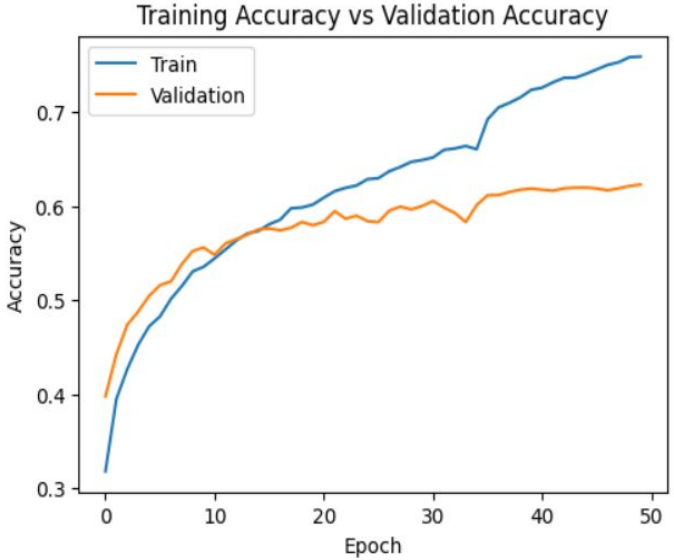
FER	SER
Train Accuracy = 90.21 % Validation Accuracy = 64.85 %	Train Accuracy = 75.88 % Validation Accuracy = 62.27 %



Model Accuracy



FER



SER

Model Comparison

FER models

1. **FER library**
   This library is based on MTCNN (Multi-Task Cascaded Convolutional Networks). Which is a popular facial feature extraction algorithm specifically designed for real-time face detection and alignment in images.

2. **DeepFace library**
   DeepFace is a lightweight face recognition and facial attribute analysis (age, gender, emotion and race) framework for python. It is a hybrid face recognition framework wrapping state-of-the-art models: VGG-Face, Google FaceNet, OpenFace, Facebook DeepFace, DeepID, ArcFace, Dlib and SFace.

SER models

1. **MLP Classifier**
   MLP classifier refers to a Multilayer Perceptron classifier, which is a type of artificial neural network used for classification tasks. MLP classifier is effectively used for detecting speech emotion.

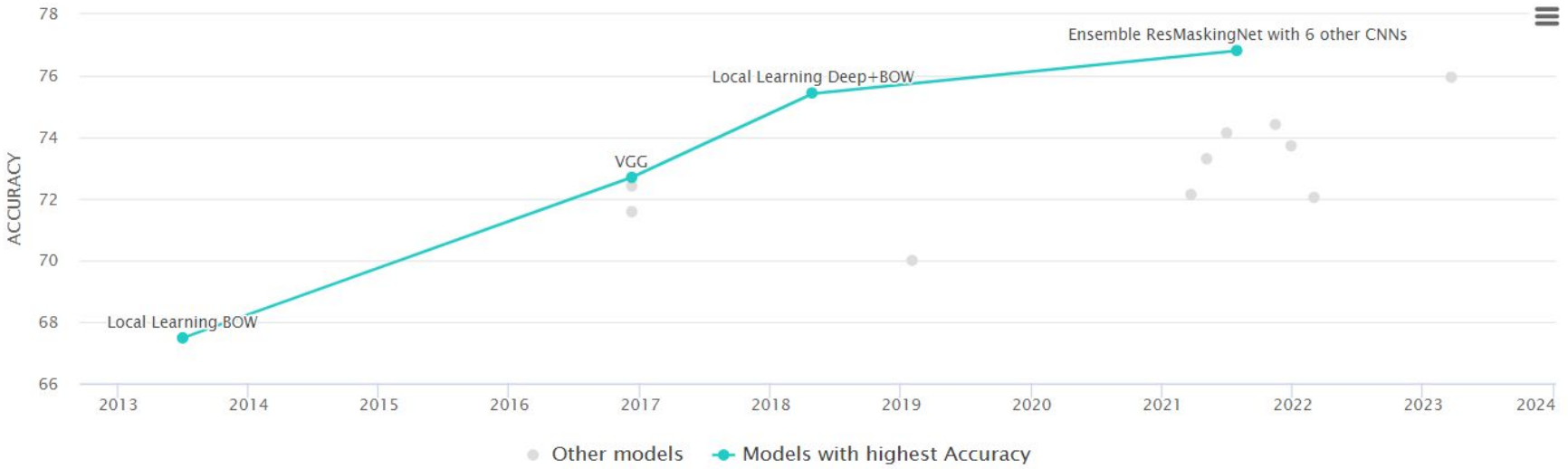
2. **CNN + LSTM**
   By combining LSTM (Long Short-Term Memory) and CNN (Convolutional Neural Network).
   LSTM is a type of recurrent neural network (RNN) architecture that is designed to overcome the limitations of traditional RNNs in capturing long-term dependencies in sequential data.

Comparisons

FER Models	Accuracy
CNN	64.85 %
FER Library	65 ± 5 %
DeepFace Library	97.44% (only on (LFW) dataset)

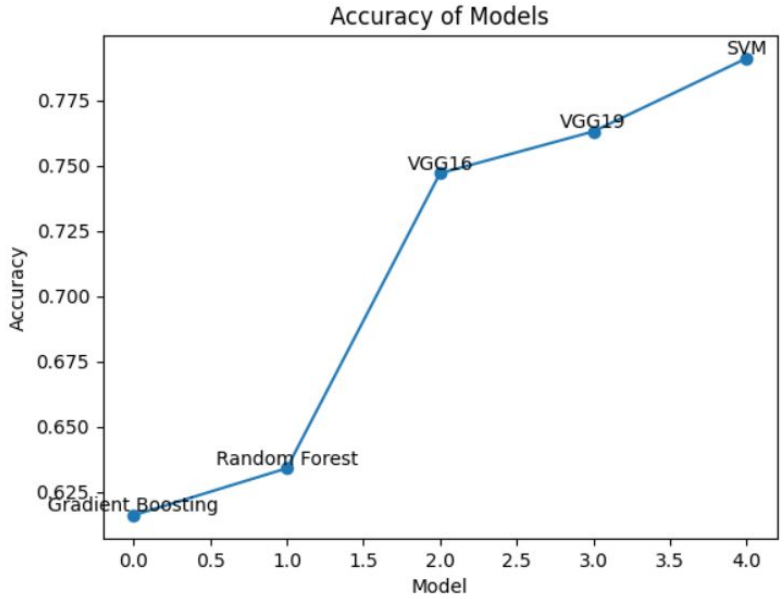
SER Models	Accuracy
CNN	62.27 %
MLP Classifier	50.93 %
CNN + LSTM*	66.91 % (claimed)

FER Comparisons



Comparison of different models on FER-2013 dataset can be found here

SER Comparisons

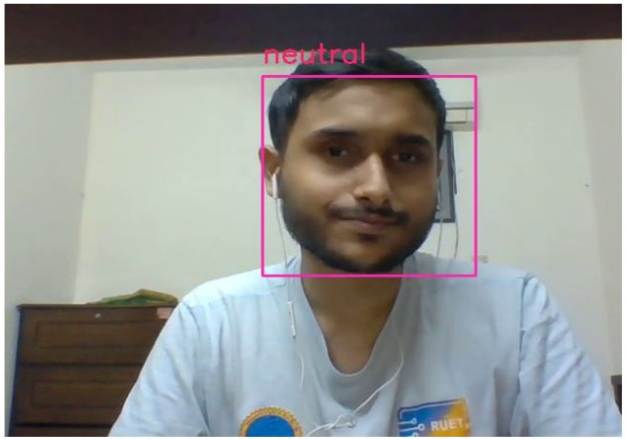


Demonstration

MINI GENIE

Detailed Report

Name: Sumit Alam Khan
Email: sumitalam@iut-dhaka.edu



Audio

Angry:	23.9 %
Sad:	0.41 %
Happy:	48.9 %
Fear:	20.08 %
Disgust:	6.45 %
Surprise:	0.01 %
Neutral:	0.26 %

Video

Angry:	5.45 %
Sad:	23.26 %
Happy:	6.16 %
Fear:	7.76 %
Disgust:	0.14 %
Surprise:	1.05 %
Neutral:	56.18 %

Detected Emotion: **Neutral**

Suggestions

Codes

Code

Github Repository – Project Link
Project Colab File – Colab Notebook for project
Main FER Colab File – FER Notebook
Main SER Colab File – SER Notebook

Limitations

Limitations

1. Fails to recognise facial expressions when the environment is blurry or dark.
2. Limited understanding capability for complex human emotions from facial or voice expression.
3. Lack of diverse and large-scale datasets.
4. Result highly dependent on classification of dataset
5. Heavy computation power is required to perform the prediction.
6. Features extracted from speech are highly dependent on various factors i.e. gender, ethnicity, tone, pitch etc.

Future Work

Future Works

1. Make the prediction realtime
2. Faster prediction for each frame
3. Successful prediction even when the frame is not clear
4. Better coordination between the voice and audio prediction
5. Successful prediction even with noisy audio.

THANK YOU!

FEEL FREE TO ASK US ANY QUESTION