Emotion detection via Voice and Speech Recognition

Dr. Rohit Rastogi

(Associate Professor, Dept. of CSE, ABES Engineering College Ghaziabad, U.P., India) rohitrastogi.shantikunj@gmail.com, 8076772048; 9818992772

Tushar Anand

(Student, B. Tech. Second Year, Dept. of CSE, ABES Engineering College Ghaziabad, U.P., India) anand12tushar@gmail.com, +91-7017915257

Shubham Sharma

(Student, B. Tech. Second Year, Dept. of CSE, ABES Engineering College Ghaziabad, U.P., India) sharma.shubh1209@gmail.com, +91-8766327963

Sarthak Panwar

(Student, B. Tech. Second Year, Dept. of CSE, ABES Engineering College Ghaziabad, U.P., India) Exclusivethought@gmail.com, +91-9560003472

ABSTRACT

Emotion detection from voice signals is needed for human-computer interaction (HCI), which is a difficult challenge. In the literature on speech emotion recognition various well know speech analysis and classification methods have been used to extract emotions from signals. Deep learning strategies have recently been proposed as a workable alternative to conventional methods and discuss several recent studies have employed these methods to identify speech-based emotions. The review examines the databases used, the emotions collected, and the contributions to speech emotion recognition.

The Speech Emotion Recognition Project was created by the research team. Which recognizes Human speech emotions. The research team developed our project using Python 3.6. RAVDEESS dataset was also used since it contained 8 distinct emotions expressed by all speakers. The RAVDESS dataset, Python programming languages, and Pycharm as an IDE were all used by the author team.

Keywords**--** Emotion Recognition, Speech Features, MFCC (**Mel-frequency cepstral coefficients)**, Deep Learning, RAVDESS Dataset, Librosa, sklearn, numpy, MLP classifier

PROBLEM STATEMENT

The Manuscript deals with the exploration and normalization of the data. As a performance measure for conversational analysis, SER (speech Emotion Recognition) may be used to categorize calls based on emotions and assess customer happiness, which enables businesses to enhance their services**.** The challenge is to design automated software for this purpose.

MOTIVATION OF STUDY

In this era of technology, one of the major concerns is the self's emotions. To overcome this problem, this research work analyses a person's speech and determines emotion. The motivation for this paper is to face the problem which is emotion itself.

Human beings as a species of higher intelligence show various emotions. Due to this, it is necessaryto understand the emotions which are conveyed in speech. The basic human emotions can be categorized as happiness, sadness, fear, disgust, anger, and surprise. Furthermore, these are further classified into complex emotions such as awe, guilt, envy, etc. It is therefore in our interest to understand these emotions.

OBJECTIVES OF RESEARCH

The area of voice recognition that is expanding in acceptance and reputation is emotion recognition. This assignment attempts to apply deep learning to detect the sentiments from the data, even though there exist methods for understanding sentiment using a machine learning approach.

In this research project, the research team has built a model that may recognize emotions from sound files using an unsupervised learning algorithm known as an MLP-Classifier. The objective of speech emotion recognition is to detect the presence of frustration or annoyance in the speaker's voice by using the librosa libraries in python and the RAVDESS dataset.

SCOPE OF STUDY

This project works on how one can use audio files to detect emotions. Various audio files are processed, searched, and then resulted in different sets of emotions like sad, happy, and nervous. One of the Purposes of this study is to make human-system interaction more effective.

As of now, the system is working on the detection of audio files that is capable of detecting single audio files but not grouped audio files. More accurate implementation of the detection of voice can be done by clearing the audio files that are mixed with the background disturbances and also the pauses in between the audio files that lower the accuracy of our results. The features of different types of voices of different domains can also be loaded to make the best effective output of emotions through voice.

STAE of ART and TOPIC ORGANIZATION

This study, first of all, gives a general idea of speech emotion recognition and how it is applicable in smart cities. The author's team explained how the speech recognition system works and various emotions which can be detected. For the endorsement of this study, the author team reviewed eight research papers of concerned topics etc.

The author team has described the methodology in which they have presented the different methods used for the study. This study used ML based critical analysis of the data obtained from the RAVDESS database. The manuscript has tabular, graphical representation of data using graphs and spectrograms.

Within the recommendation area, which is one of the foremost critical parts of the research studies, recommendations for particular applications to address the issues and limitations distinguished within the appraisal have been displayed. The novelty area alludes to components that are new within the manuscript. In the last, the conclusion section represents the final assessment and describes the overall findings of the study.

ETHICAL COMMITTEE AND FUNDING

The experiments don't include any human-related experiments and so no ethical constraints have been violated. Though the subjects performing the study were humans and air quality directly affects them but the study doesn’t violate any health-related measures. The Project is not funded by any agency.

ROLE OF AUTHORS

Dr. Rohit Rastogi acted as team leader and coordinated among all co-authors. He prepared the topic introduction and background study. He also prepared the structure of the manuscript and ensured the quality of the research. Mr. Shubham did the analysis part. Mr. Sarthak and Mr. Tushar performed the backend task of implementation that consists of downloading the dataset and assembling them to the right path location. Mr. Shubham handled the task of splitting the dataset into training and testing parts. Mr. Sarthak performed the initialization of the MLP Classifier and trained the model. Mr. Shubham completed the graphs-related work, and tested the accuracy of the model to come to a final conclusion.

INTRODUCTION

Speech is the primary tool used by humans to interact and convey information. But the interesting fact is to know about the type of information that is really delivered i.e. detecting emotion is among the most crucial marketing tactics in the world today. For this reason, the research team decided to work on a project in which researchers can control numerous AL-related applications by being able to manage someone’s mood merely by their voice.

In this article, the research team has tried to combine prosody non-verbal aspects of language that allow people to convey or understand emotion and deep learning in order to create a model that understands human emotions through speech.

Examples include the ability of call centers to play music during tense exchanges. Another example may be a smart automobile that slows down when the driver is scared or furious. Because of this, this kind of application has a lot of promise in the smart world and might potentially increase consumer safety while benefiting businesses.

Speech Recognition and World Wide Applications

Today, the most common application of Speech recognition is in mobile devices. Speech recognition has become a crucial element of many smartphones currently on the market, from voice calling to asking Siri what the weather will be like on Monday.

Speech recognition technology is also used for voice calling, speech-to-text conversion, call routing, and voice search. Users can also utilize speech recognition in computer word processing programs like Google Docs or Microsoft Word, where they can alter and say what they wish to appear as text.

It's similar to being able to listen to someone and recognize words and phrases, then translating them into sentences that assist you to grasp what they're saying (Xiaobo, B. et al., 2018) (as per figure 1).

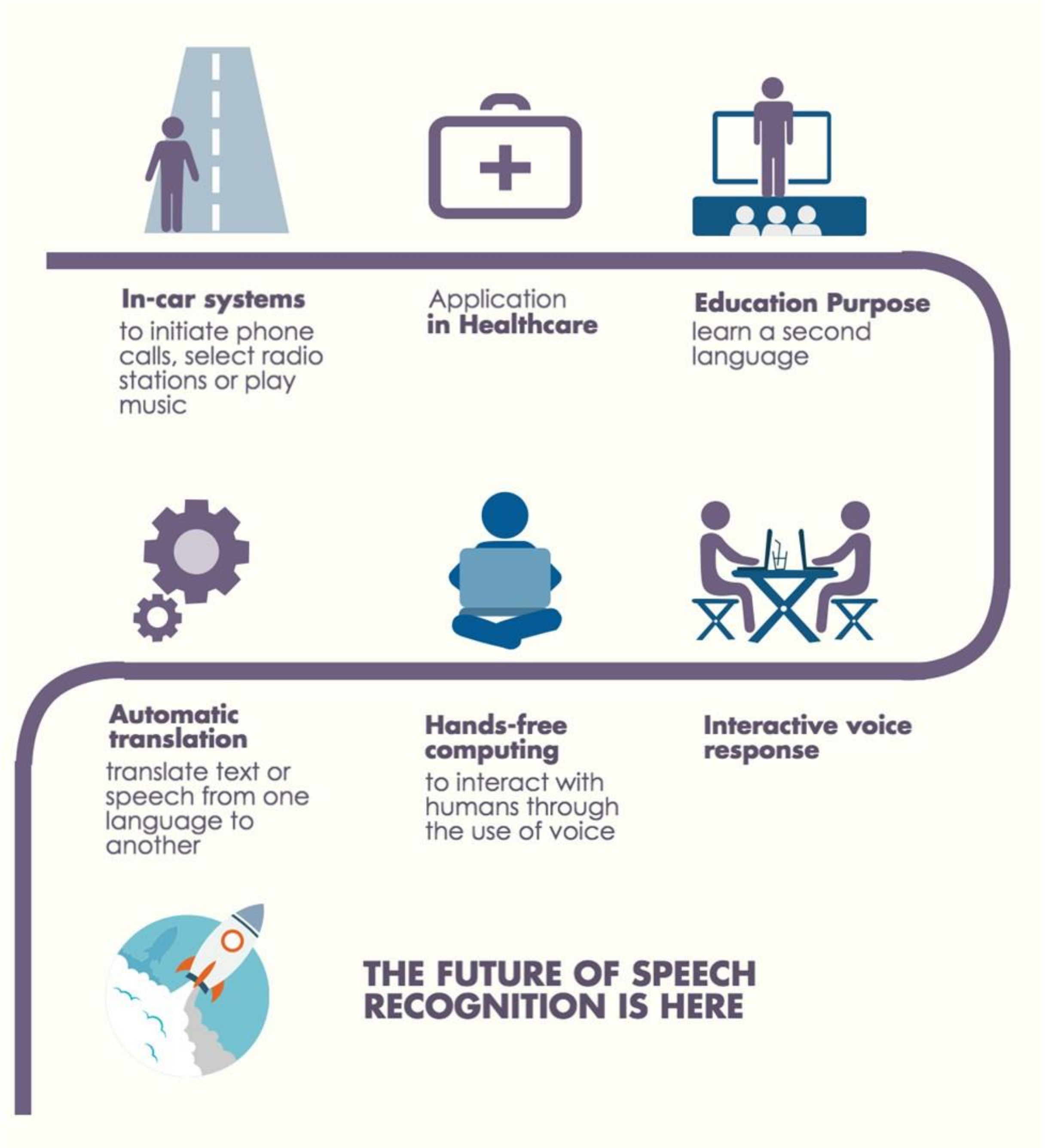

In-car systems
to initiate phone calls, select radio stations or play music
Application in Healthcare
Education Purpose
learn a second language
Automatic translation
translate text or speech from one language to another
Hands-free computing
to interact with humans through the use of voice
Interactive voice response
THE FUTURE OF SPEECH RECOGNITION IS HERE

Figure 1. The Future of Speech Recognition

Source: *(1*tvseNZieCHWaSntjG-ER7g.png (1151×1275) (medium.com))*

Emotion Recognition and 21st CenturyLife-Style

Healthcare, marketing, fraud detection, and manufacturing are just a few of the industries that might benefit from the application of emotion recognition technologies. In urgent care centers where individuals don't make appointments, healthcare practitioners can utilize emotion detection AI to prioritize individual patients by monitoring facial expressions in the reception room. The most uncomfortable people could be given top priority, while those who are less ill would have to wait fora gap inservice.

Before introducing a product, managers like to discover how it will operate. By evaluating the facial movements of testers using their goods or viewing their commercials, emotion recognition technology may help businesses get something out of focus groups. Emotion recognition technology is very useful for the automotive industry. Auto mobiles that warn drivers when they fall asleep or start to drift off might help avoid serious accidents. The warning might potentially be sent off by intense emotions like road rage. In the case of vehicles with self-driving or autopilot functions, this may be extremely useful. The autopilot can be used while informing the driver if the human operator becomes extremely emotional orfatigued.

When a consumer submits a claim, insurance firms employ voice analysis to determine if they are being truthful or not. Up to 30% of consumers have acknowledged lying to their auto insurance provider to obtain coverage, according to independent polls (Morre, S. et al., 2018) (as per Figure 2).

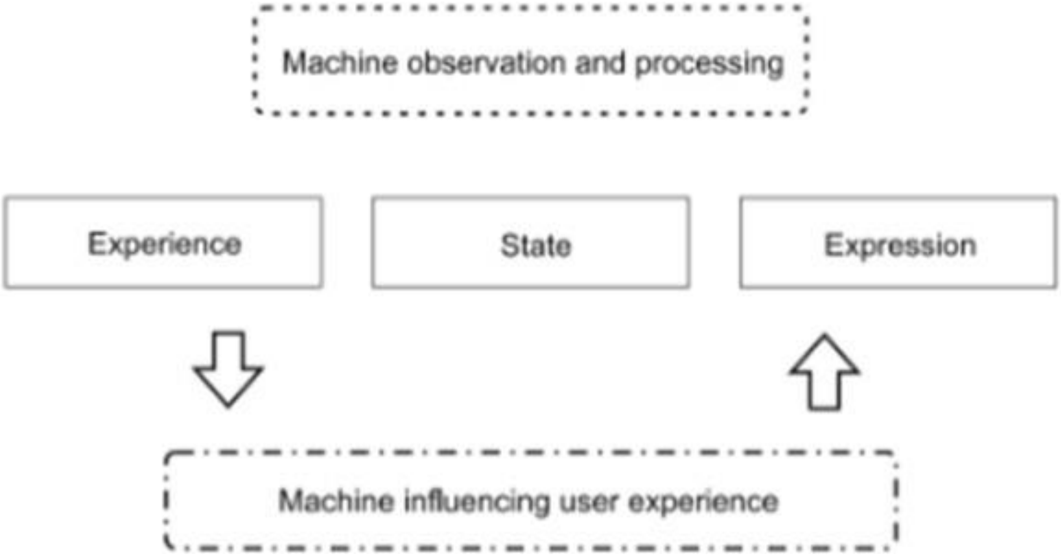

Machine observation and processing
Experience
State
Expression
Machine influencing user experience

Figure 2. The framework of affective computing.

Source: *(<u>10.1109/MIPRO.2016.7522336</u>)*

Speech Recognition with Emotion Recognition and Accuracy Standards

Emotion recognition from speech has been gaining popularity in recent years. It has huge potential and advantages in various sectors such as teaching, banking, and many more. Speech Recognition with Emotion Recognition has two major parts: Feature extraction from speech and Emotion machine classifier.

“The problem of Speech Emotion recognition is solved by classifying an order, in which input is a sequence whose length varies and a single output is obtained.” (L. Kerkeni et al,2018).

In recent years, researchers have proposed various classification algorithms in Speech Emotion Recognition (SER) such as Hidden Markov Model (HMM), Support Vector Machine (SVM), and Neural Networks (NN). Researchers are going to use Multilayer Perceptron (MLP) for Speech Emotion Recognition. The various classification algorithms have different accuracy standards for different emotions and different datasets (as per figure 3).

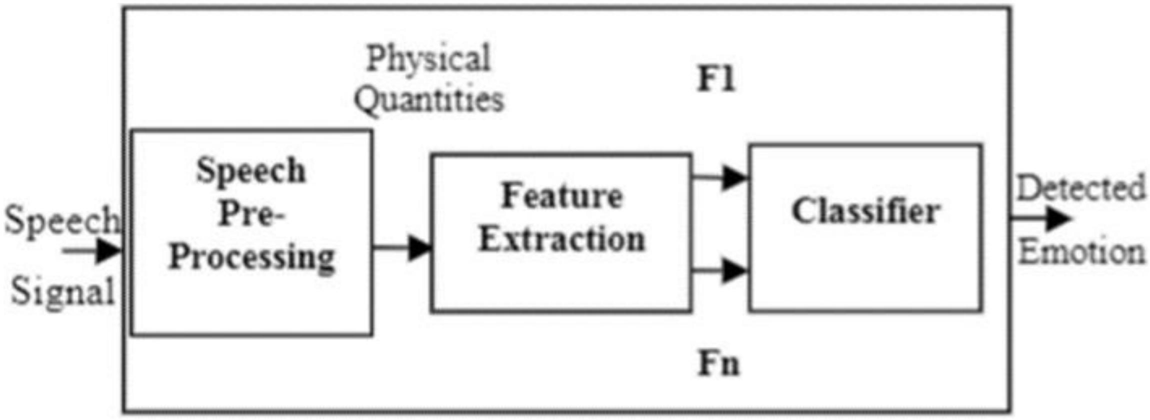


Figure 3. Process of Detecting Emotions from Image

Source:*(https://ieeexplore.ieee.org/document/8433554/figur)*

1.1 Smart City Designs and Emotion Capture Scenario

In this age, cities are becoming more advanced and smarter. Smart cities include smart infrastructure, smart healthcare, smart technology and smart energy. For a city to achieve a status of a smart city, IOT is necessary to connect various things. Speech Emotion recognition can play a vital role in making the system fast and more efficient by recognizing emotion.

Speech Emotion Recognition has various applications in a smart city. Using Speech Emotion Recognition, a customer’s emotion can be determined after a service, which could be helpful in getting a review. SER can be helpful in a distress scenario. If a person is frightened it can be recognized and action can take place in times of emergency like burglary or an accident.

"SER serves an important role in developing smart services which are applicable in surveillance, healthcare, audio forensics, affective computing, and human-machine interaction" (Badshah, A.M., et al., 2019) (as per figure 4).

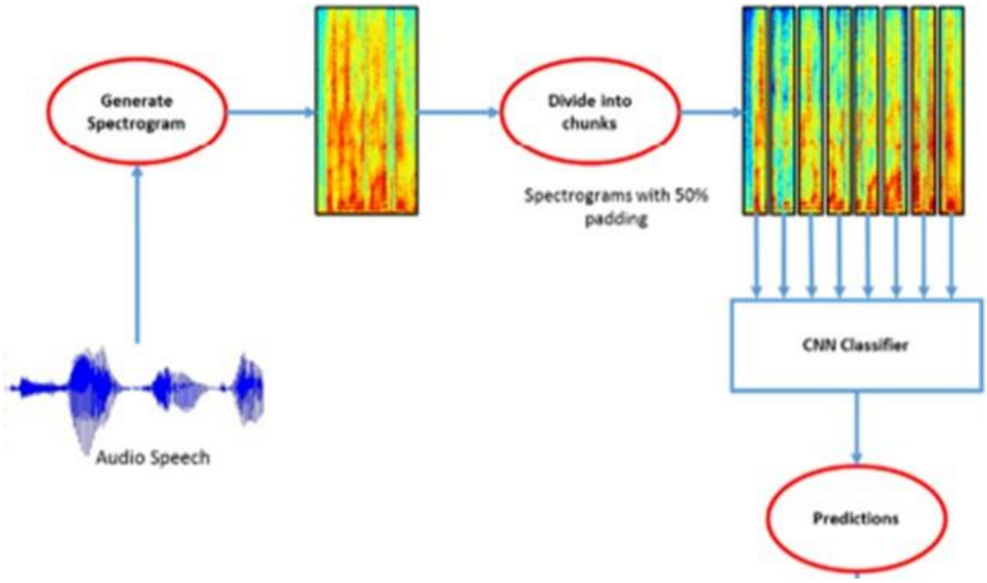


Figure 4. Emotion Prediction from speech

Source: ***(10.1007/s11042-017-5292-7)***

Knowledge Pyramid and Knowledge Extractions by Speech Recognition Based Systems

For any system knowledge management is needed to work efficiently and it is required to ensure that right things are available at the right place. The knowledge pyramid is a representation of relation between four different factors- Data, Information, Knowledge, and Wisdom.

In this system data is a set of different types of voices. Information consists of analyzing the data for further processing by the system. Knowledge consists of how one can apply one's collected information to achieve desired goals and the last is wisdom, that is what one is capable of doing now and what can be achieved in future (Demicran, S. et al., 2014) (as per figure 5).

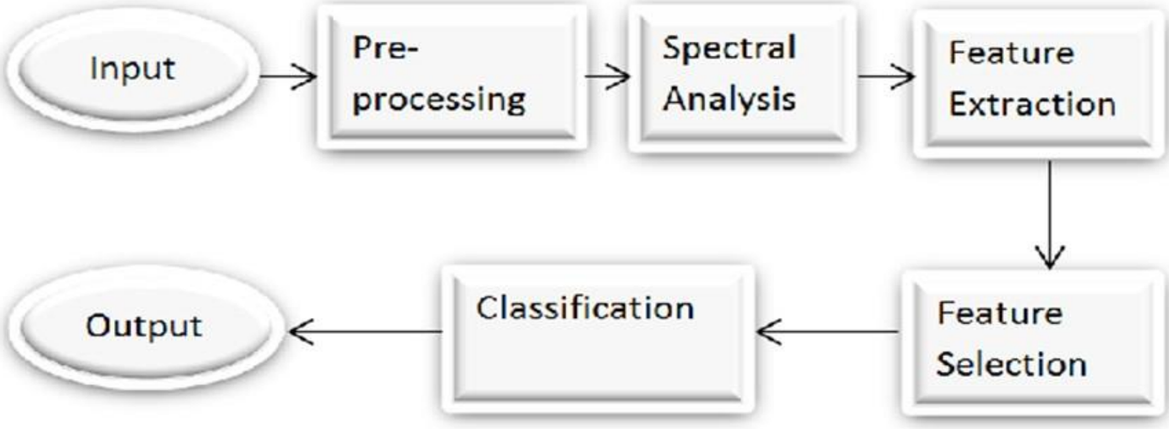


Figure 5. Extracting Knowledge from Data.

Source:*(https://d3i71xaburhd42.cloudfront.net/0cf8730a3100e7f93f288da2a2c5d7d007048683/1-Figure1-1.png)*

1.2 Impact of AI, ML and Big Data Based Systems in Global DevelopmentIndex

In this fast growing world there is a requirement of that model which helps one to walk in parallel with the world. AI and ML helps to enhance the creativity level and can help to manage large data in the properway.

The development of mobile phone software and various devices has made a huge change in the medical field. AI, ML and Big data management are also contributing in giving facilities in wireless-based applications that help to reach remote areas. As the population is rising, countries are also funding and investing in these types of modern technologies. ML is that part of AI that is capable of learning automatically and also improves its functionality with increase in time. With help of these modern techniques can be developed performance measuring devices or indicators that make one's work easy.

AI can take up tasks that are involved with high risk and work in places where humans cannot stay for a long time. It is also estimated that AI and ML may contribute to an additional increase in global GDP by 1.2% annually (Z. Khan, et al., 2020) (as per figure 6).

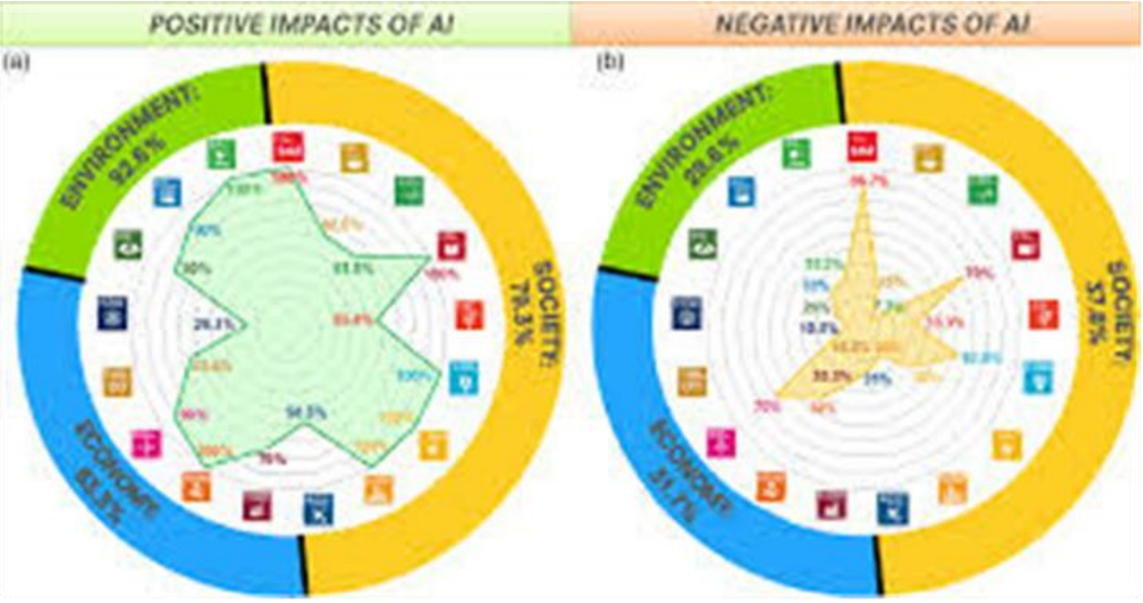


Figure 6. Impact of AI on Global Development

Source:*(https://media.springernature.com/lw685/springer-static/image/art%3A10.1038%2Fs41467-019-14108-y/MediaObjects/41467_2019_1Fig1_HTML.png?as=web*

2 LITERATUREREVIEW

To improve the understanding and importance of Speech Emotion Recognition, our team reviewed various papers. The key to understanding SER, in another study, the speech signal was used to extract the standard emotional speech characteristics such as Perceptual Linear Prediction cepstral

Coefficients (PLP) and LPCC. In this paper, the researcher team builds an SER using deep learning on the RAVDESS dataset and achieves the result with high accuracy.

C.Huang, (2015) and his team emphasized a method that automatically extracted the emotional characteristic parameter from emotional voice signals using deep belief networks (DBNs), one of the deep learning models. This approach can precisely extract emotional characteristic factors, which obviously increases the accuracy of emotional speech recognition. However, training the DBN feature extraction model took 136 hours, which was more time than other feature extraction techniques.

For training and recognition purposes, this text selected 1200 phrases containing the four fundamental emotions of grief, rage, surprise, and happiness. For training and testing purposes, this article uses 40% and 60% of the voice data, respectively. The study put out a technique for realizing the emotional elements that were automatically retrieved from the text. To extract voice emotion characteristics, a 5-layer deep network was trained usingDBNs.

They will carry out more research on voice emotion identification using DBNs in the future and enlarge the training data set. Their ultimate goal is to do research on how to increase the accuracy of voice emotion identification (C. Huang et al., 2015).

R.A Khalil et al. (2019) demonstrated how deep learning algorithms like DBM, RNN, DBN, CNN, and AE have received a lot of attention in recent years. These deep learning techniques and their layer-wise architectures are demonstrated through the categorization of a variety of natural emotions, including enjoyment, happiness, sorrow, calm, shock, bored, hate, terror, and anger.

It identifies some constructive directions for enhanced SER systems. Investigated are SER methods based on CNNs and RNNs. With LSTM network layers, the deep hierarchical CNN's structure for extracting features has been merged. According to research, CNNs feature a time-based dispersed network that produces more accurate findings. Similar to this, a system called PCA-DCNNs-SER based on a deep convolution network (D-CNN) that employed audio data as input is provided.

These tests produced more steady, accurate, and robust results for recognition in challenging situations with changing language and speaker as well as other environmental aberrations. It is simple to identify emotions like pleasure, happiness, sadness, surprise, neutrality, boredom, disgust, fear, and rage. But when real-time emotion identification is sought, it becomes challenging to accomplish so. (R.A. Khalil et al., 2019).

L. Kerkeni, et al., (2018) demonstrated how using different features and databases in Speech Emotion Recognition (SER) can produce different results and accuracy.

They used Mel-Frequency Cepstrum Coefficient (MFCC) and Modulation Spectral Features (MSF) for feature extraction from speech. The Berlin database and Spanish Databases were used.The classification algorithms that were used are Multivariate Linear Regression (MLR), Support Vector Machine (SVM), and Recurrent Neural Networks (RNN).

For training the model 70% data was used and for testing 30% data was used. From the Berlin database using MLR classifier, the average emotion detection rate for MS, MFCC, MFCC +MS features was 60.70%,67. 10% and 75.90%. While for the same features but for Spanish database results were as of 70.60%,76.08%,82.41%.From the Berlin database using SVM classifier, theaverage detection rate for MS, MFCC,MFCC+MS features was 63.30%, 56.60%, 59.50%. While for Spanish database with the same features, the results were 77.63%,70.69%.From the Berlin database using RNN classifier, the average emotion detection rate for MS,MFCC,MFCC+MS features was 66.32%,69.55%,58.51%. While for the same features but for Spanish database results were as of 82.30%, 86.56%, and 90.05%.

Based on the above result the RNN classifier with MFCC+MS feature extractor gives the highest accuracy of 90.05% for the Spanish database. This is too early to determine a system which is best for Speech Emotion Recognition but the Fourier transform method is the most used method in speech recognition (L. Kerkeni et al., 2018).

B.A. Malik and his team (2017) proposed a feature learning system powered by a discriminativeCNN which uses spectrograms to recognize emotion fromspeech.

Short term Fourier transform is used to generate spectrograms from speech input. Convolutional Neural Networks were used for classification of spectrograms. Emotion was predicted using a majority voting scheme from multiple spectrograms. Berlin Emotional Database was used for training andevaluation.

The accuracy for emotion prediction using CNN with rectangular kernels using Berlin Emotional Database were highest for angry which was 99.32% and lowest for happy which was 52.45%. If additional labelled data is provided and a much deeper CNN with rectangular kernels can be trained, the suggested approach can be improved even more. Using spectrograms, an experiment showed that rectangle kernels and max pooling processes are better suited for SER. (Badshah, A.M. et al., 2017).

H. Aouani and Y. Ben Ayed (2020) demonstrated how one can recognise emotions from speech to visualise the output from speech. There are many benefits of Speech Emotion Recognition explained such as-In Educational field, Automobile, Security, Communication and Health.

It has an emotion recognition system that uses parameters like 39MFCC, HNR, ZCR, TEO using Support Vector Machines and then there's usage of Auto-Encoder. There is also a parameter of Harmonic to Noise Rate (HNR). SVM is used for classifications of emotions and the autoencoder helps in the feature selection method.

There is an RML Emotion Database which contains 720 emotion expression samples that were used in testing taken from Ryerson Multimedia Lab. This dataset was furnished in six different languages. After achieving aseries of experiments they achieved a better identification rate. Firstly, they presented the performance of a system based on a fusion of HNR. Secondly, it is the application of auto-encoder dimensions. The result of this system shows effectiveness of achieving good results (Aouani, H. et al., 2020).

Woo, B.S. and team (2021) emphasized the need to understand emotions from speech or voice for better understanding. Generally emotions are accompanied by different changes in one's body. To make better results they were required by a high speech database. They constructed a Korean Emotional Speech Database (K- EmoDB) and used it with the RNN network. To check the sudden changes in voice it was loaded with MFCC, Chroma, spectral features, harmonic features and others. The feature extraction tool Essential is used for harmonic feature extraction which is available as a free open source tool. A LSTM model is used so that they can recognize emotion from speech. Generally emotions can only be recognized when one listens to the complete audio or speech.

The first experiment using the K-EomDB model database has achieved approximately 65.89% accuracy and the second experiment when LSTM was used the accuracy was 62.63%. They noted a feature benefit that if one uses RNN model then it can increase the performance. Although they are still dealing with the problem of finding a variety of emotions (Woo, B. et al., 2021).

3 METHODOLOGY, SETUP and DESIGN of EXPERIMENT

Setup

3.1 Name of Algorithms Used

The algorithms used here include MFCC (Mel Frequency Cepstral), Chroma Feature Extraction, and Mel (Mel Spectrogram Frequency).

3.2 Types of Databases

Natural, Simulated (Acted), and Elicited (Induced) emotional speech databases are the three types of databases that are utilized to construct speech emotion recognition systems.

Natural Database- The majority of natural speech datasets come from talk shows, contact centre recordings, radio conversations, and other similar sources. These unscripted talks are sometimes referred to as real-life speeches. The data is more difficult to get while processing.

Simulated Database- Professional or semi-professional actors’ record performed speech databases in sound-proof studios. When compared to other ways, creating such a database is quite simple.

Elicited Database- Elicited speech databases are made by putting speakers in a simulated emotional scenario that can elicit a variety of emotions. The emotions are near to real ones, despite not being fully evoked.

3.3 Dataset

The study team scoured the internet and discovered many dataset sets, some of which are shown below:

1. Ryerson Audio-Visual Database of Emotional Speech andSong.

2. Crowd-sourced Emotional Multimodal ActorsDataset.

3. Surrey Audio-Visual ExpressedEmotion.

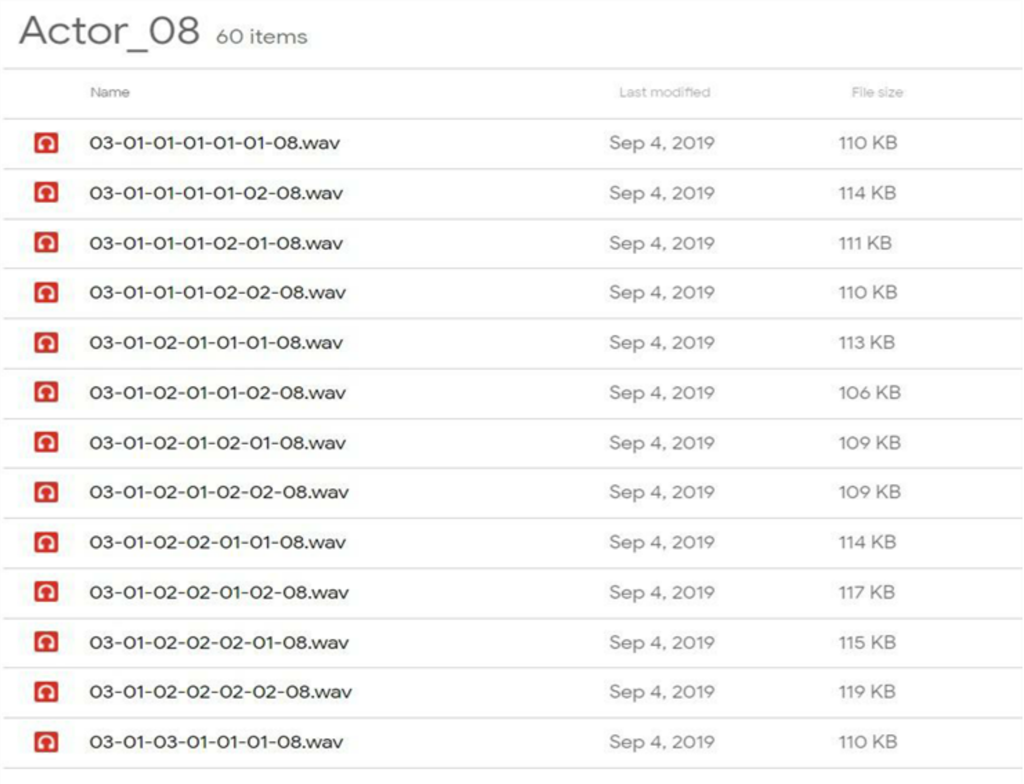
4. Toronto emotional speech set.

The research team uses the Ryerson Audio-Visual Database of Emotional Speech and Song is used by the study team (RAVDESS). It can be downloaded for free and has 24 experienced actresses (12 Females and 12 male). The author team personally checked and confirmed the authenticity of the emotions recorded by the voice artists in the RAVDESS dataset is correct.

3.4 RAVDESS Data Set Attributes and URL

Both in speech and in the lyrics, there are expressions of calmness, happiness, joy, sadness, anger, fear, surprise, and contempt (as per Fig. 7).

Speech-emotion-recognition-ravdess-data.zip - Google Drive.



Actor_08 60 items

Name	Last modified	File size
03-01-01-01-01-01-08.wav	Sep 4, 2019	110 KB
03-01-01-01-01-02-08.wav	Sep 4, 2019	114 KB
03-01-01-01-02-01-08.wav	Sep 4, 2019	111 KB
03-01-01-01-02-02-08.wav	Sep 4, 2019	110 KB
03-01-02-01-01-01-08.wav	Sep 4, 2019	113 KB
03-01-02-01-01-02-08.wav	Sep 4, 2019	106 KB
03-01-02-01-02-01-08.wav	Sep 4, 2019	109 KB
03-01-02-01-02-02-08.wav	Sep 4, 2019	109 KB
03-01-02-02-01-01-08.wav	Sep 4, 2019	114 KB
03-01-02-02-01-02-08.wav	Sep 4, 2019	117 KB
03-01-02-02-02-01-08.wav	Sep 4, 2019	115 KB
03-01-02-02-02-02-08.wav	Sep 4, 2019	119 KB
03-01-03-01-01-01-08.wav	Sep 4, 2019	110 KB

Figure 7. Audio Files Sample

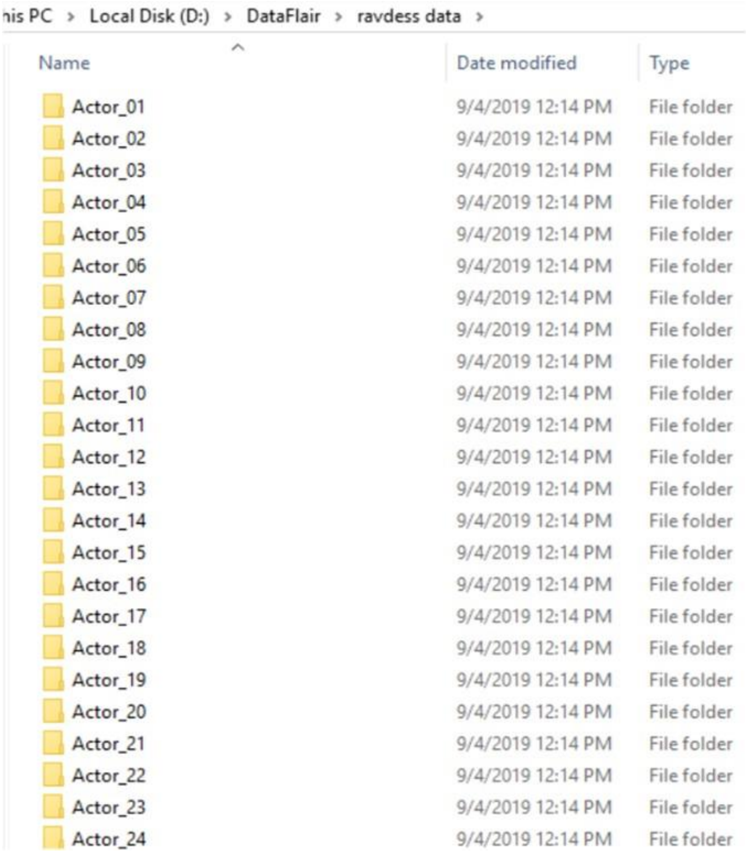
3.5 Metadata and Sample Dataset

Pitch, energy, and intensity are all key factors in expressing the emotional content of a speech. The speech attribute that were used to determine the emotions were -

Pitch- It provides the wave's greatest peak, allowing us to gauge our emotional condition.

Energy- The most effective variable for recognizing emotions.

Intensity- The parameter is utilized to determine the physical energy and volume of speech (as per Fig. 8).



Name	Date modified	Type
Actor_01	9/4/2019 12:14 PM	File folder
Actor_02	9/4/2019 12:14 PM	File folder
Actor_03	9/4/2019 12:14 PM	File folder
Actor_04	9/4/2019 12:14 PM	File folder
Actor_05	9/4/2019 12:14 PM	File folder
Actor_06	9/4/2019 12:14 PM	File folder
Actor_07	9/4/2019 12:14 PM	File folder
Actor_08	9/4/2019 12:14 PM	File folder
Actor_09	9/4/2019 12:14 PM	File folder
Actor_10	9/4/2019 12:14 PM	File folder
Actor_11	9/4/2019 12:14 PM	File folder
Actor_12	9/4/2019 12:14 PM	File folder
Actor_13	9/4/2019 12:14 PM	File folder
Actor_14	9/4/2019 12:14 PM	File folder
Actor_15	9/4/2019 12:14 PM	File folder
Actor_16	9/4/2019 12:14 PM	File folder
Actor_17	9/4/2019 12:14 PM	File folder
Actor_18	9/4/2019 12:14 PM	File folder
Actor_19	9/4/2019 12:14 PM	File folder
Actor_20	9/4/2019 12:14 PM	File folder
Actor_21	9/4/2019 12:14 PM	File folder
Actor_22	9/4/2019 12:14 PM	File folder
Actor_23	9/4/2019 12:14 PM	File folder
Actor_24	9/4/2019 12:14 PM	File folder

Figure 8. Sample Data Set Distribution

3.6 Dataset Size (Storagespace)

This large dataset consists of 7356 files that 247 individuals assessed 10 times for emotional sincerity, intensity, and validity. The complete dataset comes in at 24.8 GB and includes 24 actors, but the author team used only the audio data available in the RAVDESS dataset.

3.7 Functional and Non Functional Requirements

Functional requirements are product features or functions that developers must implement to enable users to accomplish their tasks. So, it's important to make them clear both for the development team and the stakeholders. Generally, functional requirements describe system behavior under specific conditions.

Non-functional requirements or NFRs are a set of specifications that describe the system's operation capabilities and constraints and attempt to improve its functionality. These are basically the requirements that outline how well it will operate including things like speed, security, reliability, data integrity, etc.

3.7.1 Hardware Requirement

Processing Power: 1.7 GHz or above

Memory: 2GB or above

Storage: For applications 100 MB or above and for databases 600 MB or above

Sound: Sound card required Microphone for listening speech

3.7.2 Software Requirement

IDE: Any IDE which supports python

Frontend: Kivy, a python framework for creating UI

Backend: Python 3.6 or higher versions

Libraries additionally installed are librosa, soundile, sklearn, numpy, pickle, and pyaudio.

3.7.3 Network Requirement

Cloud or Distributed Environment or can be used by a Single user

3.7.4 OS Requirement

Windows 10 is used for back-end work but other versions like Windows 7 or 8 or 11 will also work. It is providing background for running python software and downloading the dataset. Other operating systems like Linux, and macOS can also be used.

3.7.5 Database Requirement

Database is required so that we can easily store the information and can access it later without any problem. The research team uses the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Structured Query language (SQL) is a programming language that develops databases.

3.7.6 Storage Requirement

System's storage is required to access audio files downloaded from RAVDESS Dataset. For applications 100 MB or above and for datasets additional 25 GB will be the basic storagerequirement.

3.8 Front End

For the front end part, the research team is looking to design an application for this mowhatdel. The Researcher team is thinking of using Kivy GUI which is a python based framework for developing an application.

3.9 Back End

For the backend, Python Programming language is used. Python is a high-level, interpreted, and general-purpose programming language. Python is an interpreted language which can be used in machine learning. It has additional libraries such as librosa for analyzing music, sound file for using sound files from datasets, NumPy for numerical calculations, pickle for serializing and desterilizing data objects, pyaudio for taking audio files from user's microphone, and sklearn contains ML algorithms.

3.10 Steps of Execution

1. Load the data as Inputspeech.

2. Extracting features fromspeech.

3. Split Dataset for training andtesting.

4. Initialize MLP classifier to train themodel.

5. Getting emotions from input data asOutput.

5. Check the accuracy of the model.

3.11 Different Diagrams

As presented below:

3.11.1 Flow Chart of the Activity

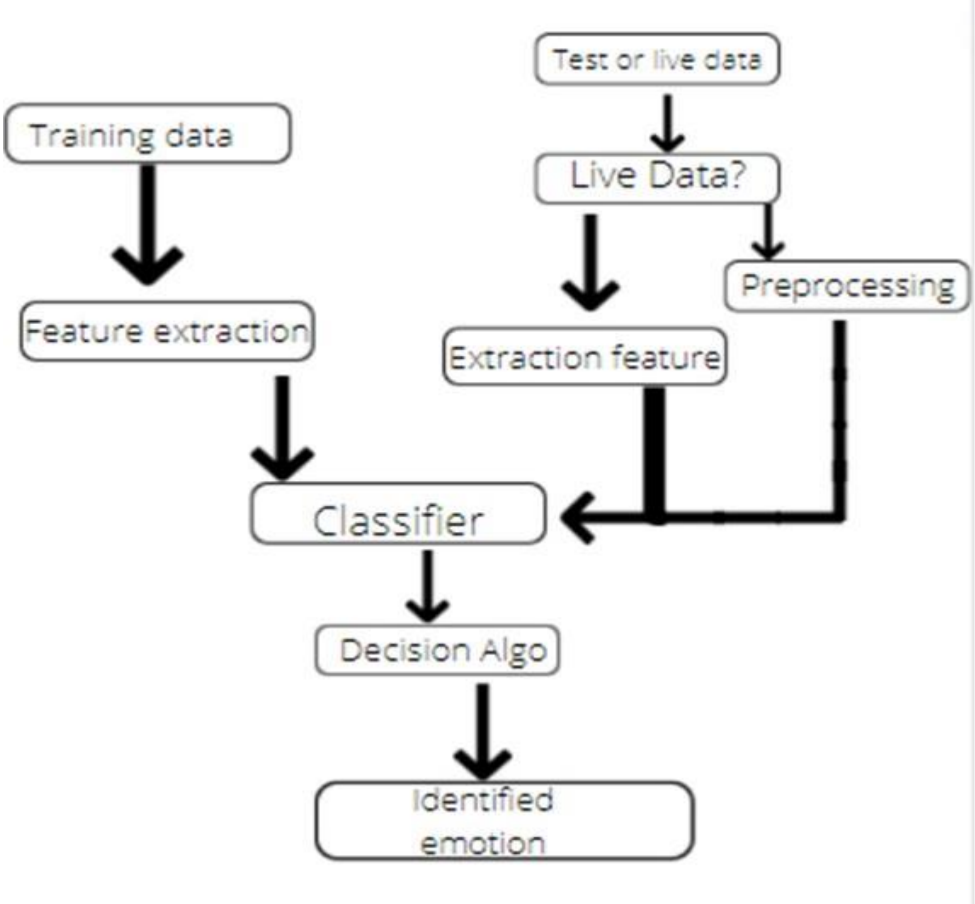


Figure 9. Flowchart of Emotion Recognition

The dataset in this Flowchart diagram contains sound files that load training data into the feature extractor. For live data, it will first pre-processes then extract features. For the training model,features are extracted and saved for the classifier and then the analyzing identifies emotion from speech (asper the Figure9).

3.11.2 ER Diagram

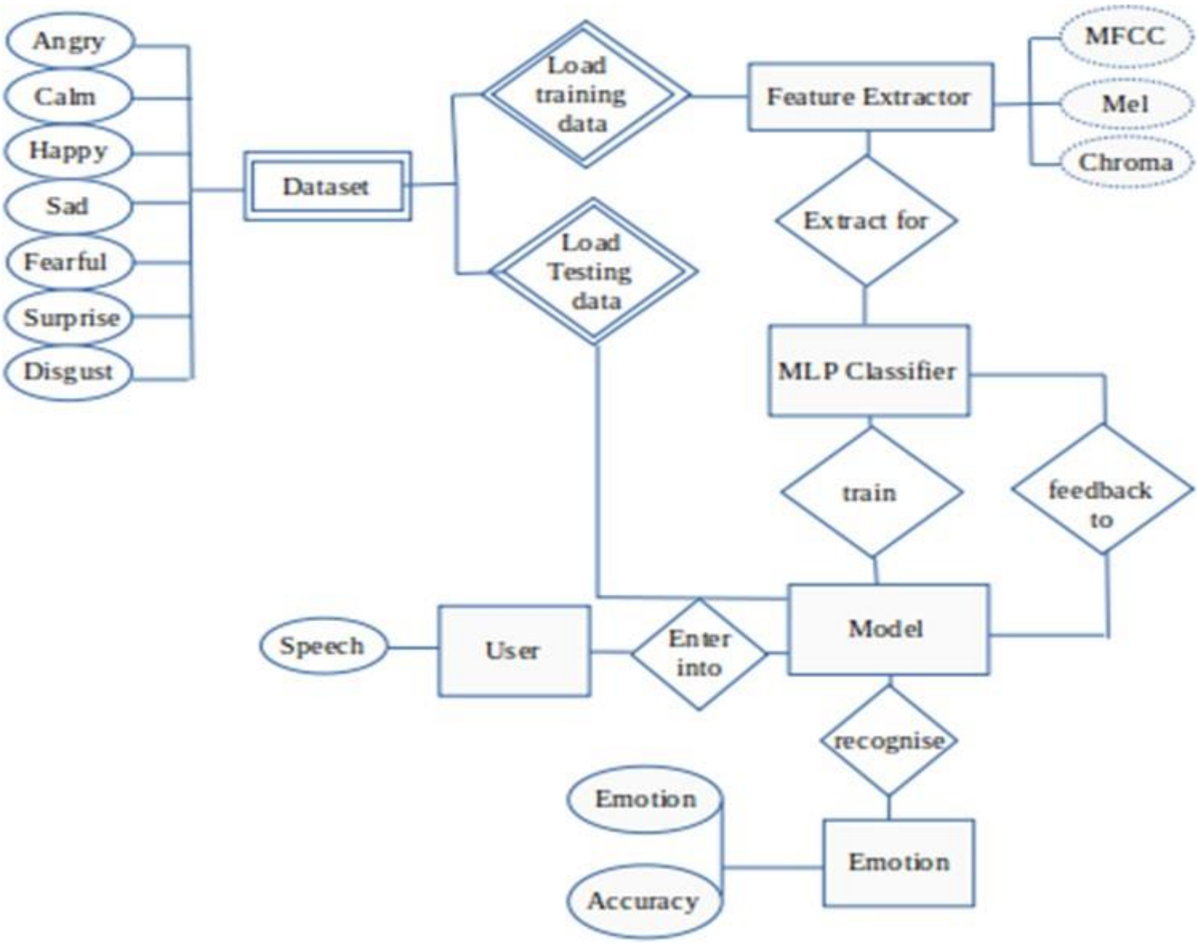


Figure 10. ER Diagram describing connection with dataset.

In this ER diagram, the dataset contains sound files which load training data into feature extractor. Features are extracted and saved for classifier for training model. Testing data is used on model for accuracy standards and speech from user is entered to recognize emotion and user data is again sent for training model (as per Fig. 10).

3.11.3 Object Oriented Class Diagram

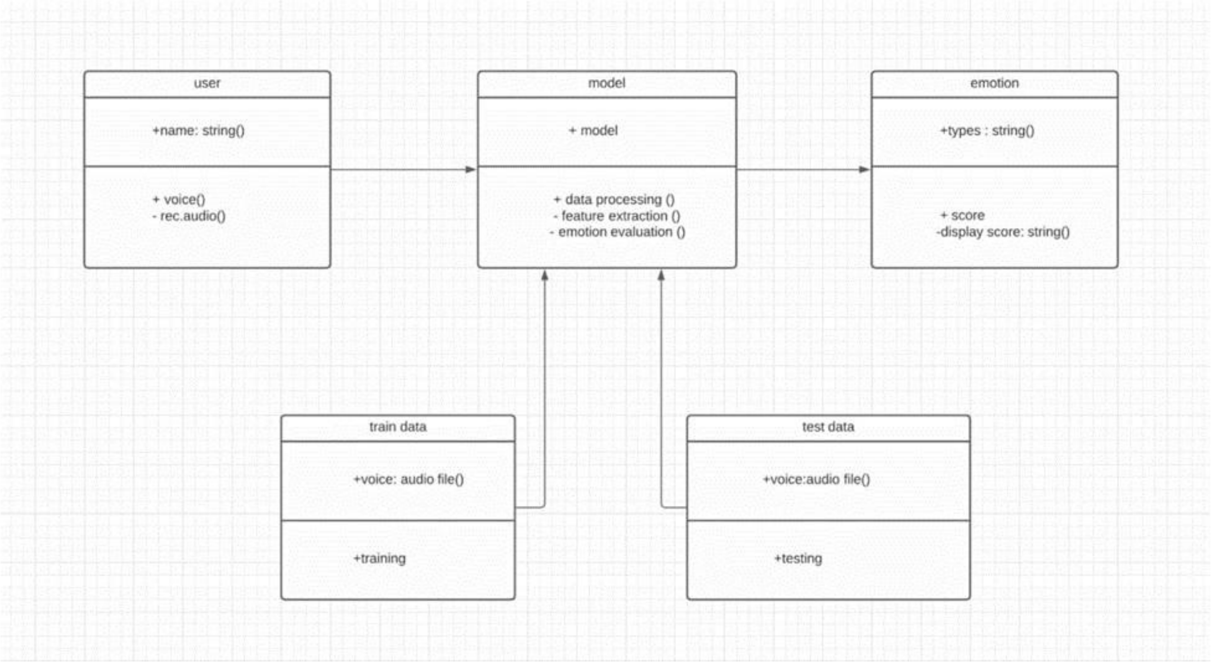


Figure 11. Object oriented class diagram with different attributes and classes.

The basic structure of object-oriented modeling is the Class diagram. The above diagram consists of 5 classes’ user, model, emotion, train, and test. It further consists of different attributes such as audio, model, evaluation, and data. Under attribute, it shows the operations that are to be executed by the model. It has operations like feature extraction, prediction, recognition, etc. The straight line between different classes is showing associativity of each class with the other (as per Fig. 11).

3.11.4 Data Flow Diagram (level 0, level 1, level2)

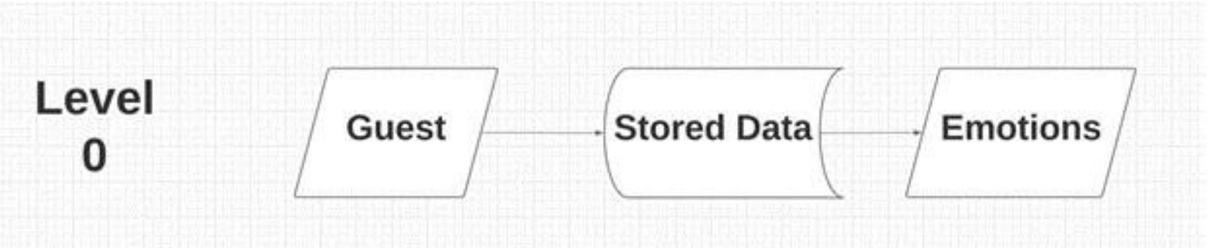


Figure 12. Level 0 of Data Flow Diagram for Working Procedure of Extracting Emotions from Voice.

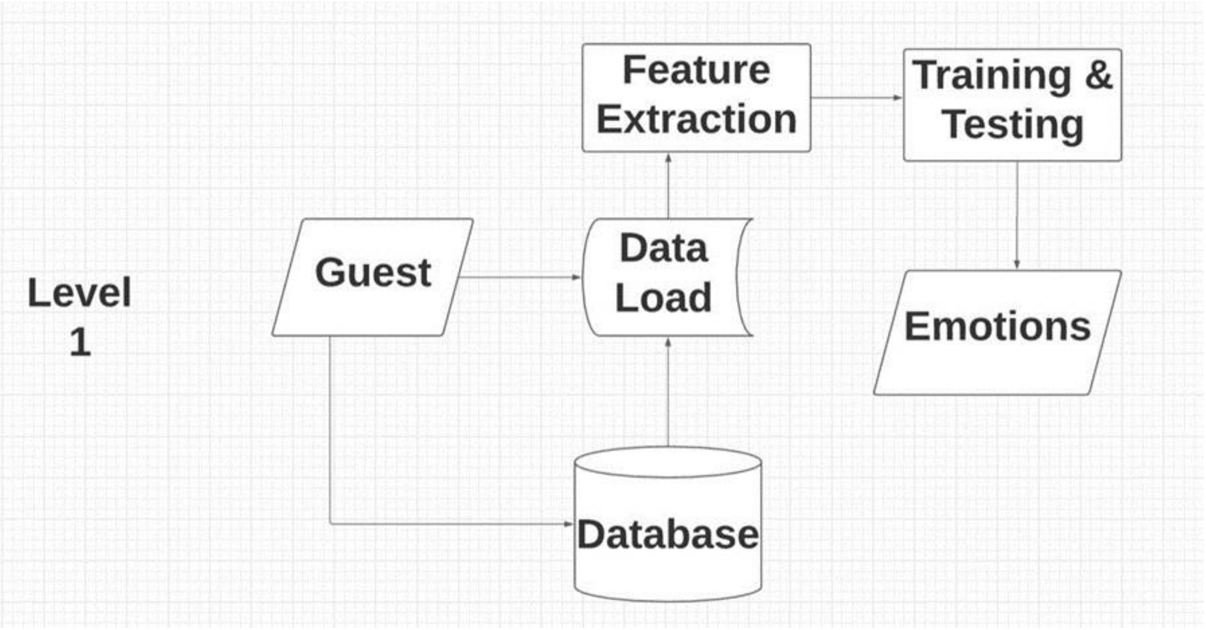


Figure 13. Level 1 of Data Flow Diagram for Working Procedure of Extracting Emotions from Voice.

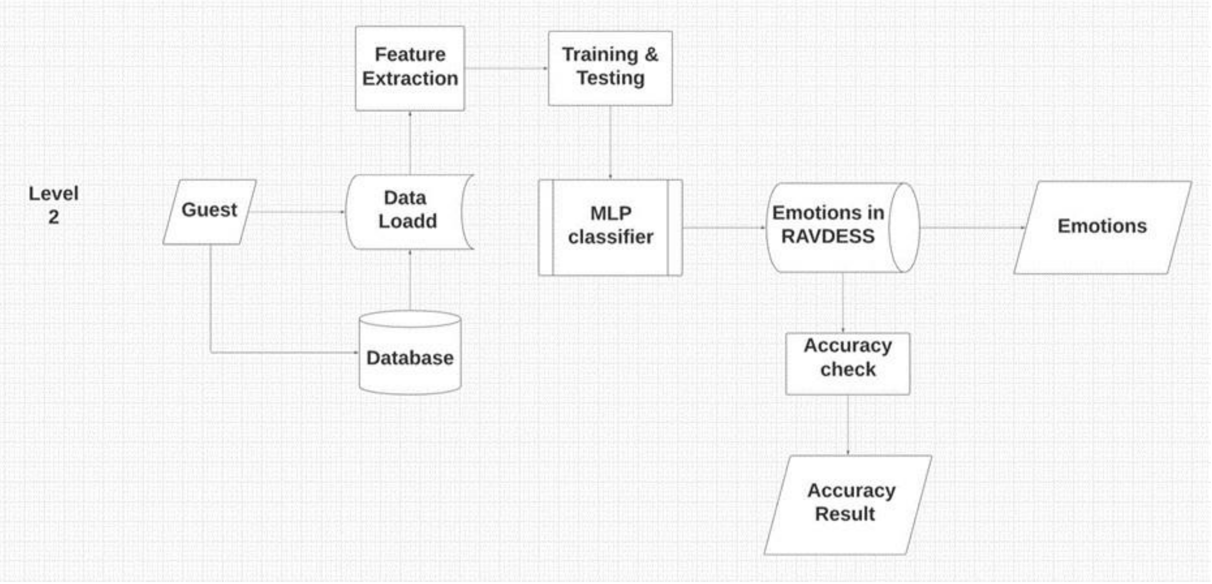


Figure 14. Level 2 of Data Flow Diagram for Working Procedure of Extracting Emotions from Voice.

Diagrams representing 3 levels of the Emotions Recognition process. Level 0 only shows how one's data is used to recognize emotions from it. Level 1 is showing the feature extraction and Training-Testing process and finally the Emotion recognition step. Level 2 is a more deep evaluation showing each process briefly. Once the data is loaded it goes through a feature extraction process, training and testing process, and then MLP Classifier starts working. Finally, we get Emotions and on parallel Accuracy check (as per Fig.12, 13, 14).

3.11.5 Use Case Diagram

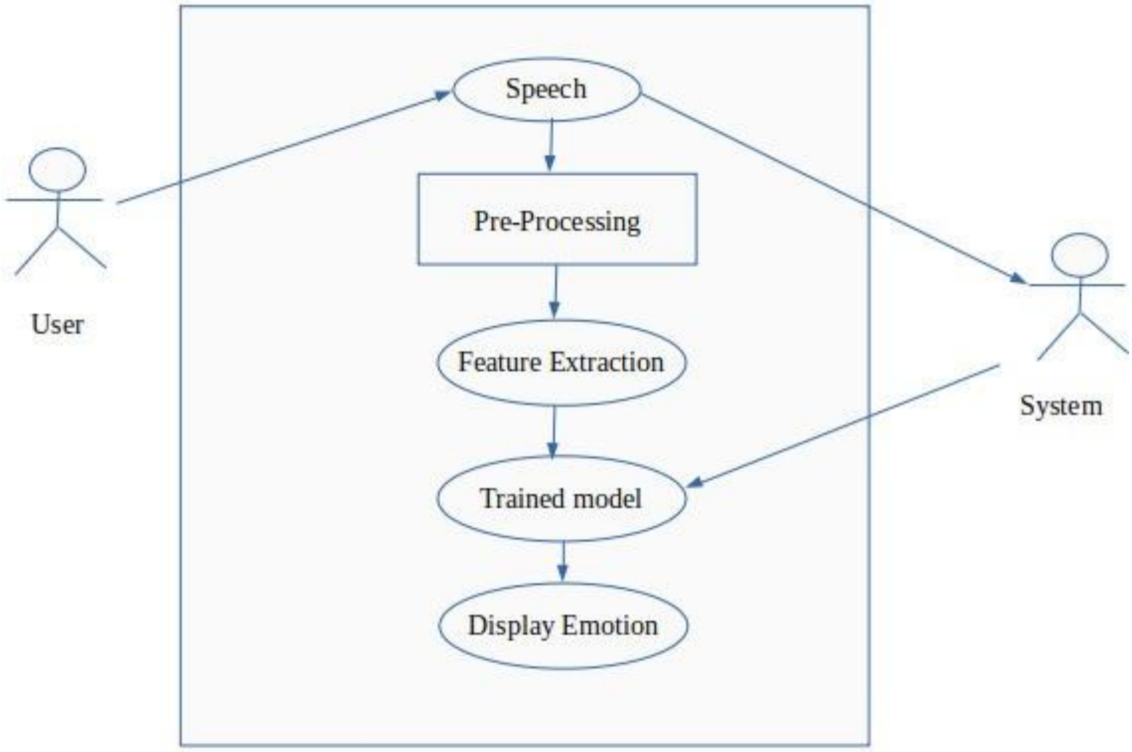


Figure 15. User Case Diagram for Extraction Emotion from User Speech

Figure presented above explains that whenever a user tries to speak, it will be predicted as it suser voice or Well-trained audio if it is sample test audio then will directly reflect the classifier otherwise goes through of pre-processing for feature extraction that means it will pre-process live data before extracting features. Features are taken from the training model, kept for the classifier, and then analysis is used to determine the emotion in the speech, further goes to system and system to detect the emotion and display it to the user (MFCC) (as per fig 15).

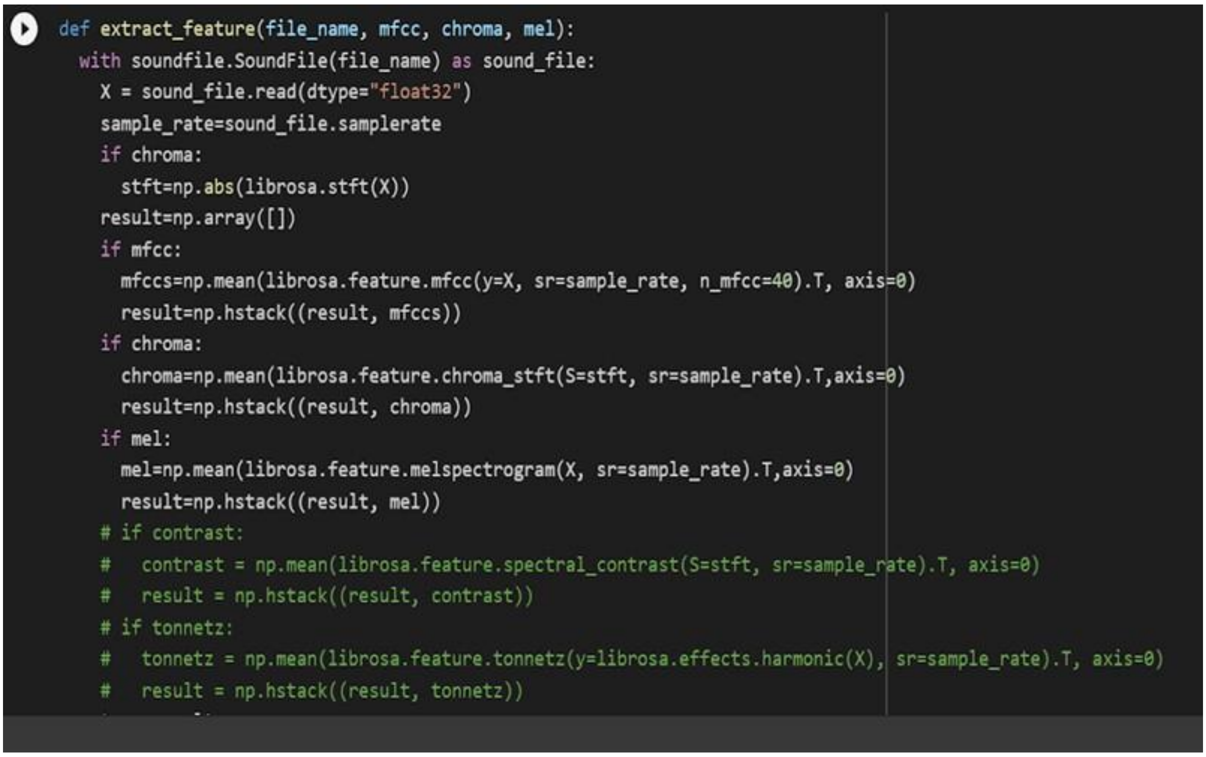
4 Results and Discussions

This has been presented in two parts:

- The implementation results
- Discussions on Findings

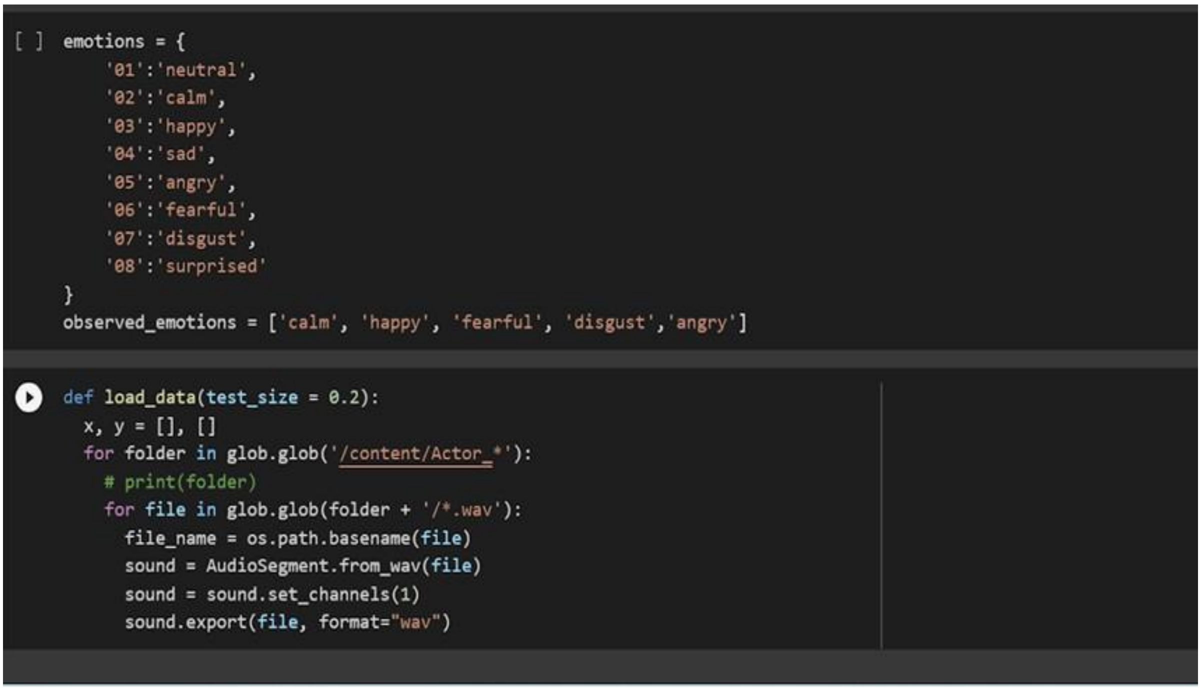
4.1 Coding Snapshots

Pl. refer figure 16, 17 and 18 for the important coding snippets of this research work in Python.



```
def extract_feature(file_name, mfcc, chroma, mel):
  with soundfile.SoundFile(file_name) as sound_file:
    X = sound_file.read(dtype="float32")
    sample_rate=sound_file.samplerate
    if chroma:
      stft=np.abs(librosa.stft(X))
    result=np.array([])
    if mfcc:
      mfccs=np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T, axis=0)
      result=np.hstack((result, mfccs))
    if chroma:
      chroma=np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T,axis=0)
      result=np.hstack((result, chroma))
    if mel:
      mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
      result=np.hstack((result, mel))
    # if contrast:
    #   contrast = np.mean(librosa.feature.spectral_contrast(S=stft, sr=sample_rate).T, axis=0)
    #   result = np.hstack((result, contrast))
    # if tonnetz:
    #   tonnetz = np.mean(librosa.feature.tonnetz(y=librosa.effects.harmonic(X), sr=sample_rate).T, axis=0)
    #   result = np.hstack((result, tonnetz))
```

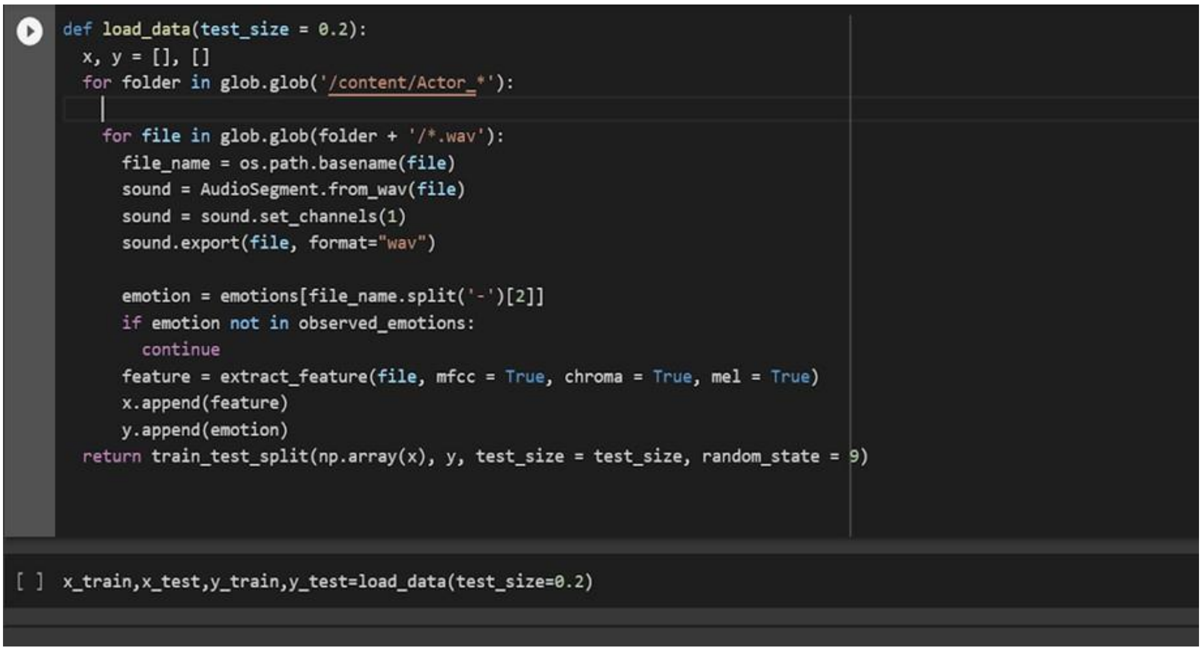
Figure 16. Coding Snapshot for MLP Classifier



```
[ ]  emotions = {
         '01':'neutral',
         '02':'calm',
         '03':'happy',
         '04':'sad',
         '05':'angry',
         '06':'fearful',
         '07':'disgust',
         '08':'surprised'
     }
     observed_emotions = ['calm', 'happy', 'fearful', 'disgust','angry']
```

```
def load_data(test_size = 0.2):
  x, y = [], []
  for folder in glob.glob('/content/Actor_*'):
    # print(folder)
    for file in glob.glob(folder + '/*.wav'):
      file_name = os.path.basename(file)
      sound = AudioSegment.from_wav(file)
      sound = sound.set_channels(1)
      sound.export(file, format="wav")
```

Figure 17. Coding Snapshot for Emotion Detection



```
def load_data(test_size = 0.2):
  x, y = [], []
  for folder in glob.glob('/content/Actor_*'):

    for file in glob.glob(folder + '/*.wav'):
      file_name = os.path.basename(file)
      sound = AudioSegment.from_wav(file)
      sound = sound.set_channels(1)
      sound.export(file, format="wav")

      emotion = emotions[file_name.split('-')[2]]
      if emotion not in observed_emotions:
        continue
      feature = extract_feature(file, mfcc = True, chroma = True, mel = True)
      x.append(feature)
      y.append(emotion)
  return train_test_split(np.array(x), y, test_size = test_size, random_state = 9)
```

```
[ ]  x_train,x_test,y_train,y_test=load_data(test_size=0.2)
```

Figure 18. Coding Snapshot for Test Data

The Codes in the Python are representing the accuracy is used as the performance metric for the conclusion of this study. Accuracy is the total predictions that the model got right. Higher the accuracy of the classifier, better the classifier is inpredicting the correct emotion.

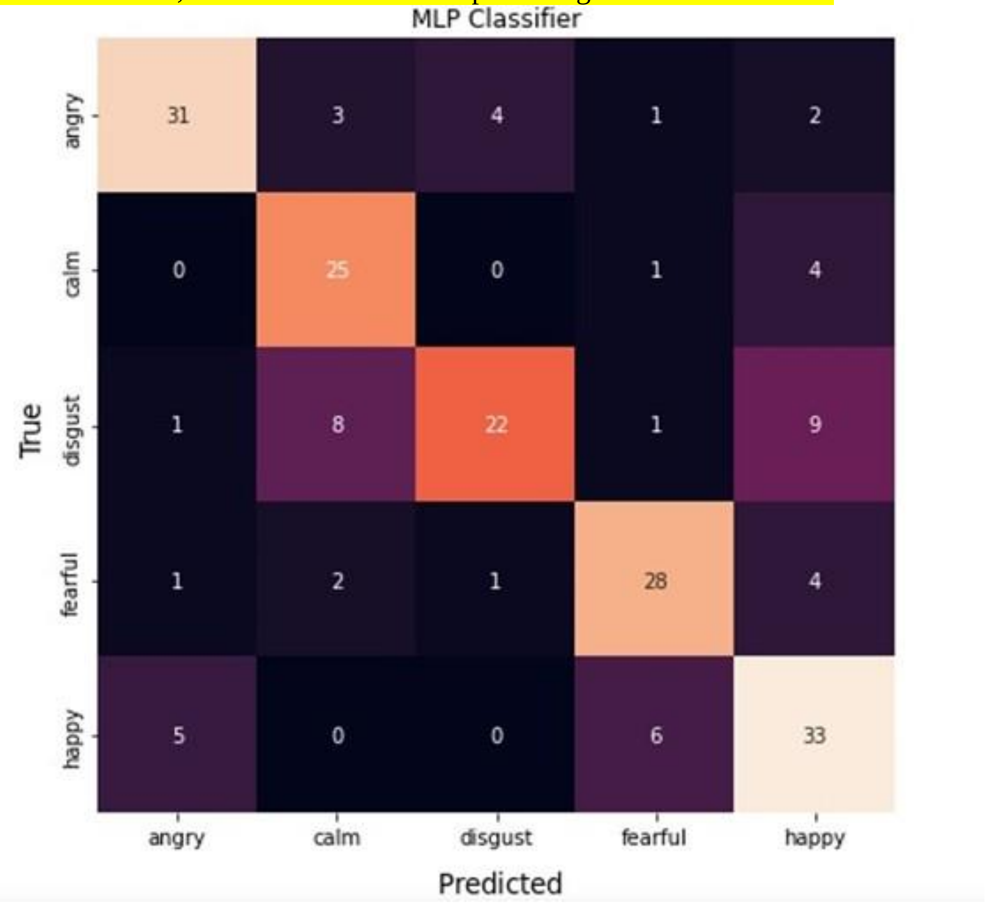


Figure 19. Confusion matrix of MLP Classifier Algorithm.

The MLP Classifier confusion matrix is represented above. The above figure is the confusion matrix of the Multi Layer Perceptrons (MLP) Classifier. In the Confusion matrix. Y-axis represents the actual emotions and the X-axis represents the predicted value by the model. According to theabove figure, the

highest accurately predicted emotion is angry and
happy out of the other emotions (As per Figure 19).

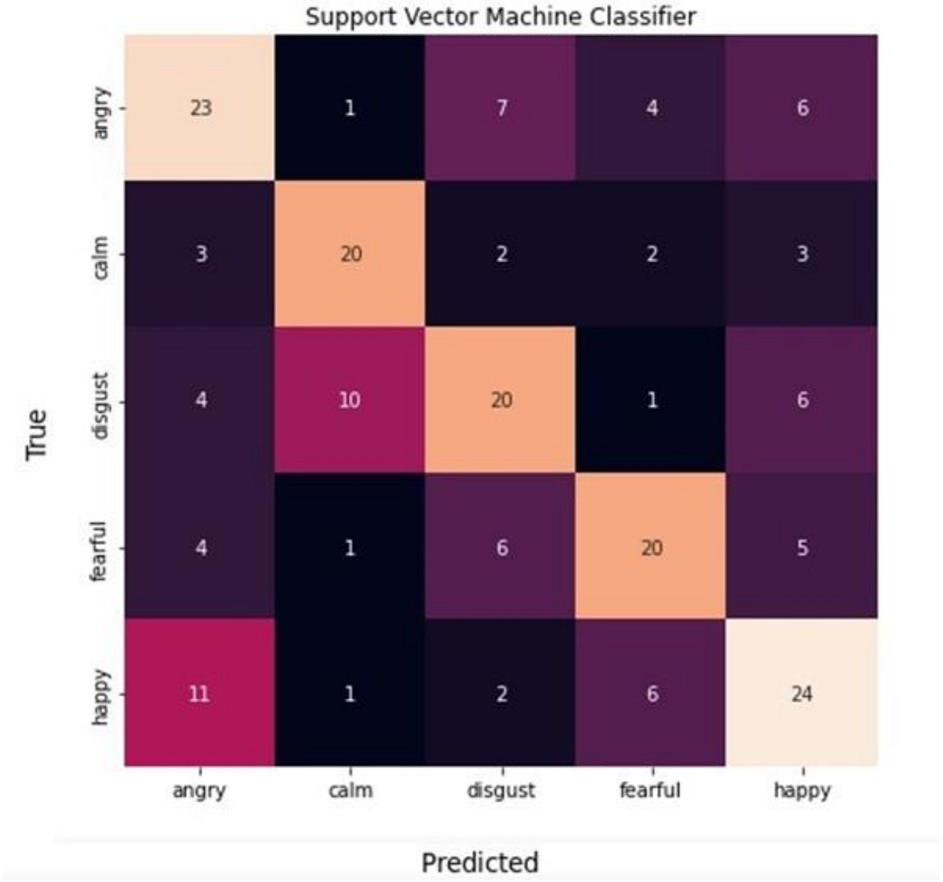


Figure 20. Confusion matrix of SVM Classification Algorithm.

The SVM Classifier confusion matrix is represented above. The above figure is the confusion matrix of the Support Vector Machine (SVM) Classifier. In the confusion matrix. Y-axis represents the actual emotions and the X-axis represents the predicted value by the model. According to the above figure, the highest accurately predicted emotion is angry and
Happy out of the other emotions (As per Figure 20).

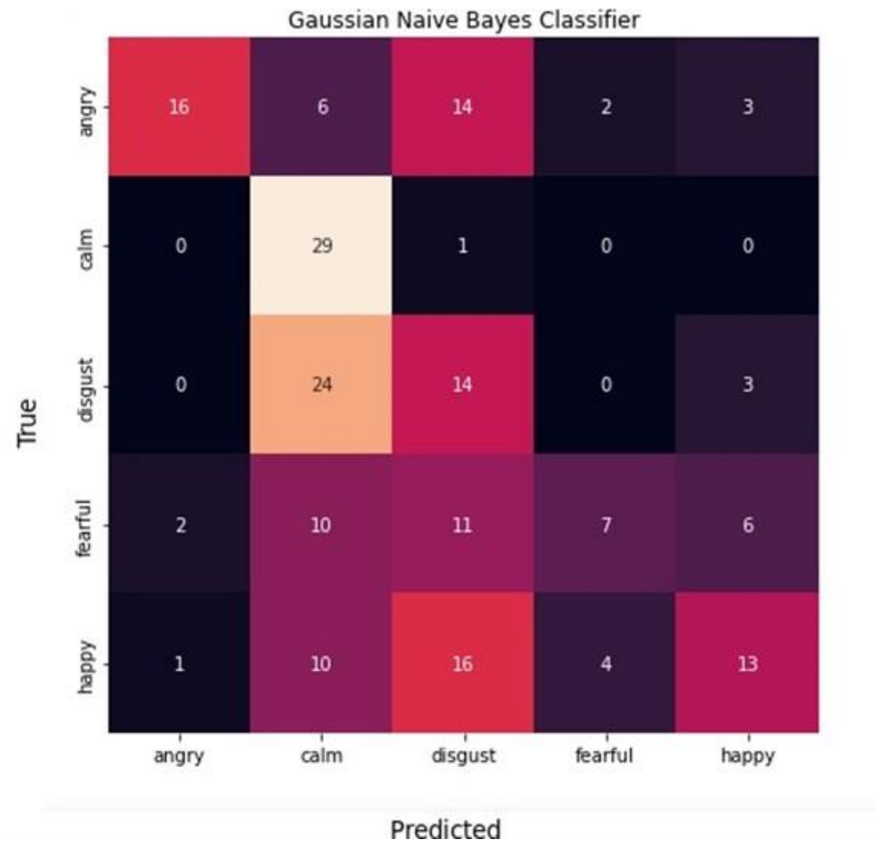
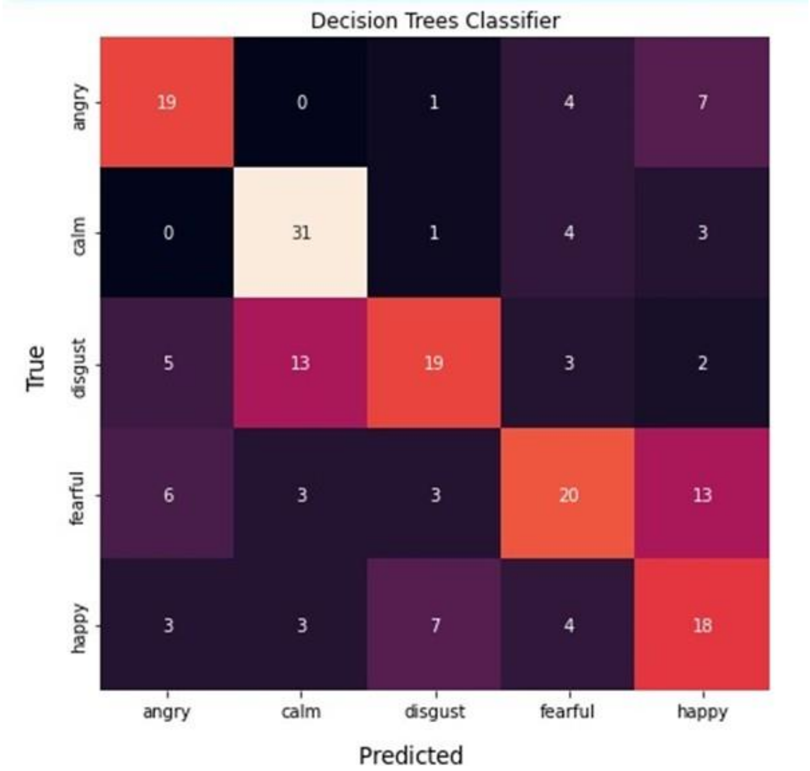


Figure 21. Confusion Matrix of Gaussian Naïve Bayes Classifier.

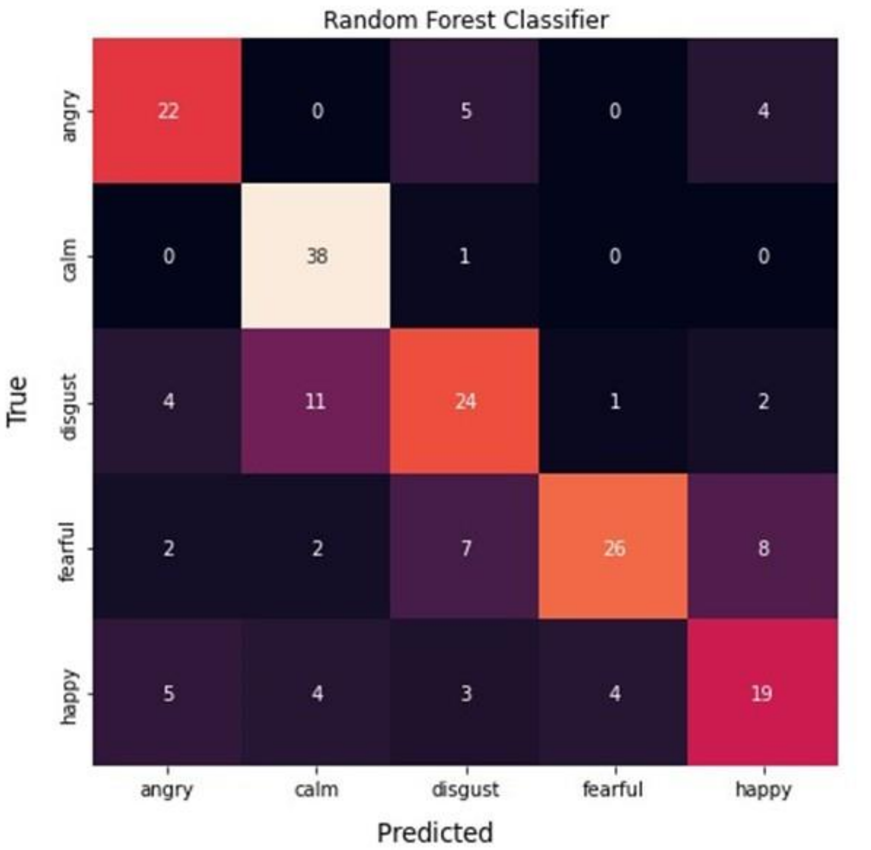
The Naïve Bays Classifier confusion matrix is represented above. The above figure is the confusion matrix of the Gaussian Naive Baye’s (GNB) Classifier. In the confusion matrix. Y-axis represents the actual emotions and the X-axis represents the predicted value by the model. According to the above figure, the highest accurately predicted emotion is Calm and Angry out of the other emotions (As per Figure 21).



True \ Predicted	angry	calm	disgust	fearful	happy
angry	19	0	1	4	7
calm	0	31	1	4	3
disgust	5	13	19	3	2
fearful	6	3	3	20	13
happy	3	3	7	4	18

Figure 22. Confusion matrix of Decision Tree Classification Algorithm.

The above figure is the confusion matrix of the Decision Tree Classifier. In the confusion matrix.Y-axis represents the actual emotions and the X-axis represents the predicted value by the model.
According to the above figure, the highest accurately predicted emotion is Calm and Fearful out of the other emotions (As per Figure 22).



True \ Predicted	angry	calm	disgust	fearful	happy
angry	22	0	5	0	4
calm	0	38	1	0	0
disgust	4	11	24	1	2
fearful	2	2	7	26	8
happy	5	4	3	4	19

Figure 23. Confusion matrix of Random Forest Classifier Algorithm.

The above figure is the confusion matrix of the Random Forest Classifier. In the confusion matrix.Y-axis represents the actual emotions and the X-axis represents the predicted value by the model. According to the above figure, the highest accurately predicted emotion is Calm and Fear out of the other emotions (As per Figure 23).

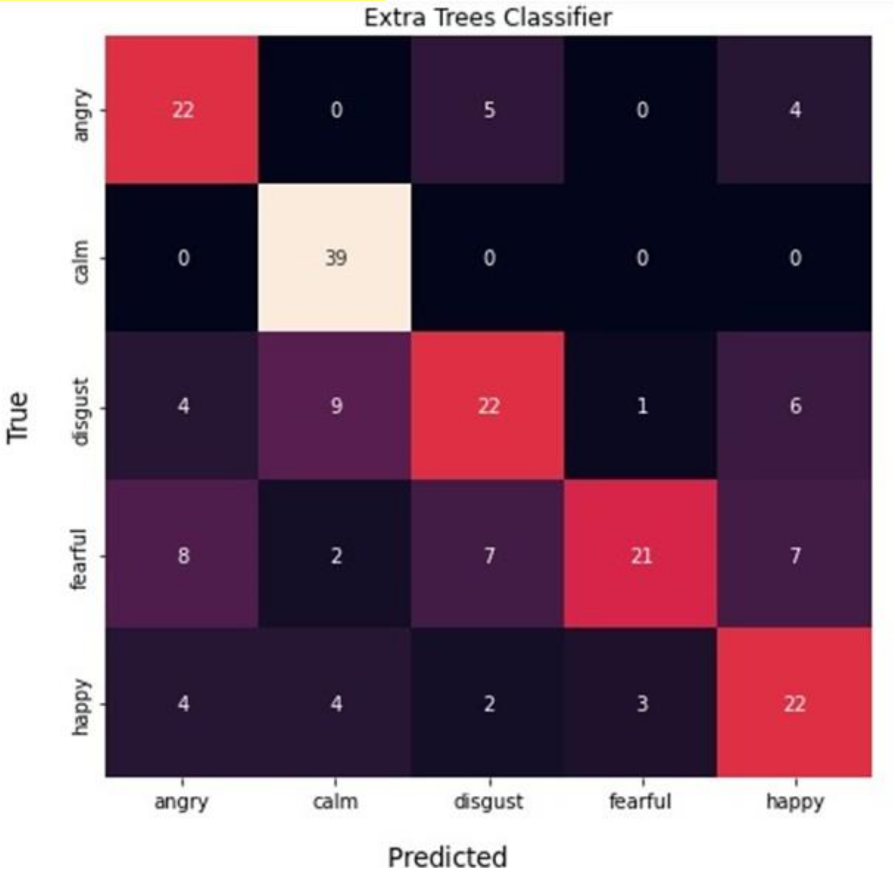


Figure 24. Confusion Matrix of Extra Tree Classifier.

The above figure is the confusion matrix of the Support Vector Machine (SVM) Classifier. In the confusion matrix. Y-axis represents the actual emotions and the X-axis represents the predicted value by the model. According to the above figure, the highest accurately predicted emotion is Calm out of the other emotions (As per Figure 24).

4.2 Discussions

From above implemented visualizations, one can easily compare the results within the works and with existing and already done efforts in this domain (Pl. refer Table 1 and Table 2).

The Decision Tree classifier works on creating two nodes, one as Decision node and second as Leaf node and these nodes further produce output. The researcher team applied the algorithms on the dataset collected and also performed the real time testing on the voice that gives desired output in the form of emotions. The MLP classifier had higher average recall and F1-score while the Extra Trees Classifier had a higher average Precision. A no. of comparision charts have been presented in support of this fact (as per Figures 25, 26 and 27).

Comparison Charts:

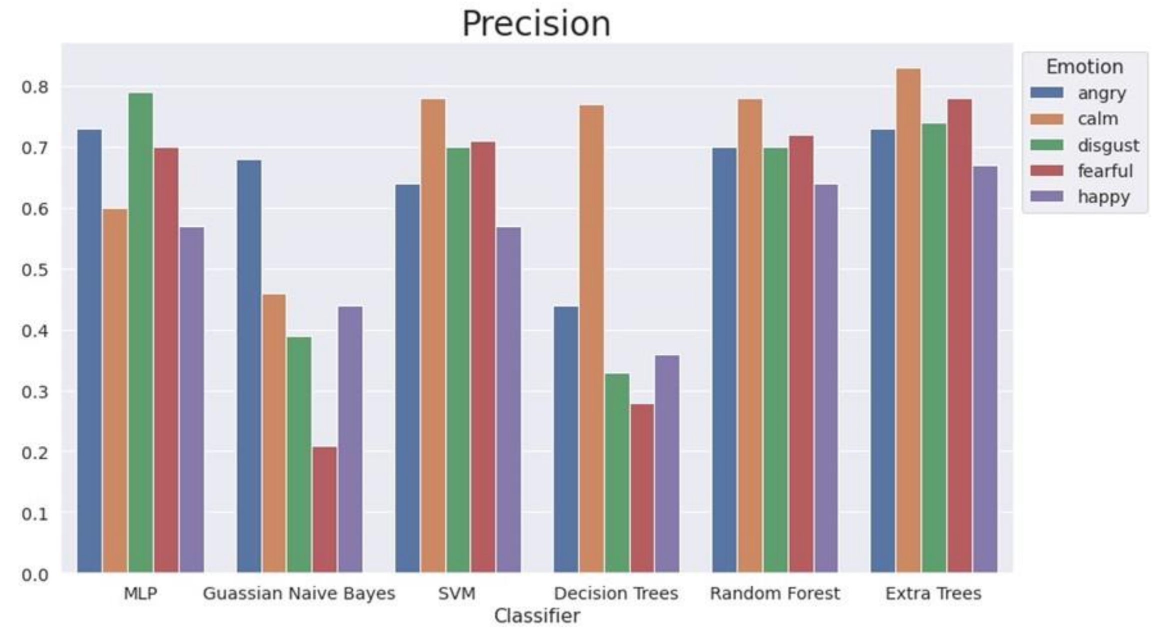


Figure 25.Comparison Charts of Precision of Six Algorithms.

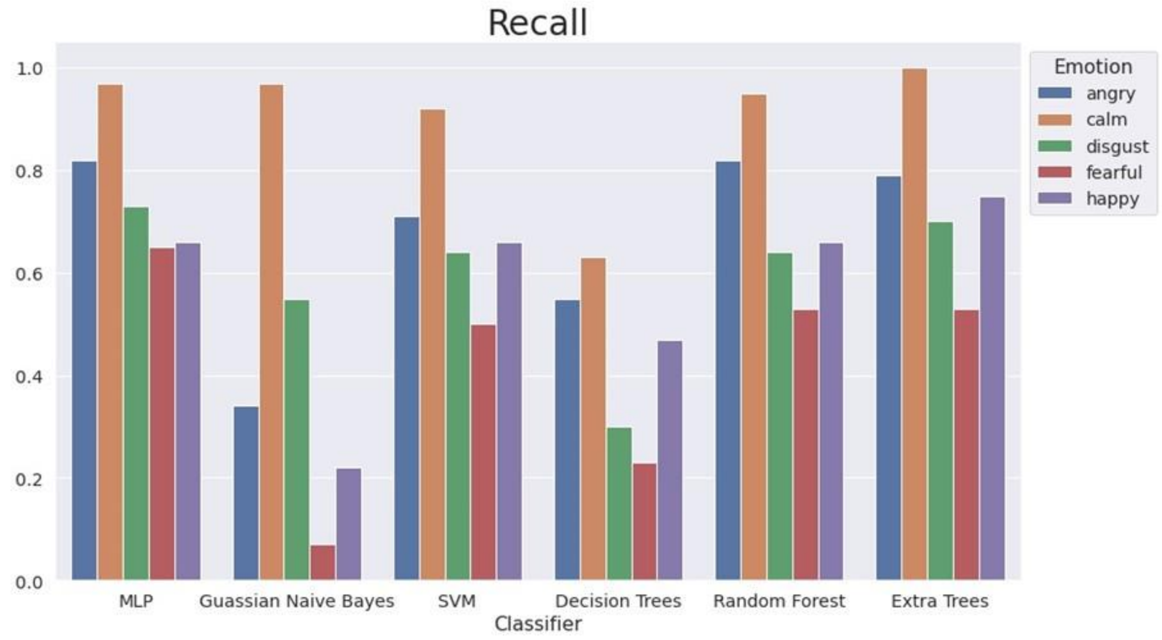


Figure 26. Comparison Charts of Recall of Six Algorithms.

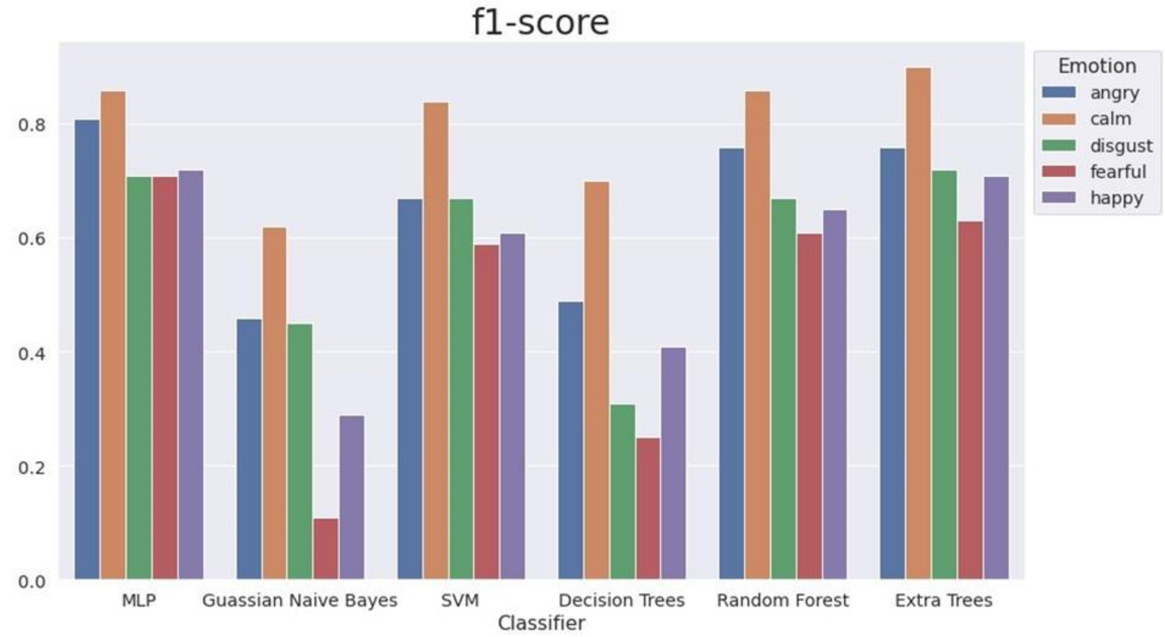


Figure 27. Comparison Charts of f1-score of Six Algorithms.

Table 1. Tabular Summary for Algorithms and Their Accuracy

Features	**Multi-Layer Perceptron**	**Support Vector Machine**	**Gaussian Naive Bayes**	**Decision Tree Classifier**	**Random Forest**	**Extra Trees Algorithm**
Accuracy (%)	75.00	51.04	64.58	52.60	65.10	67.28
Best Predicted Emotion	Anger, Happy	Anger, Happy	Calm	Calm	Calm	Calm
Worst Predicted Emotion	Calm, Disgust	Calm, Disgust, Fear	Fear, Happy	Anger, True, Fear, Happy	Anger, Happy	Anger, Fear, Disgust, Happy

The conversion of Speech to Emotion is conducted by performing an algorithm on a dataset full of different audio files. The names of the algorithms used were SVM, Gaussian naive Baye's, MLP, Random Forest, Decision Tree and Extra Tree. The accuracy obtained for the above algorithms are 75.00%, 51.04%, 64.58%, 52.60%, 65.10%, 67.28% for MLP, Support Vector Machine, GNB, Decision Tree, Random Forest, Extra Trees Classifier respectively. Out of the above classifiers, the MLP classifier had the highest accuracy in predicting correct information (as per Table 1).

Table 2. Comparative Analysis of Presented Work with Similar Existing Works

Work Done by Authors	**Technology Used**	**Advantage and Limitations**
C. Huang (2015)	Deep Belief Networks with DBN features of 136 hours	Precisely extract emotional characteristic factors but limited accuracy
RA Khalil (2019)	CNN, RNN and LSTM networks	More steady, accurate, and robust results but challenging in real time emotion identification
L. Kerkeni, (2018)	MLR, SVM and RNN	Best Accuracy in Spanish Dataset but less in others
B.A. Mailk (2017)	CNN with Short term Fourier Transform	Berlin Emotional Database were highest for angry which was 99.32% and lowest for happy which was 52.45%.
Our Presented Approach	SVM, Gaussian naive Baye's, MLP, Random Forest, Decision Tree and Extra Tree	Face Emotions are used while spepch but The majority of speech emotional databases fall short when it comes to mimicking emotions in a genuine and understandable way.

5 NOVELTIES

- The Research paper consists of well-established speech analysis and classification techniques.
- It consists of a Dataset that is loaded with 8 different emotions from different speakers. There is the use of an unsupervised learning algorithm known as an MLPclassifier.
- Clearing the audio files is a major task in SER that may affect recognizing the emotionsas

they are mixed with background noises orpauses.

6 RECOMMENDATIONS

This paper focuses on Emotion Recognition from speech. Emotion can be recognized in other ways too. Using Natural language processing, speech from the user can be translated into the text to make an understanding of the text and find concealed emotions conveyed from words of speech. Most of the emotions of a human being can be easily seen on the face of a person which can then be used toderive features of the face and recognize emotions. Alongside facial emotion recognition, body language can also be taken as a good way to recognize emotions.

7 FUTURE RESEARCH DIRECTIONS AND LIMITATIONS

7.1 Limitations

This paper uses a particular dataset. As everyone has a different accent, it's difficult for the system to interpret everyone's emotions, which may result in an error. To get emotion recognition it should be part of the regular mother tongue English**.** It can't detect all types of emotion accurately**.** The majority of speech emotional databases fall short when it comes to mimicking emotions in a genuine and understandable way.

7.2 Future Directions

- The next research should be conducted on the big data sets with more timeduration.
- The dataset should be considered real time people voice input without delay in future research.
- In order to make the programme more user-friendly, future research should train the application in more languages, such as Hindi, Urdu,etc.

8 CONCLUSIONS

Through this manuscript, the author team successfully created a Speech Emotion Recognition System. This project successfully detects emotion from speech entered by the user. The team used python 3.8 to create their research portal. RAVDESS Dataset was used as it contains 8 different emotions from all speakers. UI has been created in Kivy Python Framework. In this project, Librosa library is used to analyse audio files and

extract features for emotion recognition, pyaudio for recording sound from user, numpy for calculations and sklearn library contains tools for machine learning.
Kivy was used for creating front end and python was used for backend. It is a ready to install application which can be installed and used.

Multi Layer Perceptron is an algorithm which works as a neural network whereas SVM works. Gaussian Algorithm is based on applying Baye's theorem with strong assumptions. Random forest is like creating a number of decision trees on the provided dataset so that one can predict emotions.

The Manuscripts presents a way to upcoming researches for execution on the big data sets with more time duration of speech contents. It also emphasizes that the dataset should be considered real time people voice input without delay in future research and in order to make the programme more user-friendly, future research should train the application in more languages, such as Hindi, Urdu, and regional languages of south Asian continent etc.

REFERENCES

Aouani, H., B. Ayed, Y. (2020). Speech Emotion Learning with Deep Learning, *24th International Conference on Knowledge-based and AI & Engineering Systems, Volume no. 176*, Page no. 251-260. https://doi.org/10.1016/j.procs.2020.08.027

Badshah, A.M., Rahim, N., Ullah, N., (2019). Deep features-based speech emotion recognition for smart effective services. *Multimed Tools Appl*, 78**,** 5571–5589. doi:10.1007/s11042-017-5292-7

C.Huang, W.Gong, W.Fu, and D.Feng. (2015) *Research of Speech Emotion Recognition Based on Deep Belief Network and SVM, Mathematical Problems in Engineering*. Vol. 2017, Article ID 749604, 7 pages. https://rg/10.115doi.o5/2014/749604

Demicran, S., Kahramanli, H. (2014). Feature extraction from Speech Data for Emotion Recognition, *Journal of Advances in Computers Networks , Volume no.2*, Issue no.1, Page no. 28. Doi: 10.7763/JACN.2014.V2.76.

Kerkeni L., Serrestou Y., Mbarki M., Raoof K. and Mahjoub M. (2018). Speech Emotion Recognition: Methods and Cases Study.In*Proceedings of the 10th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, ISBN 978-989-758-275-2, pages 175-182. DOI: 10.5220/0006611601750182.

Moore, S. (2018). 13 Surprising Uses For Emotion Ai Technology, gartner.com,https://www.gartner.com/smarterwithgartner/13-surprising-uses-for-emotion ai-technology

R.A. khalil, E. Jones, M.I., Babar, T., Jan, M.H., Zafar, T., Alhussain, (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access, Volume 7,* Page no. 117327-117345.10.1109/ACCESS.2019.2936124

Woo, B., S., Seok-Pil, L. (2021). A Study on Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms, *Applied Sciences, Volume no. 11,* Issue no. 1890. https://doi.org/10.3390/app110418980

Xiaobo, B. (2016). Application of Speech Recognition Technology in Speech-Related Disabilities*:* An Analysis and Forecast, *HMB431*, Medium (Access date 23 May 2023). https://medium.com/@TEAM.BULLS/application-of-speech-recognition-technology-in-speech-related-disabilities-an-analysis-and-5b102da57f86

Z. Khan, F., R. Alotaibi, S. (2020). Applications of AI and Big Data Analytics in Health, *College of Computing and Information Technology, Volume no 2020,* Article Id.8894694., https://doi.org/10.1155/2020/8894694

ADDITIONAL READINGS

1.A Review of Generalizable Transfer Learning in Automatic EmotionRecognition.

doi:10.3389/fcomp.2020.00009

2.A Waveform-Feature Dual Branch Acoustic Embedding Network for EmotionRecognition.

doi:10.3389/fcomp.2020.00009

3. Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction andCompanding.

doi:10.3389/fcomp.2020.00014

4. Speech Emotion Recognition with deeplearning.

URL:https://doi.org/10.1016/j.procs.2020.08.027

5. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, andclassifiers

URL:https://doi.org/10.1016/j.specom.2019.12.001

6. A Research of Speech Emotion Recognition Based on Deep Belief Network andSVM.

URL:https://doi.org/10.1155/2014/749604

7. Speech Emotion Recognition: Methods and CaseStudy.

URL:https://dx.doi.org/10.5220/0006611601750182

8. A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep LearningAlgorithms

URL:https://doi.org/10.3390/app11041890

9. Speech Emotion Recognition algorithm based on deep learning algorithm fusion of temporal and spatialfeatures

doi:10.1088/1742-6596/1861/1/012064

Key Terms and Definitions

1. Neural Networks - A neural network is a series of algorithms that endeavours to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates.

2. Multilayer Perceptron - A multilayer perceptron (MLP) is a feedforward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and outputlayers.

3. Perceptual Linear Prediction cepstral coefficients - A new technique for the analysis of speech, the perceptual linear predictive (PLP) technique, is presented and examined. PLP analysis is computationally efficient and yields a low-dimensional representation ofspeech

4. DBNs - In machine learning, a deep belief network is a generative graphical model, or alternatively a class of deep neural network, composed of multiple layers of latent variables, with connections between the layers but not between units within eachlayer.

5. CNNs and RNNs - The ability to process temporal information data that comes in sequences, such as a sentence. Recurrent neural networks are designed for this very purpose, while convolutional neural networks are incapable of effectively interpreting temporalinformation

6. SAVEE - It is an emotion recognition dataset. It consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total. The sentences were chosen from the standard TIMIT corpus and phonetically-balanced for eachemotion**.**

7. Mel spectrogram - It logarithmically renders frequencies above a certain threshold (the corner frequency). For example, in the linearly scaled spectrogram, the vertical space between 1,000 and 2,000Hz is half of the vertical space between 2,000Hz and4,000Hz.