BMS COLLEGE OF ENGINEERING

(Autonomous Institute, Affiliated to VTU)
Bull Temple Road, Basavanagudi, Bengaluru - 560019

A project report on

"ANALYSIS OF SPEECH PATTERNS IN CHILDREN FOR DETECTING POTENTIAL DEPRESSION"

Submitted in partial fulfillment of the requirements for the award of degree

BACHELOR OF ENGINEERING IN

INFORMATION SCIENCE AND ENGINEERING

By

RASHMI V BHAT – 1BM16IS071

SANJANA M MOODBAGIL – 1BM16IS077

SUPREETA ANAND BYATNAL – 1BM16IS090

Under the guidance of

NALINI M K ASSISTANT

PROFESSOR

Department of Information Science and Engineering

2019-2020

BMS COLLEGE OF ENGINEERING

(Autonomous Institute, Affiliated to VTU) Bull Temple Road, Basavanagudi, Bengaluru – 560019

Department of Information Science and Engineering

C E R T I F I C A T E

This is to certify that the project entitled **“ANALYSIS OF SPEECH PATTERNS IN CHILDREN FOR DETECTING POTENTIAL DEPRESSION**” is a bona-fide work carried out by **Rashmi V Bhat (1BM16IS071), Sanjana M Moodbagil (1BM16SI077), Supreeta Anand Byatnal (1BM16IS090)** in partial fulfillment for the award of degree of Bachelor of Engineering in **Information Science and Engineering** from **Visvesvaraya Technological University, Belgaum** during the year **2019-2020**. It is certified that all corrections/suggestions indicated for Internal Assessments have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the Bachelor of Engineering Degree.

Nalini M K	**Dr. M Dakshayini**	**Dr. B. V Ravishankar**
Assistant Professor	**Professor and HOD**	**Principal**

Examiners

Name of the Examiner	**Signature of the Examiner**
1.	
2.	

II. ABSTRACT

Anxiety and Depression in children can emerge at a very early stage and often go undiagnosed. Symptoms are frequently disregarded until the time children can communicate their discomfort easily. Whenever left untreated, it is regarded as internalizing disorders, which result in extended periods of adverse outcomes, including substance abuse and suicidal thoughts. These prolonged childhood depressions can quickly turn to teen depression. Hence, to avoid internalizing disorders, the detection of depression, and anxiety at an early stage is necessary. A design of dynamic relationship between characteristics of voice audio signals and diagnosis needs to be developed for identifying kids with internalizing illness. For this task, an approach which is driven by data, such as machine learning, is best performed by analyzing the participant 's reaction to a speech challenge, a behavioral activity designed to evoke anxiety. In a speech task, we use the recording of children's speech for analysis and ML to identify and diagnose clinically induced symptoms for anxiety and depression in children aged 3 to 7 years. To classify young children with mental illness, investigation of the speech data from mood enhancement tasks with machine learning is done. The more discriminative symptoms of internalizing behavior was analyzed in detail, showing that children suffering from depression exhibit unusually low pitch voices, with clear speech inflections and high- pitched reactions to unwanted stimulus. For this task, we analyze the performance of several statistical models, investigate how the low quality data can impact the performance of the model, and after comparison, the model that performs the best can be used to detect depression among children.

III. ACKNOWLEDGEMENT

We would like to express our deepest gratitude to our final year project guide, Mrs. Nalini M K for making it possible for us to finish this project on time, whose commitment, encouragement and motivation has helped us organize our work. We acknowledge with much appreciation her crucial role in helping us from the beginning, motivating us towards this topic, guiding us with full effort to accomplish all our objectives.

The final year project has been prepared and submitted as a part of the course, ‘PROJECT WORK’. We would also like to express our sincere gratitude towards the project work coordinators, Mrs. Nalini M K and Mrs. Sindhu K for their support and active participation. This project could not have been completed without their constant motivation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

Mental health is an essential aspect of the overall health of a child. It has a very tightly wound relationship with the physical health and ability to perform their best at school, work, and develop an overall healthy emotional outlook on life. It is easy for parents to identify the basic needs of childlike clothes that are warm when it is cold, food that is nutritious and good for their health, sleep to get energized. However, it may not be obvious to recognize the emotional needs of a child. Hence it becomes imperative to identify how a child is doing in terms of his/her mental health as good mental health is vital for the child to develop new skills and grow in the society. Depression is more than just typical blues or sadness that a child faces. It affects the ability to function normally in day to day life. Anxiety and Depression can be seen in a child of age as young as four years. Due to the inability to express their emotions clearly at this very young age, the symptoms often get overlooked. Studies show that about 2 percent of the children of preschool and school-age children are affected by depression. There are various types of symptoms that the child tends to exhibit, some of which include having an irritable mood, anger, loss of interest, physical body aches, and many more clinical signs. Just as in adults, various factors contribute to the risk of developing depression in children, some of which are as below.

- Physical health – Physical health and mental health are directly related. Children who suffer from chronic diseases are most likely to be depressed.
- Surroundings – A stressful home life or a stressful peer group contribute to the risk of developing depression.

- Family history – If a parent or any of the family members suffer from depression, it is most likely for the child to develop depression at a very young age.

Identifying that a child is developing the symptoms of depression is harder than it sounds. From the psychological viewpoint, it is said that one of the apparent signs of depression is the way a human expresses their emotions in their speech, assuming that depression often affects the acoustic qualities of the speech. We can perform a computer based, automated analysis of the speech to ease the process, help the clinicians to efficiently diagnose and hence treat the illness at a much earlier stage. Deep Learning methods are often seen to give fruitful results when it comes to the analysis of speech signals due to their unstructured nature. Therefore, in our proposed model, we present a machine learning tool that can objectively screen children for anxiety and depression, which is a valuable asset for early childhood intervention to prevent future problems of mental health issues.

1.2 MOTIVATION

Depression is a severe medical illness that affects the way we think, feel, and act. For this task, we analyze the performance of several statistical models, investigate how the low quality of data affects the performance of the model and hence compare the performances and use the best available model to test and diagnose depression in children . Parents are often unaware of symptoms of depression and anxiety at early stages. More than 20 percent of kids, however, tend to develop an internal disorder during childhood. This influences the growth and development of the child and predicts, if left unchecked, severe health problems later in life. Thus, there are social burdens which require the need for early detection of such disorders in the mental health of a child.

1.3 OBJECTIVE

To identify children with depression and anxiety disorders accurately using a three-minute speech. To reduce the waiting time at hospitals, to improve the accuracy of detecting depression in young children using machine learning approaches, and to build a fast, automated and low-cost software to detect depression in young children.

1.4 SCOPE

With the NMHP and its implementing arm, the District Mental Health Program (DMHP), more room is available for improving and scaling depression treatment through centered and committed interventions. Testing and treatment for mental illness can be applied at different levels of health care, occupational and educational settings using standardized methods and common protocols. Accessibility to depression treatment can be enhanced by providing services and ensuring continuous supply of basic drugs at primary health care (both urban and rural) settings.

1.5 DOMAIN KNOWLEDGE

1.5.1 MACHINE LEARNING:

Machine learning is the process through which a machine learns from past experiences or rather existing data/patterns, so that it can predict what could be output based on certain inputs. Although focus is usually put on selecting the most effective learning algorithm, developers find that few of the most important problems emerge from none of the equally performing machine learning algorithms available. Mostly it is a concern with training results, but this often occurs when dealing in different areas of machine learning.

Work performed while operating on actual projects also fuels development in the field and there are two reasons for this: firstly, Trend to explore strengths and shortcomings of current approaches. secondly, Researchers and engineers who collaborate with specialists in the area and use time and experience to boost device performance.

Often this occurs by accident. One example may be model sets, or variations of several learning algorithms to increase accuracy. Teams vying for the 2009 Netflix Price noticed that they had their strongest performance by mixing their learners with learners from other teams, contributing to an enhanced selection algorithm.

1.5.2 DEPRESSION IN CHILDREN:

Depression is a mood disorder with low mood, loss of energy, fatigue, helplessness, lack of hope, guilt, decreased sleep and appetite. It has both physical and emotional symptoms.is the most prevalent question in mental wellbeing, not just among adults, but also among children and adolescents. The World Health Organisation (WHO) study 'Health for World Youth' identifies depression as the leading cause of illness and impairment for children and teenagers between the ages of 10 and 19, and homicide is the main cause of death after road accidents and HIV/AIDS.

This study body is now recognized, but autism and other mental health problems in children were almost unknown a few decades earlier. Depression in children always has been a difficult subject for society as well as for healthcare professionals to understand. Children who exhibited signs that are now recognized as depression were thought to have behavioral problems out of which they eventually would grow. There are still some misconceptions about this: that kids are too early to be depressed; that true depression happens only in adulthood; and that the grief can develop only for certain phases which they can come out of really easily.

Depression in childhood is real and, as a parent, they can encourage their child to understand it and take care of it themselves. Depression has both a psychological and behavioral origin. Children with a family history of mental illness are at greater risk of themselves experiencing depression. Depression may arise when a child becomes physically fragile and faces family or social tension, too. But it appears different from an indication of distress among children: the child may appear irritable, or show physical problems. This attitude can be spotted at home and at school; at home, the child may not be interested in their hobbies, would not want to play with their friends or a pet, would not want to communicate with parents or siblings. The child's work may suffer at school, the child may skip school frequently or display disciplinary problems.

A problem that sometimes pops up is that all troubled children require antidepressants. The answer is quite complicated. Without medication and psychotherapy sessions that equip them with better coping strategies, a child with mild symptoms and identifiable stressors can recover. Medication may be required in moderate and severe cases. Unfortunately, only one out of five children with depression seek and receive help. Untreated depression in childhood raises the likelihood of depression in adulthood. The WHO estimates that depression would impact more individuals than any other health condition by 2030. It is important for them to discuss this troubling figure and inform parents about the misconceptions and reality of childhood depression.

It isn't always easy to communicate with depressed children. But if a parent sees differences in the actions of his or her child correlated with depression, a reasonable approach to continue the discussion will be to let the child know that you have seen improvements and that you are ready to respond to everything they may have to tell. Thinking regarding depression may provide a psychological advantage in itself and although the child does not instantly open up, they can gradually wake up to it.

1.6 EXISTING SYSTEM

The existing system implemented machine learning algorithms on child speech data set to compare the accuracy between high-quality and low-quality data. They used a 3-minute speech task that employed Trier-Social Stress Task along with ML to diagnose anxiety and depression. They employed SVM-Linear model which gave 80% and 53% accuracy with high quality and low quality data respectively.

1.7 PROPOSED SYSTEM

The system proposed in this project is machine learning based that employs speech analysis of a child's voice and uses four classification models to detect children with anxiety and depression. An application is created which processes any new data and classifies them as depressed or normal. Any new speeches can be recorded in the application itself and simultaneously used for prediction. The model is trained using four models that gives better accuracy than the existing system. This approach takes less computation time and is cost effective. Three among the four models gives better accuracy than the existing system.

We have also proposed a transfer learning model where a pre-trained WaveNet model is fit with CNN model and compared this with a CNN model using MFCC extracted features.

CHAPTER 2

LITERATURE SURVEY

Often, depression goes unnoticed in children. In our paper, we intend to diagnose depression in children by analyzing their speech. In [1], the author has laid down that the acoustic properties of speech can be taken as possible cues of depression as there is enough evidence to establish the same. Thus, specific vocal parameters can be used to classify depressed speech.

Mental health can be analyzed with the help of the vocal attributes of a person [2]. The vocal effect describes the mental and emotional state in which the speaker is and the overall condition

Expression of emotions is a way of communicating in daily life but mental/emotional disorders like Depression or Anxiety would be classified as potentially crushing illnesses. The author [2] has further enunciated the importance of inclusion glottal features in voice analysis and also analyzed it. The voice track produces a range of format frequencies (FMT) such as F0, F1, F2, and so on based on the speaker's enunciation of voice.

For voice analysis, the spectral information is considered. The Voice activity detector (VAD) proves to be a great tool in extracting voice signals from the input audio.[3] Segment selection from the voice spectrum where the active audio regions that provide maximum discrimination between neutral and depressing speech. Feature extraction also plays a critical role in the diagnosis of depressed speech. Features such as cues, energy, and pitch would be customarily considered. The Mel Frequency Cepstral Coefficients (MFCCs), the Spectral Centroid Frequencies and Amplitudes (SCF / SCA), delay in the linear predictive group give us the detailed spectral information.

Today, clinal depression diagnosis requires structural interviews and standardized questionnaires [4]. Hence there is a scope to automate the whole process, which also proves to be efficient. Data can be extracted with the help of wearable sensors that sense vocal reactions from children to various tasks [4]. Signal features are derived from the voice signal which includes mean, root mean square (RMS), skew, kurtosis, range, maximum, minimum, standard deviation, peak to RMS amplitude, signal power within frequency bands 0-0.5 Hz, 0.5- 1.5 Hz, 1.5-5 Hz, 5-10 Hz, 10-15 Hz, 15-20 Hz, all frequencies higher than 20 Hz, and features extracted from the autocorrelation of the signal. This comprises a total of 29 features from each of the six-time series. Various machine learning models such as support vector machine with a linear kernel (SVM), decision tree (DT), k-nearest neighbor (kNN), and logistic regression (LR) models are proposed for the diagnosis.

A similar approach is followed by the author [5] for diagnosis where a 3-minute speech task, Trier Social Stress Task for children (TSST-C), and machine learning are employed from children of ages 3 to 7 years old. Audio data is processed via VAD that discriminates audio presence and absence in the voice signal. Within each phase, parameters of audio signals were computed for each speech epoch. Features included speech epoch duration, mean frequency, zero-crossing rate (ZCR) of the audio signal, Mel frequency cepstral coefficients i.e (MFCC), dominant frequency, perceptual spectral centroid (PSC), spectral flatness, skew and kurtosis of the power spectral density (PSD), ZCR of the z-score of the PSD (ZCR zPSD) for all speech epochs, the first, second, and third formants, and the percentage of signal energy above 200, 500, 700, 1000, and 2000 Hz. Binary classification models such models logistic regression (LR), support vector machine with a linear kernel (SL), support vector machine with a gaussian kernel (SG), and random forest (RF) are proposed to classify depressed and neutral speech.

Additionally, with a booming trend in deep learning models, many such models have been successfully implemented in the area of speech analysis [6]. The author's proposed model is to build Multiple convoluted and recurrent layers that provide an integrated model for speech analysis for both acoustic features and also a language-based model. A high-level feature extraction approach using NMF (Non-Negative Matrix Factorization) is combined with multilingual deep neural network (DNN) training, and a CTC model is taken into consideration for speech analysis. A transfer learning approach is proposed.

Depression is sort of a soft trait biometrically which will play a vital role in different legal issues while helping in individual characterization. Since the diagnosis and detection can be a difficult task normally, there is a lot of interest being shown in automating it using various features like voice, eyes, behavior etc. Hence it is important to know which is seen to be performing better compared to the others.

In the paper [7], different features of the voice were extracted and analyzed. A two level fusion technique involving RASTA-PLP (Relative Spectral Transform - Perceptual Linear Prediction) was seen to be performing better by extracting the better quality of speech by reducing the variations in the speakers' speeches and by not loosing out on important information when compared to the normally used MFCC features. In the experiment it was seen that this method worked perfectly for training dataset but not quite for testing dataset.

There can be a lot of variations in the vocals of a person be it a child or an adult based on the severity of the depression. Vocal changes contribute a lot towards the overall change in the personality of a human. There are various factors like pitch, frequency, loudness of the voice etc. that contribute to the vocals of a person.

In paper [8], an experiment was conducted with 57 participants as samples who were of different age groups ranging from 19 to 65 years and were suffering from depression at various levels. Through this experiment, it was seen that changes in the vocal prosody could determine how severe their depression was. SP mean and variability and mean and variability of F0 were considered the measures here. Hence it was proved through this experiment that change in fluctuation and rhythm of the voice/ voice prosody is a very common symptom of depression. Therefore, by using this information it is possible for researchers to come up with different techniques that use audio and voice clips to detect and diagnose depression in early stages.

It is seen from a study conducted by WHO that about 12.8% of the people in the world suffer from Clinical depression. Hence it is very important to consider depression as a critical problem and treat it in the earlier stages just like any physical diseases are treated. A multimodal framework for detecting and estimating depression is proposed in paper [9].

In paper [9], the research done by the author mainly focuses on the three aspects :

1. To detect depression different features and techniques that are multi modal are explored.
2. As there is not much data to train the deep learning models about depression, this has to be managed.
3. Dimensional Affective analysis applied for estimating depression.

If the classification and estimation are considered and performed hand in hand, the outcomes are seen to be better. Text format can be used to classify where as voice and video will help in estimation framework. DCGAN-based data generation methods seemed to increase the performance during estimation.

Since there is a lot of effort that goes in if there it is required to clinically test for depression among the patients, a lot of research is being conducted all over the world to use artificial intelligence and deep learning algorithms to automate this diagnosis and make it easier.

One of the main symptoms of depression was seen to be frequent variation in the moods.

In the approach proposed in the paper [10], Depression AudioNet was the framework used to extract different features like MADN and MFCC. The authors also proposed that using joint tuning layers by making use of information related to contextual emotions will help increase the performance. The outcomes for MAE and RMSCE were seen to be 7.07 and 9.15 respectively.

In [11] Sahar Harati et al. proposes a model that is predictive in nature on the features that are based on emotions. The comparision is provided between the models like Support Vector Machine, K- Nearest Neighbor, Hidden Markov Model with Gaussian Mixture Emissions and Grated Recurrent Unit that is based on RNN model for features extracted using different methods like Basic feature set, Switching Linear Dynamic System and the Emotional features. The architecture proposed in [11] uses the Long Short Term Memory with 2 hidden layers. The primary outcomes of this approach shows that the model that is proposed efficiently categorizes "depressed" and "improved" stages of Deep Brain Simulation Treatment with the Area Under the Curve being 0.8.

In [12], a one dimensional CNN and Transformer methods are proposed by Genevieve Lam et al. to automate the detection of depression using data augmentation that is context aware and makes use of acoustic features. This method is based on topic modelling and it was observed that it was efficient for audio and text forms. One dimensional CNN and Transformer model separately achieved great F1 scores but when they were combined, they outperformed the existing performance and showed strong outcome of F1 Score 0.87

In [13], Ziping Zhao et al. proposes an approach which has two steps. One is to recognize the speech and the second is to detect depression from the recognized speech. The fundamental component here is the Mechanism of Attention Transfer which is made use of, for the detection of the depression in the second step. A HATN network uses hierarchical attention autoencoders to learn in the unsupervised way. BLTSM is used to process each frames individually and one by one sequentially. It can be seen that the results produced by this hybrid model are the best-known as of date.

In [14] Barkha Rani compares the performances of I Vector and Fuzzy membership on the audio signs to detect depression level. The dataset of around 20 samples is considered in which 10 are considered as training data and 10 as test data. This data is preprocessed for better performance. The f measure increases for every audio sample from the existing system. The analysis of the graphs show that Fuzzy membership method performs better than I-Vector method in terms of accuracy and performance. Hence applying Fuzzy Membership to the real time data will give better outcomes.

In [15], Zhaocheng Huang et al. proposes a method where a series of vocal events is analyzed as a speech and is used to detect depression. Natural Language Processing methods are used to tokenize this speech as words at different intervals of time. This proposed framework explores the information with the combination of landmark and vocal words. The tokens are very compact and help preserving privacy and hence can be exchanged between client and server easily. For both the data sets considered in this, landmark dependent words are proven to be the most efficient for heterogenous type of features. In this newly proposed framework SVM Linear model outperforms the existing methods.

CHAPTER 3

REQUIREMENT ANALYSIS

3.1 FUNCTIONAL REQUIREMENTS

Some of the functional requirements for the child depression application are:

1. Create a desktop software.
2. Load the voice as input.
3. System should preprocess and extract the deep features from the voice frames.
4. By applying the algorithm, the system should classify the given voice as depressed or normal.
5. Application should provide high accuracy of classification of child voice.

3.2 NON-FUNCTIONAL REQUIREMENTS

They define some of the attributes of the system that serve as restrictions or constraints on the system's design. There are certain criterias that are used to decide the system's operation and performance. Some of the examples are reliability, security etc.

Constraints within which our system works:

· Self-contained application / portable – It can be installed in different computers and can be moved from one system to the other, with the assumption that the connection of the network exists while application is used.

· Availability and Scalability – The system is available 100 percent of the time and scalable to support and service additional clients is necessary.

- Reliability – This system contains no glitches. It is reliable all the time.

- Maintainability - Ease of maintenance of the application is achieved.

3.3 HARDWARE REQUIREMENTS

● System	:	Intel i3 or more
● Hard Disk	:	120 GB.
● Monitor	:	13’’ LED
● Input Devices	:	Mouse, Keyboard
● RAM	:	4 GB

3.4 SOFTWARE REQUIREMENTS

● Operating system (OS)	:	Windows 7,8,10
● Language used for coding	:	Python 3.6
● Tools	:	Python IDE

CHAPTER 4

DESIGN

4.1 SYSTEM DESIGN

4.1.1 SYSTEM ARCHITECTURE

In the proposed system, the child dataset is collected using VAD, which detects the presence of human speech. The data undergoes various audio processing techniques such as noise cancellation, blank audio duration, etc. followed by feature extraction. Features such as MFCCs, Chrome Feature, Energy, Jitter, etc. are extracted from the audio. Machine Learning algorithms are trained and used to classify the children as depressed or mentally healthy. The below figure explains the workflow of the proposed system.

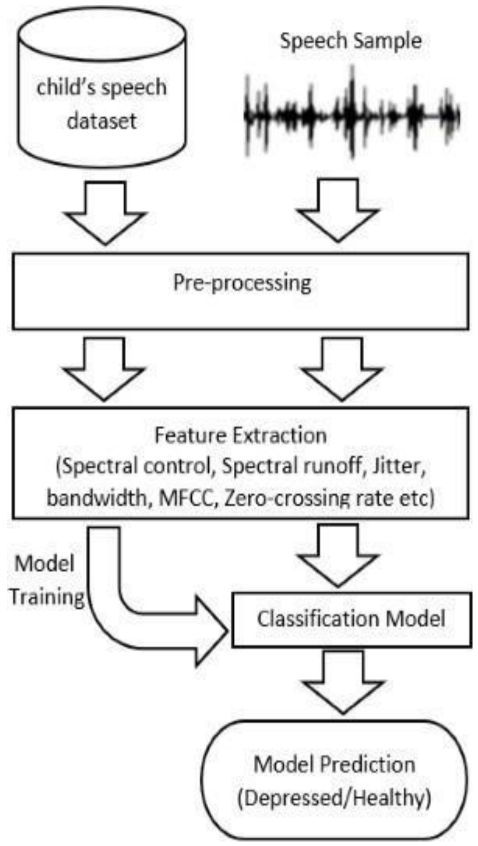


Fig. 4.1 System architecture

4.1. 2 MODULE DESIGN

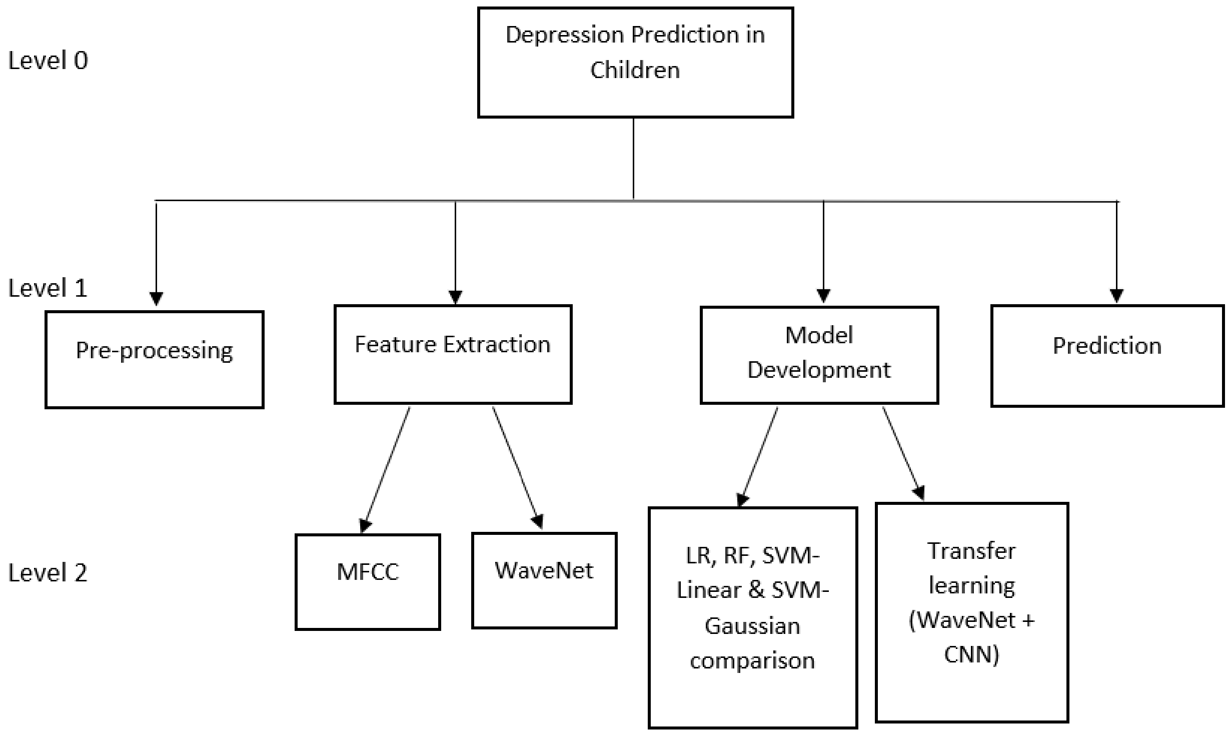


Fig. 4.2 Module Design

4.2 DETAILED DESIGN

4.2.1 DATA FLOW DIAGRAM

Level: 0

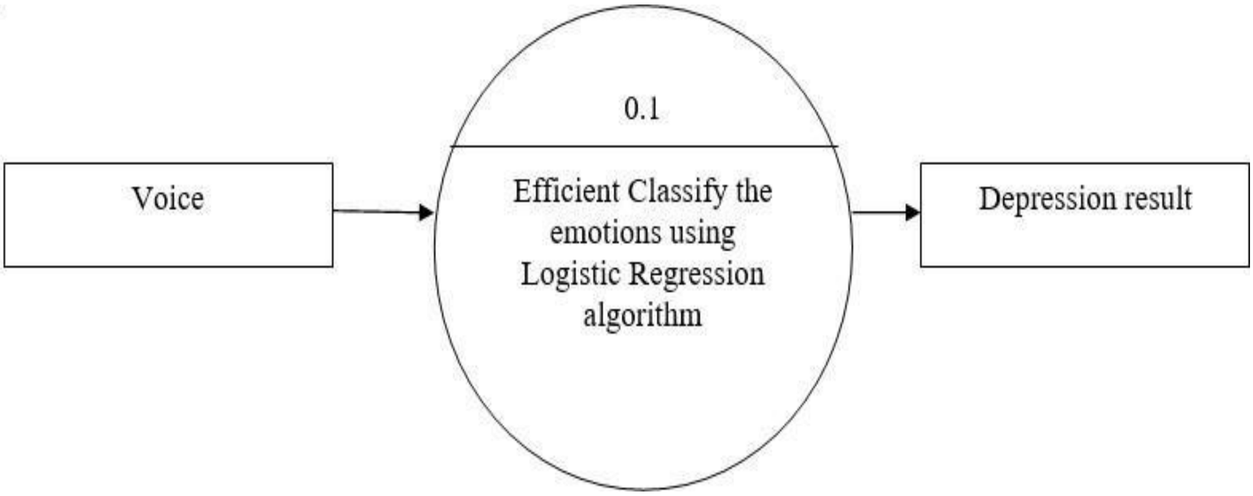


Fig. 4.3 Data Flow Diagram Level 0

Level 0 Describes the overall process of this project. we are passing voice as an input the system will classify the voice using logistic regression to determine the person is in depression or not.

Level: 1

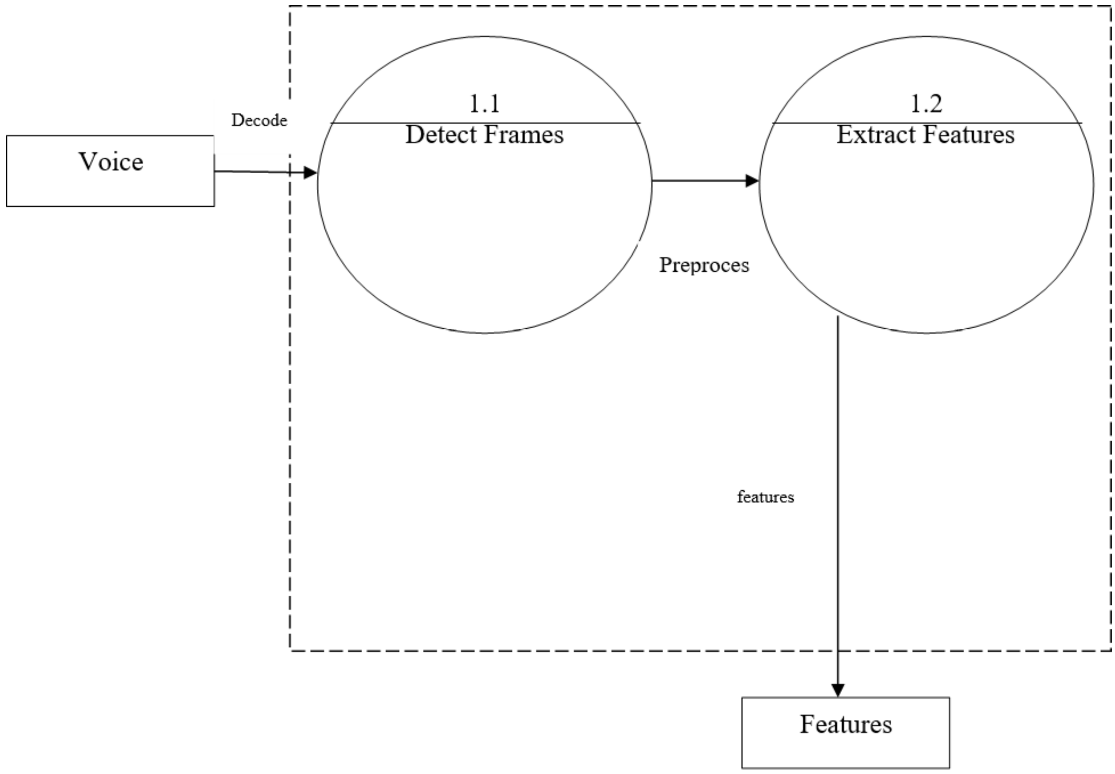


Fig. 4.4 Data Flow Diagram Level 1

Level 1 Describes the first stage process of this project. we are passing voice as a input the system will detect the frames in voice then extract the features (MFCC, Spatial frequency, etc.,).

Level: 2

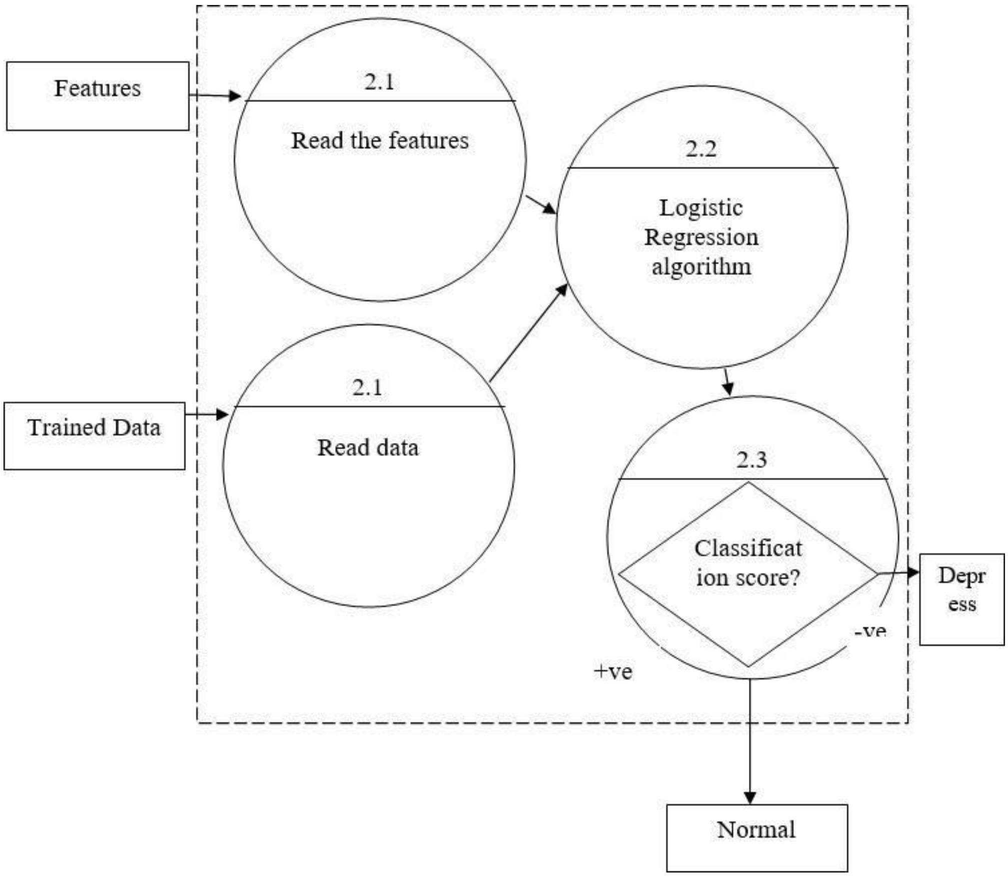


Fig. 4.5 Data Flow Diagram Level 2

Level 2 Describes the final stage process of this project. We are passing extracted features from level 1and trained data as an input the system will classify the given voice is normal or depressed using logistic regression algorithm.

4.2.2 CLASS DIAGRAM

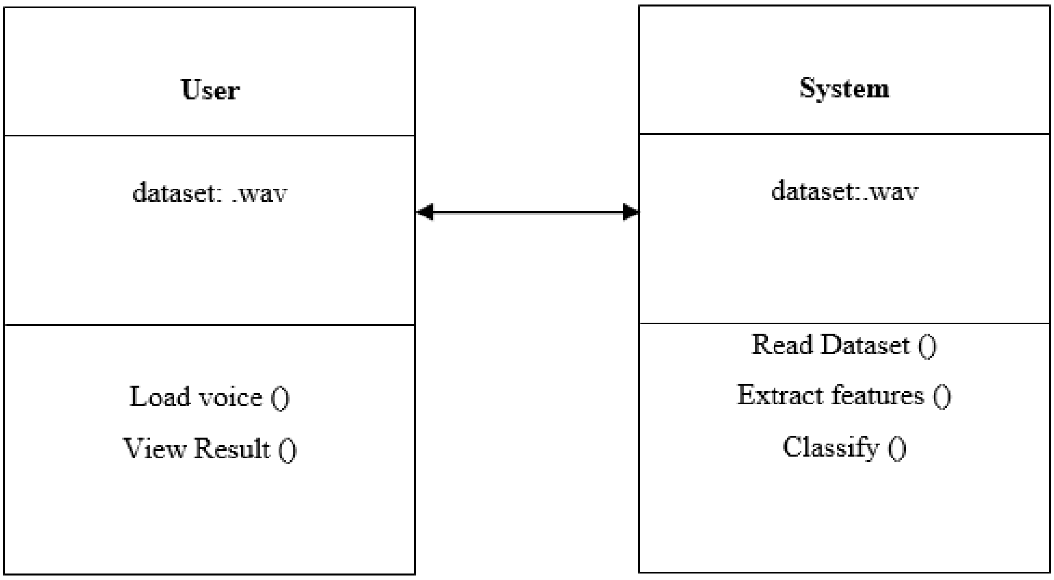


Fig. 4.6 Class Diagram

4.2.3 SEQUENCE DIAGRAM

The sequence diagram will specify the users states (active/ inactive) with our application.

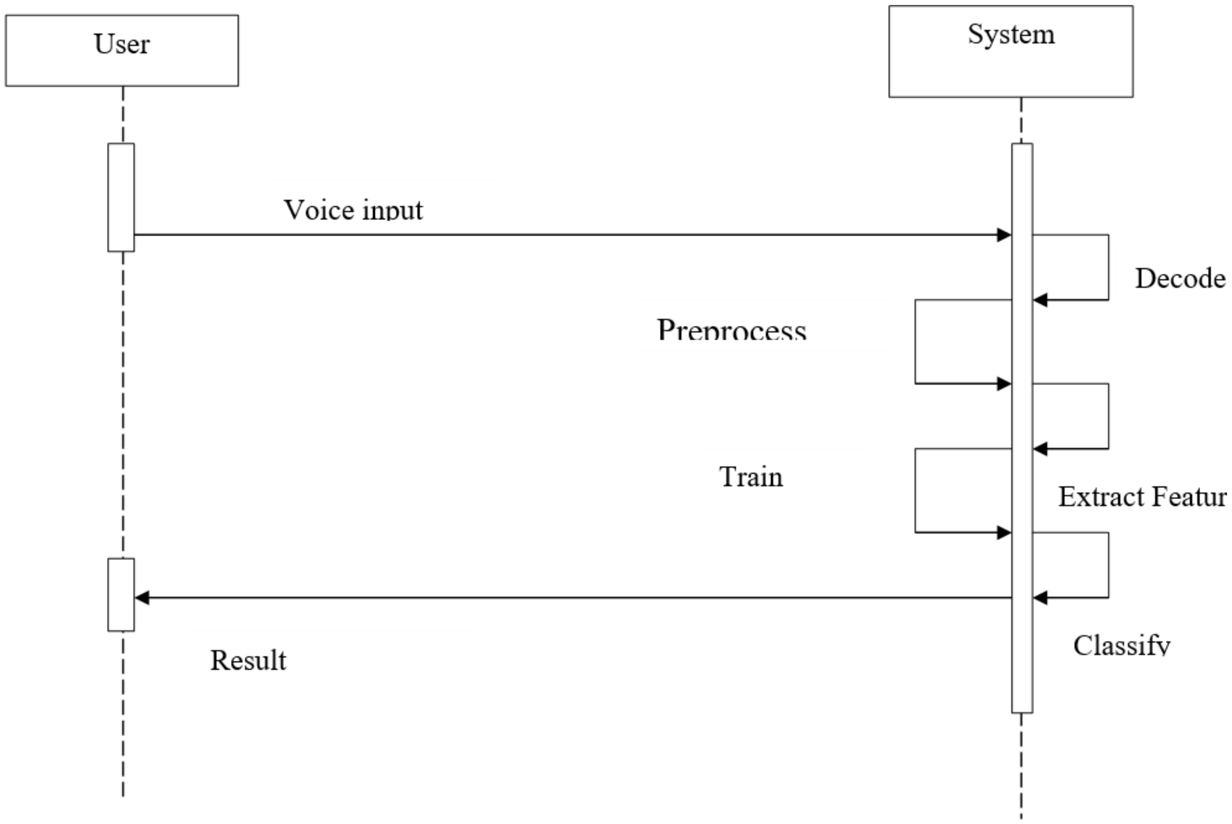


Fig. 4.7 Sequence Diagram

4.2.4 USE CASE DIAGRAM

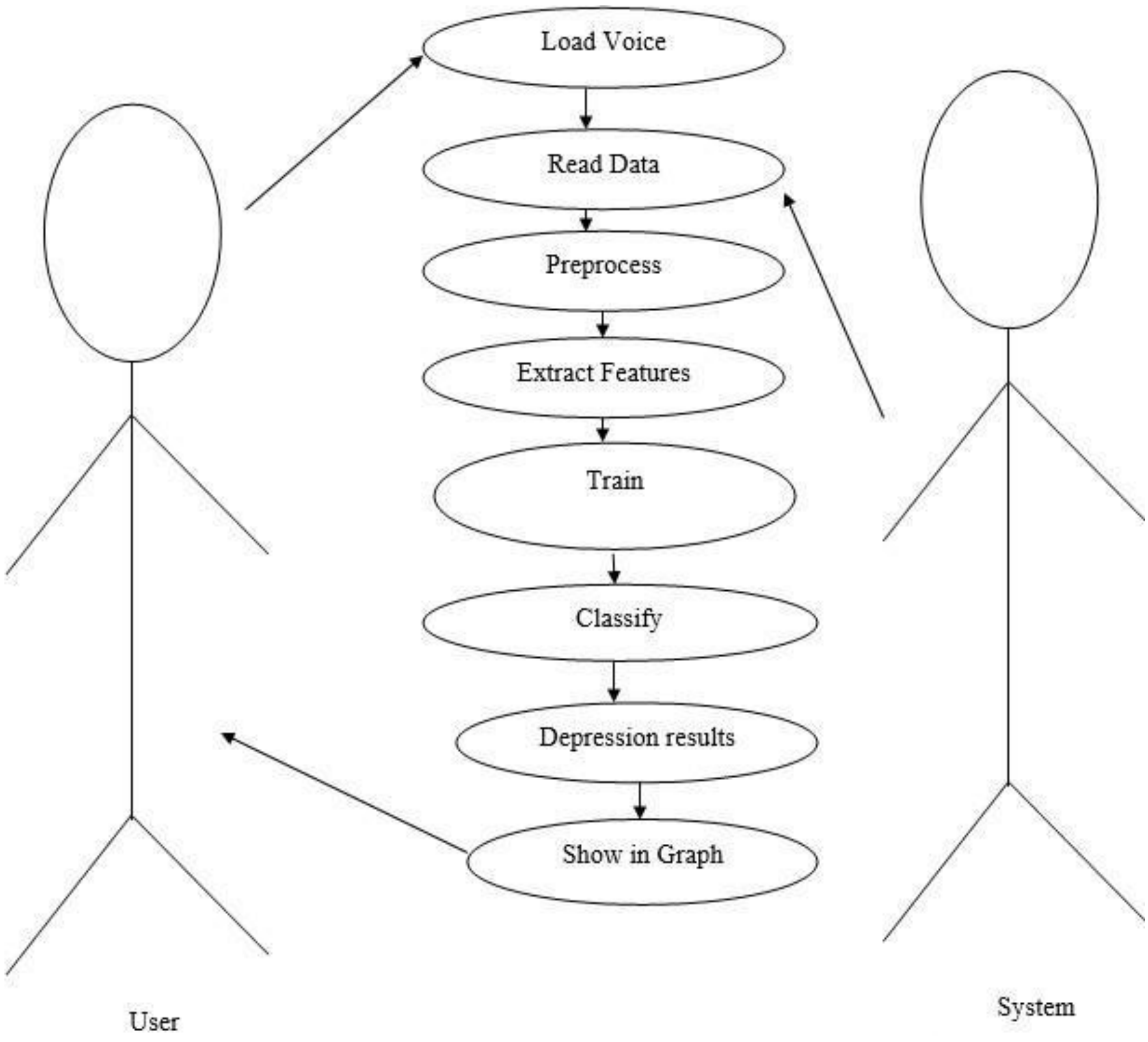


Fig. 4.8 Use case diagram

The use case diagram will specify the user interactions with our applications.

4.2.5 SCENARIOS

Pre-processing the data set to extract features and obtaining the results of the classification models.

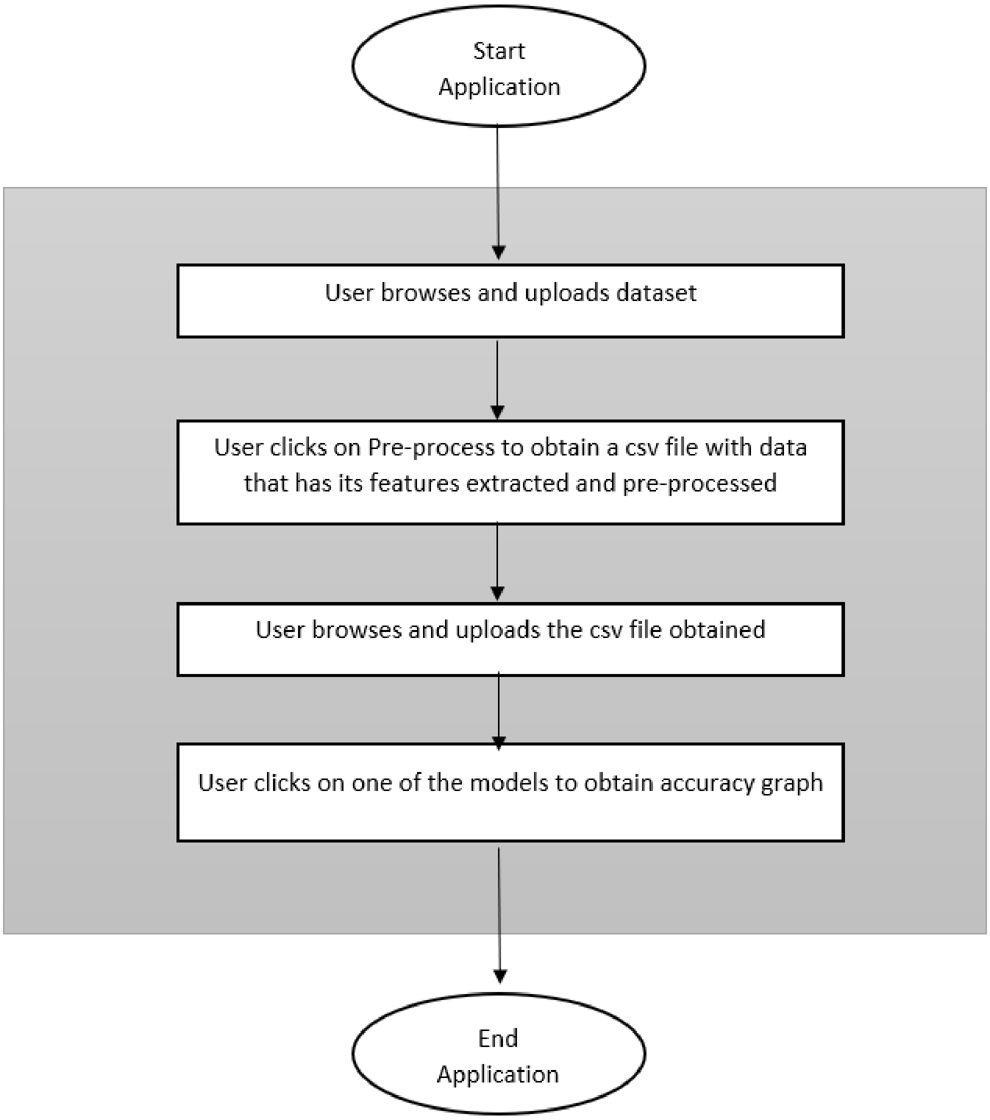


Fig. 4.9 Scenario 1 diagram

Predicting the classification output (Depressed/ Normal) on a newly recorded data.

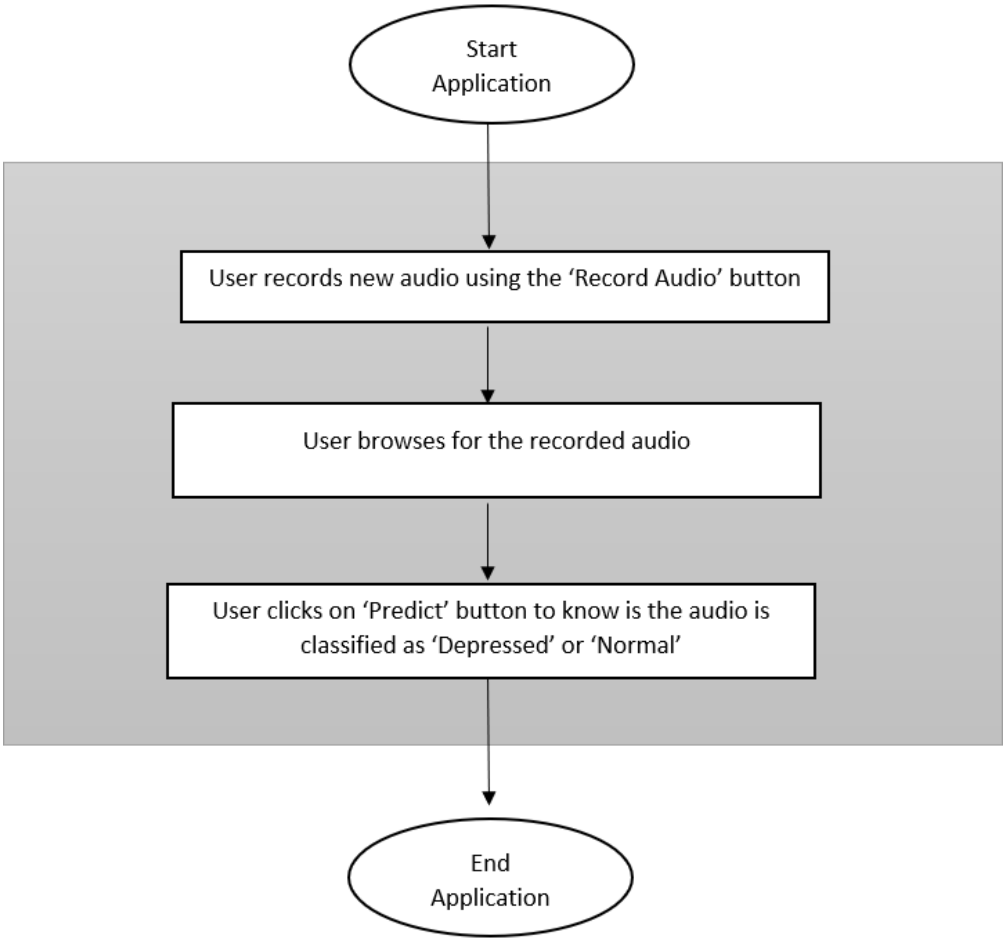


Fig 4.10 Scenario 2 diagram

Predicting the classification output (Depressed/ Normal) on an existing test data.

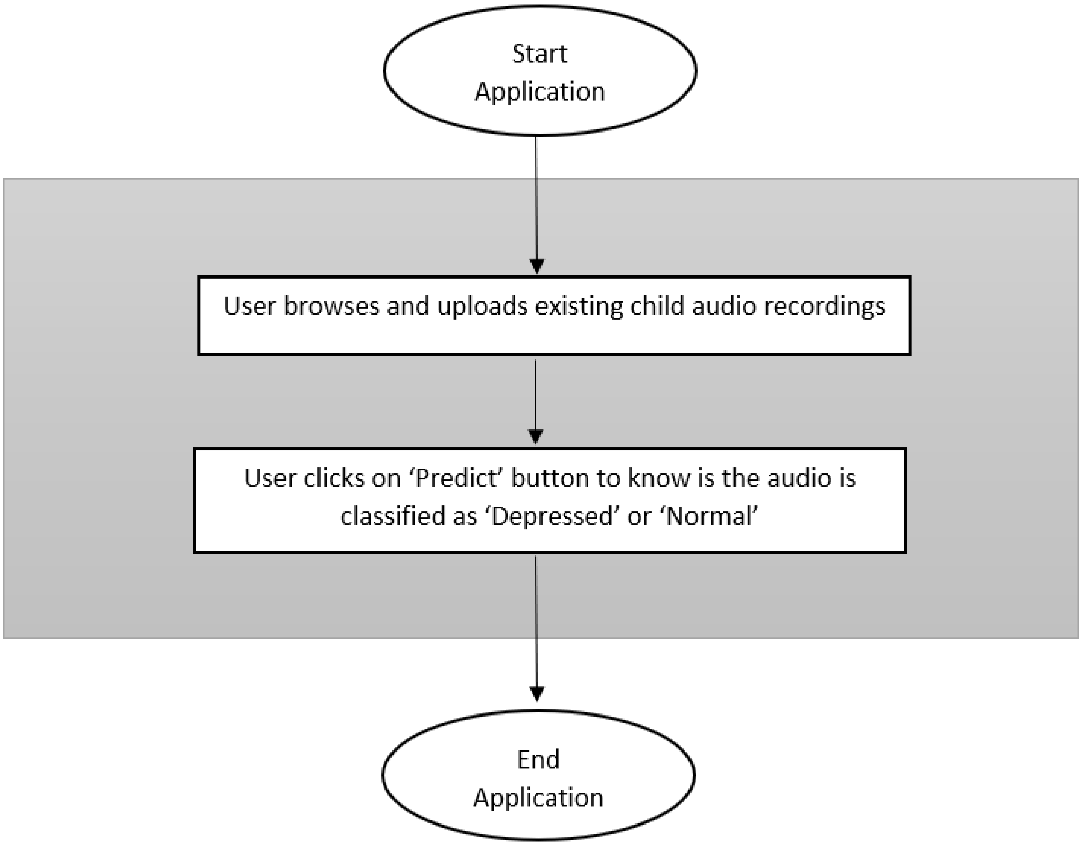


Fig. 4.11 Scenario 3 diagram

4.2.6 LOW LEVEL DESIGN

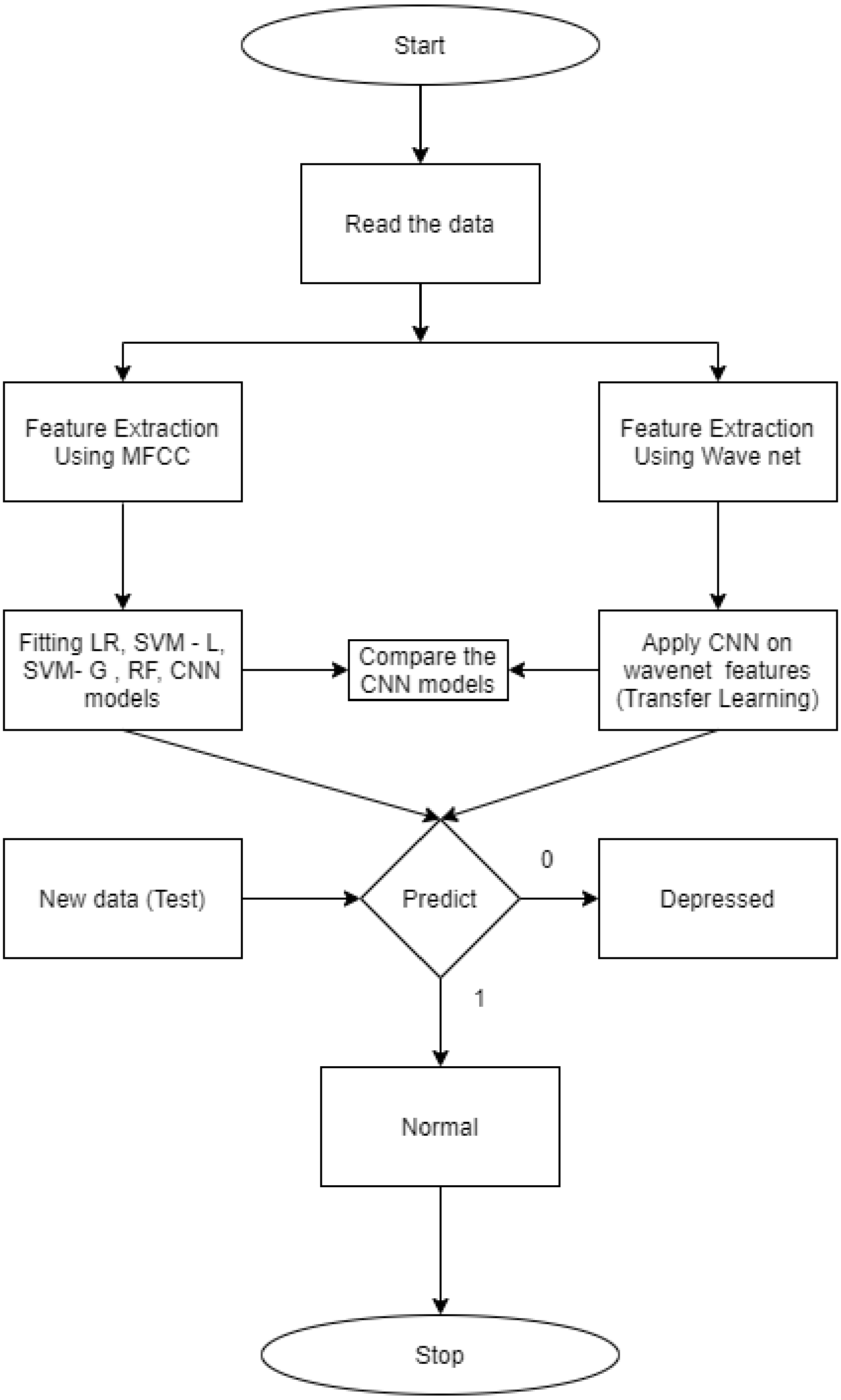


Fig 4.12 Low Level Design

CHAPTER 5

IMPLEMENTATION

5.1 MODULE 1: AUDIO DATA PREPROCESSING

5.1.1. AUDIO ANALYSIS:

Audio data analysis: It is analyzing and recognise digital device-captured audio signals, with numerous business, healthcare, productivity, and smart city applications. Applications provide consumer service reporting from customer care requests, reviewing and extracting media information, medical testing assistance and patient tracking, aid devices for individuals with hearing impairments, and public safety audio reporting.

Sample is defined as the magnitude of waves at a specified time period, where the depth of the bit defines how accurate the measurement would also be recognized as the signal dynamic range (typically 16bit, implying a sample will vary from 65.536 amplitude values).

Audio Data analysis in Python:

Sound/audio is represented as an audio signal which has factors like bandwidth, the frequencies, decibels, pitch etc. A standard audio wave signal is represented as a function of Amplitude and Time. A traditional method of audio processing includes the extraction of acoustic features specific to the mission, accompanied by making decision strategies requiring the identification, categorization and combination of information. We have quite a few libraries in python to serve the purpose

1. Librosa

In general, the librosa Python library that is used to analyze wave signals in the audio but is more musical oriented.

Visualizing audio:

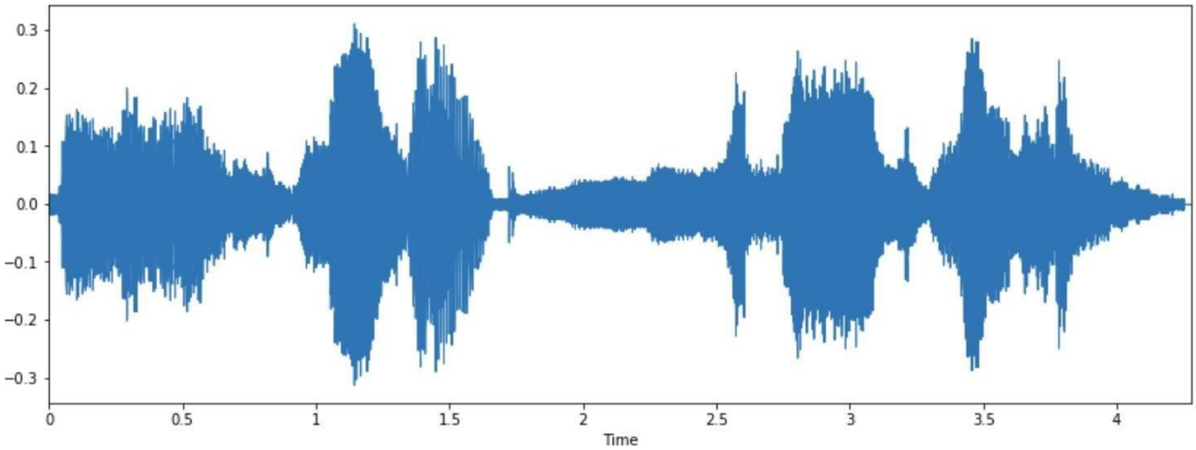


Fig. 5.1 Audio Visualization plot

Extracting features from Audio signal

Each audio wave signal has various functions. There is a need to however extract the features in accordance to the problem arising. The method by which characteristics are extracted to be used for review is called Feature Extraction. The spectral features (space- based characteristics) that are produced by utilizing the Fourier Transform to translate the time-based signal into the frequency domain, such as fundamental frequency, frequency elements, spectral centroid, spectral flux, spectral intensity, spectral roll-off etc.

1. Spectral Centroid

It specifies the frequency at which the position of the spectrum is focused towards the center. It specifies the "mass center" for audio/sound. It is similar to weighted mean:

$$f_c = \frac{\sum_k S(k)f(k)}{\sum_k S(k)}$$

where S(k) is the spectral magnitude at frequency bin k, f(k) is the frequency at bin k.

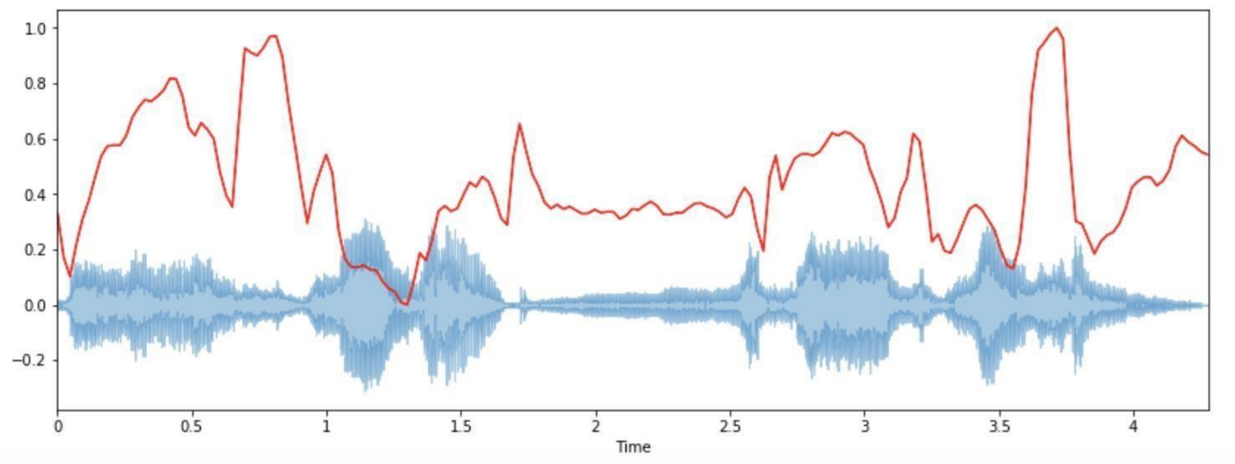


Fig. 5.2 Spectral Centroid

2. Spectral Rolloff

It is a signal type calculation. It reflects the pace at which the high frequencies decrease to 0. For the calculation, a fraction of bins in the power spectrum must be calculated wherever at lower frequencies 85 percent of its power is.

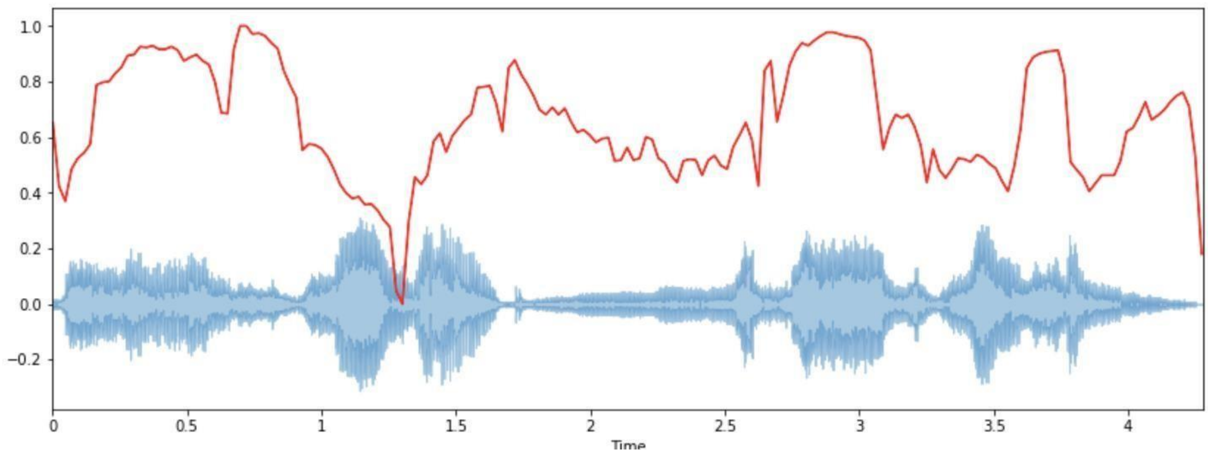


Fig. 5.3 Spectral Rolloff

3. Spectral Bandwidth

Spectral Bandwidth is known as the diameter of audio at a maximum of half the peak and is shown on the axis of wavelength by the 2 straight red lines and λSB. The p spectral bandwidth is given by:

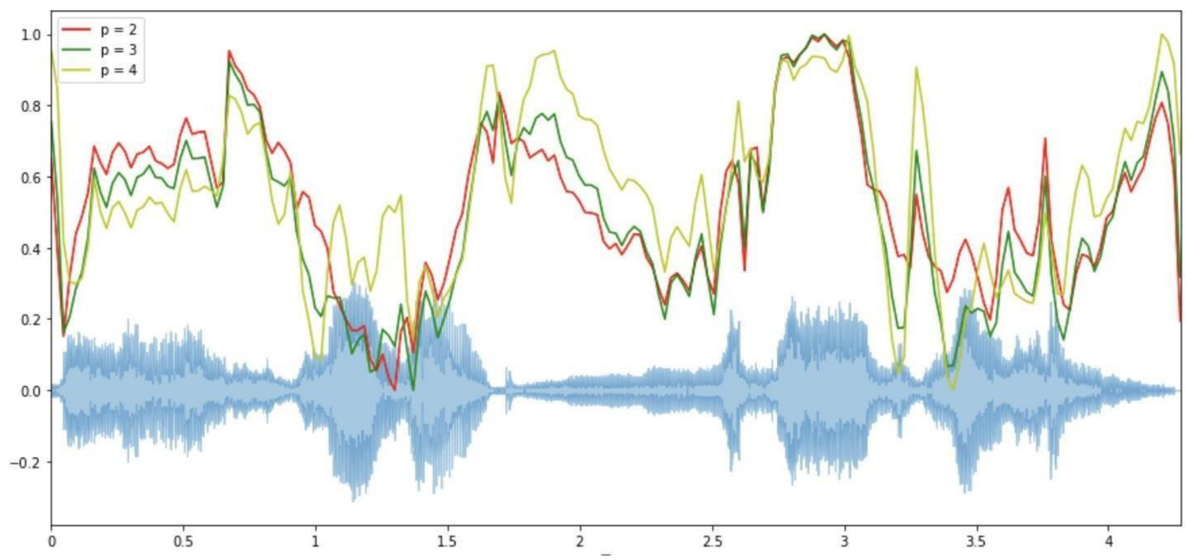


Fig 5.4 Spectral Bandwidth

4. Zero-Crossing Rate

A non complex way to measure the evenness of an audio signal is by calculating the sum of zero-crossing inside a signal sample/segment. A speech signal gradually oscillates, but an unvocal fricative would be having 3000 zero crosses per second.

$$zcr = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbb{I}\left\{ s_t s_{t-1} < 0 \right\}$$

Fig. 4. Formula to calculate the Zero Crossing Rate

s_t is the signal of length t
$II\{X\}$ is the indicator function (=1 if X true, else =0)

5. Mel-Frequency Cepstral Coefficients(MFCCs)

A signal's Mel frequency cepstral coefficients (MFCCs) are a tiny set of features that depict the overview of shape of a spectral envelope in a concise way. It molds human voice characteristics.

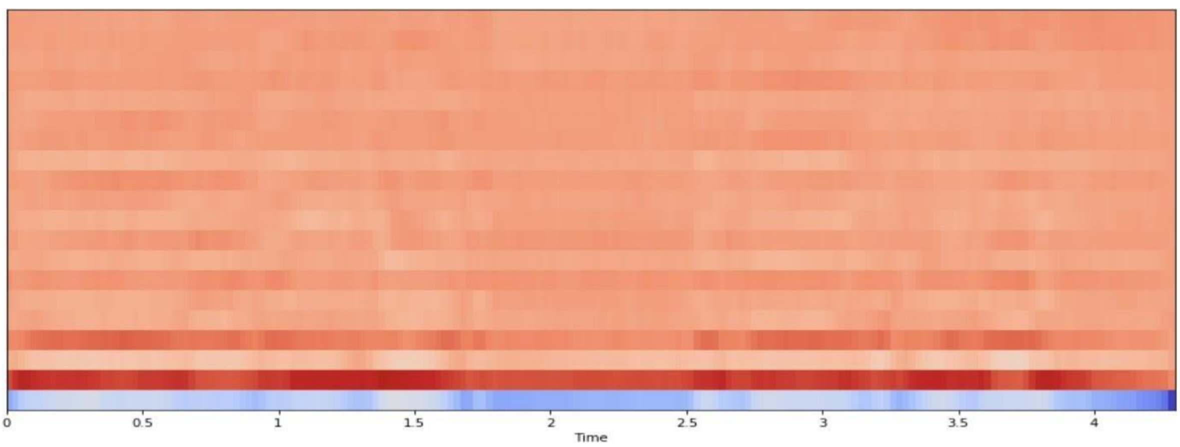


Fig. 5.5 MFCC

6. Chroma feature

A chroma function is usually a 12-element character/feature vector which indicates the amount of energy present in the signal of increasing pitch unit, {C, C #, D #, D #, E, ..., B}. In brief, It offers a rigorous way of representing a calculation of resemblance between pieces of music.

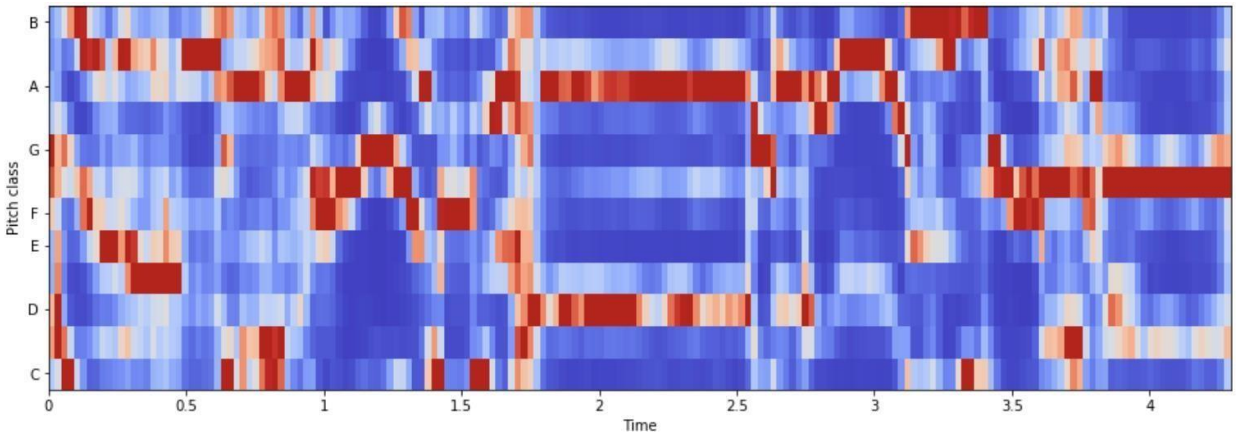


Fig. 5.6 Chroma Feature

5.2 MODULE 2: AUDIO FEATURE EXTRACTION

Pseudo code:

Procedure Feature Extraction (audio files):

Input: audio files

Output: Extract the MFCC core

features Begin

Step1: Read the dataset

files Step 2: for each

file

Step 2.1: Extract the core features from the file using librosa

library End for

Step 3: write the extracted features in the corresponding

csv file. End

Desktop application for prediction of child depression:

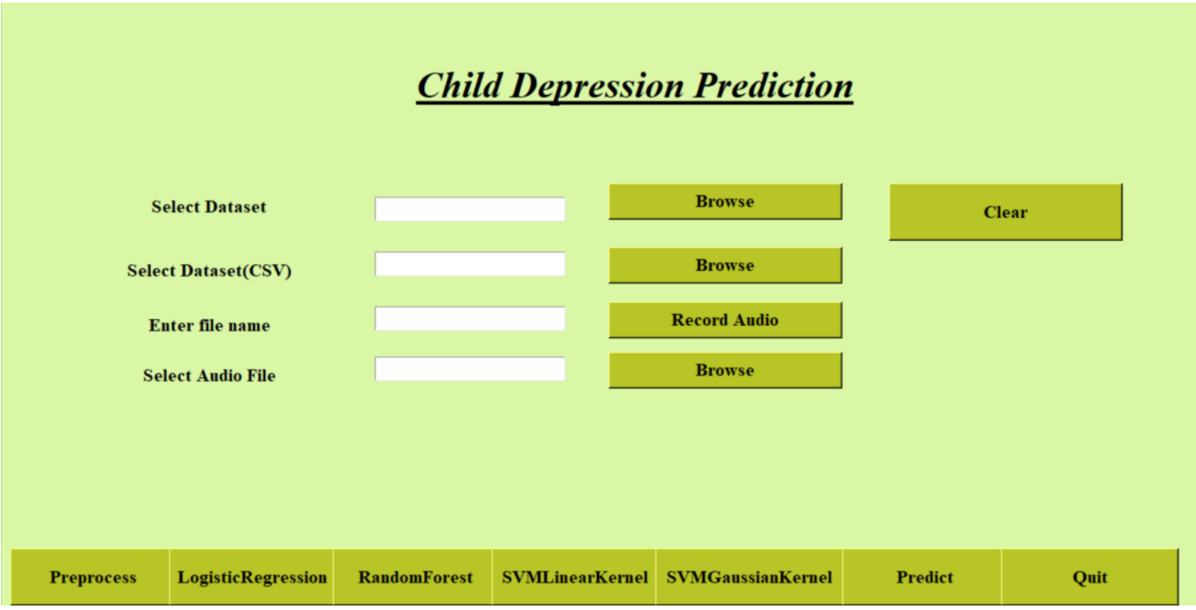


Fig. 5.7 Child depression prediction desktop application

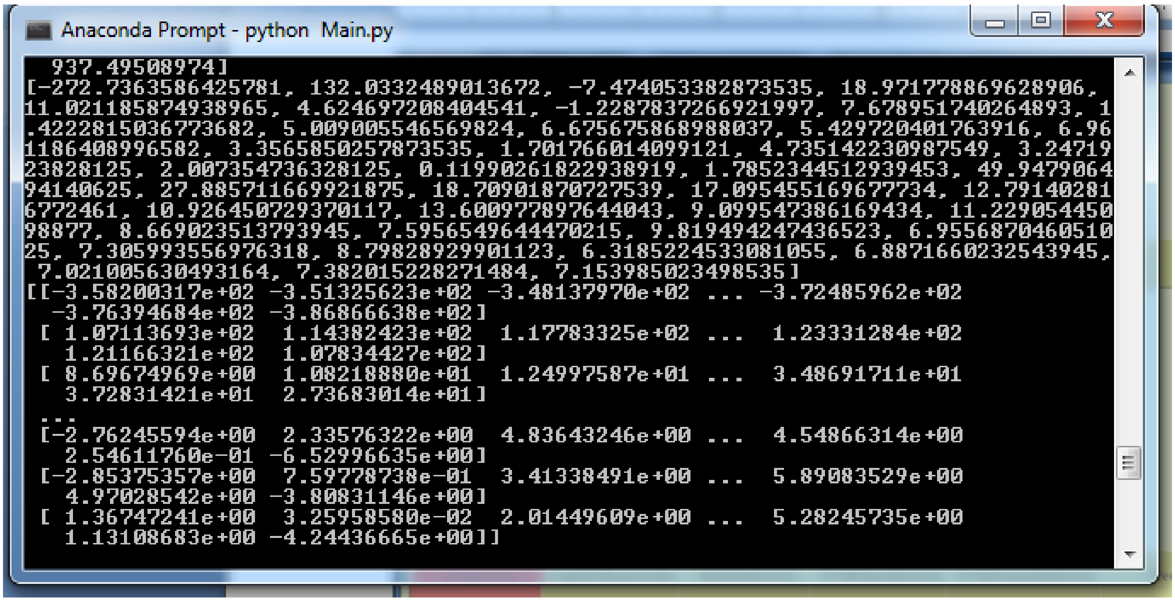


Fig. 5.8 Pre-processing

5.3 MODULE 3: MODEL DEVELOPMENT

The proposed system involves the implementation of four binary classification models. They include Random Forest, Logistic Regression, Support Vector Machine (SVM) Gaussian Kernel, and SVM Linear Kernel. These classification models relate to the audio signal attributes determined with clinical consent trained using the methods of supervised learning with data that is of high-quality audio

5.3.1 LOGISTIC REGRESSION

Logistic Regression is a model that is used to estimate a likelihood of a variable that is categorically dependent. A "dependent variable" is a binary variable that contains value encoded as either 1 or 0, which also means YES or a NO, in other words, success or a fail. In our project, if an audio recording is classified as YES (1, Success), then that child is depressed (mentally ill).

Pseudo code:

Procedure logistic regression input: Dataset

output: Classified results Begin:

Step1: Read dataset

Step 2: Split Dataset into train and test data

Step 3: Use train data to train the logistic regression model Step 4: predict the result using sigmoid function

Step5: Return the predicted class. Step 6: Calculate accuracy measures End.

5.3.2 RANDOM FOREST

It is a collection of many decision tree classifiers that are developed on sub-samples of the data set and uses averaging of these trees to predict, improve accuracy and better control over fitting of the model. Sub-sample is the same size as the actual input size but samples are drawn depending on the feature and with replacement.

Pseudo code:

Procedure Random Forest input: Dataset

output: Classified results Begin:

Step1: Read dataset

Step 2: Split Dataset into train and test data Step 3: Use train data to train the RF model Step 4: predict the result using test data Step5: Return the predicted class.

Step 6: Calculate accuracy measures End.

5.3.3 SUPPORT VECTOR MACHINE (SVM) – LINEAR KERNEL

SVM is a supervised vector ML that uses classification models to find the accuracy for two category classification problems. It uses associated learning that analyzes data used for classification and regression.

Data that is separable linearly is used with a linear kernel which means that the data can be separated using a single line. This model is particularly used with data containing a large number of extracted features. For example, it is widely known to be used for text classification as each alphabet is a new feature. For audio analysis, this model works well due to a large number of features such as energy, pitch, frequency etc., that are extracted during preprocessing.

Pseudo code:

Procedure SVM with Linear Kernel input: Dataset

output: Classified results

Begin:

Step1: Read dataset

Step 2: Split Dataset into train and test data

Step 3: Use train data to train the SVM model

Step 4: predict the result using test data

Step5: Return the predicted class.

Step 6: Calculate accuracy measures

End.

5.3.4 SUPPORT VECTOR MACHINE (SVM) - GAUSSIAN KERNEL

Gaussian Kernel, also called RBF kernel is a simple weighted linear function combination that is quantified between support vectors and a data point. It is rendered in the convolution phase by using normal distribution to weigh the adjacent pixel.

Pseudo code:

Procedure SVM with Gaussian Kernel input: Dataset

output: Classified results

Begin:

Step1: Read dataset

Step 2: Split Dataset into train and test data

Step 3: Use train data to train the SVM model

Step 4: predict the result using test data

Step5: Return the predicted class.

Step 6: Calculate accuracy measures

End.

5.3.5 TRANSFER LEARNING – WAVENET + CNN

Convolutional Neural Network (CNN) is a multilayered deep learning neural network. Although best suited for image classification, it has shown a promise in audio classification, having the ability to predict with high accuracy when trained with a large dataset.

WaveNet is a deep learning multilayered neural network used for producing raw audio waveforms and extract features. It carries out autoregressive learning with casual or dilated casual convolutions. It does predictive distribution for audio samples conditioned on the previously observed values. Which means that each time a sample is predicted, it is fed back to the model to predict the next sample. There is a convolutional sliding window that is used on the audio samples, which at every step tries to predict a sample value that it hasn't seen yet. Therefore, it builds a network around these audio samples and learns the casual relationship by trying to predict a sample using a few previous steps of the neural network. WaveNet models with a casual convolution cannot violate the order in which the input data is modelled whereas in dilated casual convolutions, a filter is applied to the network to skip some input values with certain steps. Although the output is the same value, dilated casual convolution is often used as it's more efficient considering only the important input features to the next level and allows the to operate effectively on a coarser scale.

By using the concept of Transfer Learning, we have used WaveNet feature extraction model on top of CNN and compared the results with a CNN model built using features extracted from the inbuilt MFCC in Librosa. 42 observations in 8 features are extracted using MFCC and 848 observations among all features are extracted using WaveNet. These features are fed to the same CNN classifier model.

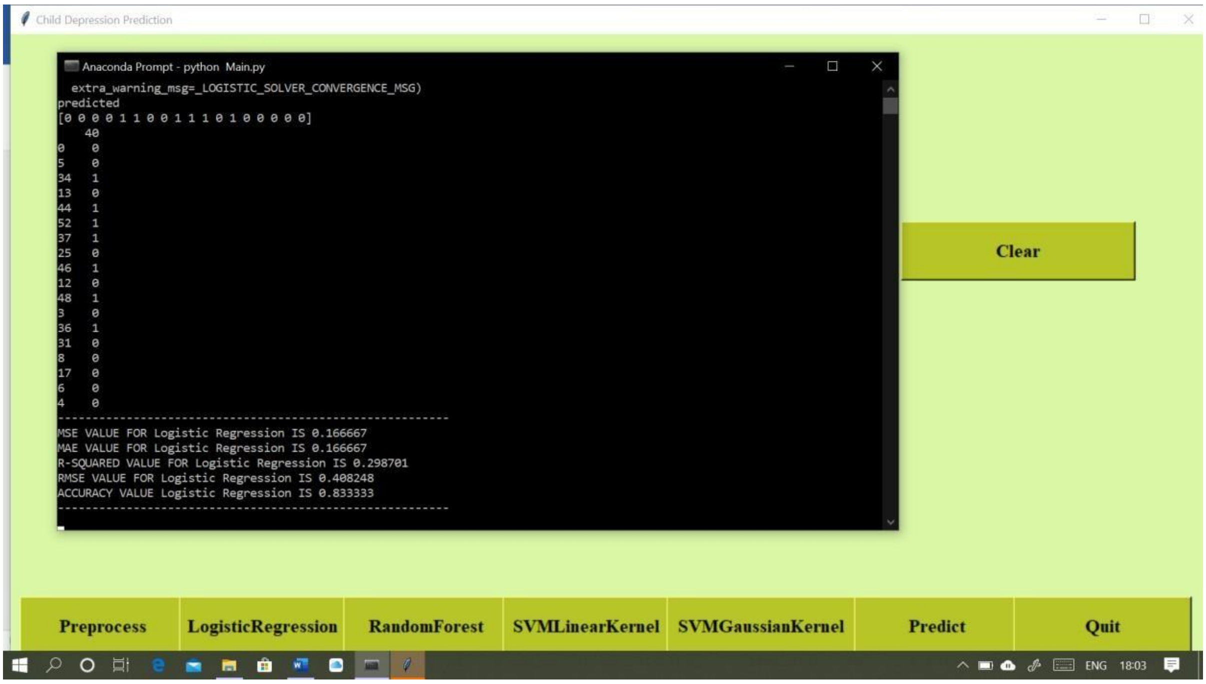
```
[19] model.summary()

Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_5 (Dense)              (None, 256)               2748672
_________________________________________________________________
activation_5 (Activation)    (None, 256)               0
_________________________________________________________________
dropout_5 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_6 (Dense)              (None, 256)               65792
_________________________________________________________________
activation_6 (Activation)    (None, 256)               0
_________________________________________________________________
dropout_6 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_7 (Dense)              (None, 2)                 514
_________________________________________________________________
activation_7 (Activation)    (None, 2)                 0
=================================================================
Total params: 2,814,978
Trainable params: 2,814,978
Non-trainable params: 0
_________________________________________________________________
```

Fig 5.9. Model Summary of CNN

5.4 MODULE 4: PREDICTION AND ANALYSIS

Logistic Regression:



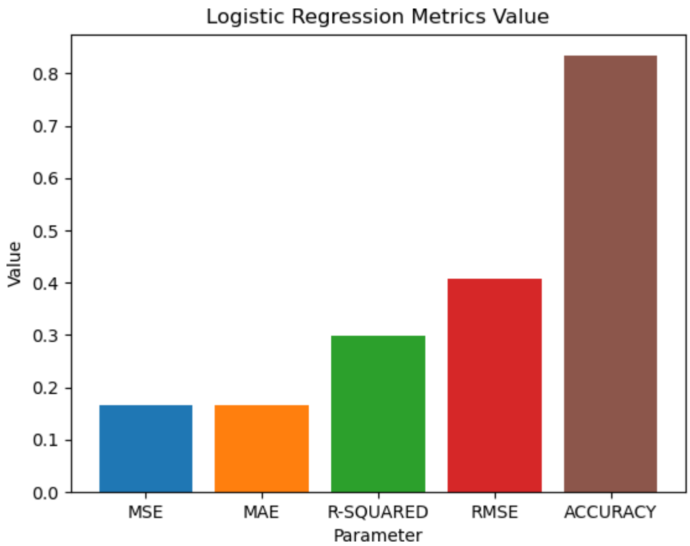
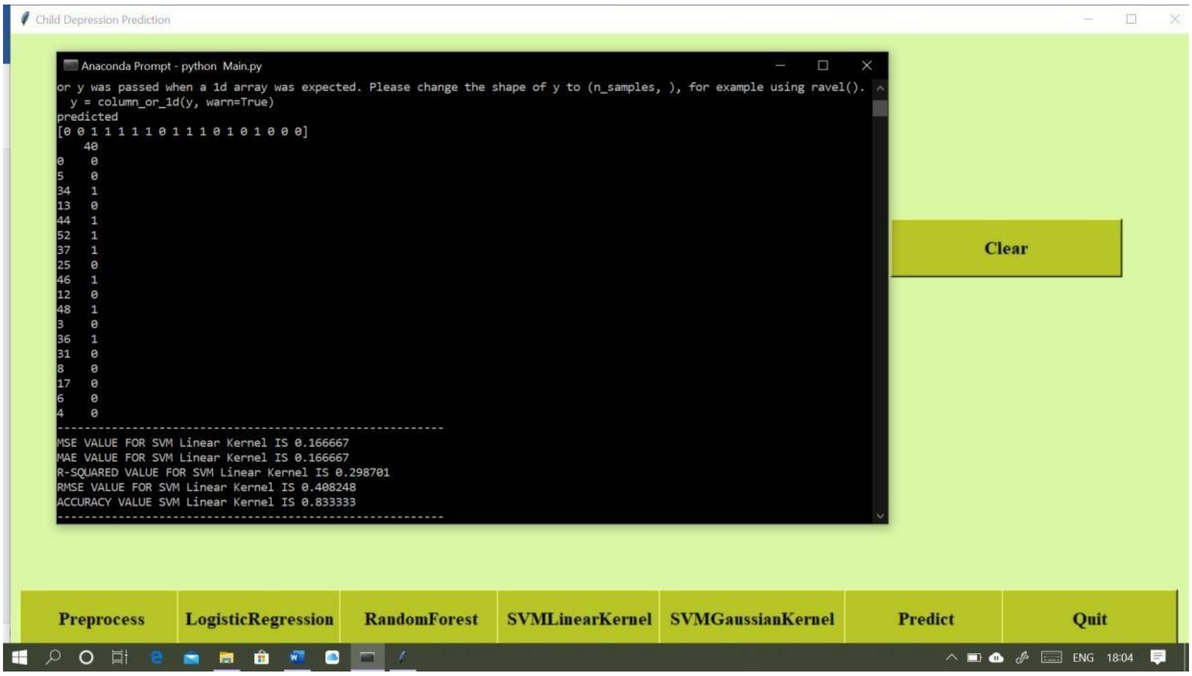


Fig. 5.10 Logistic regression prediction result

SVM Linear Kernel:



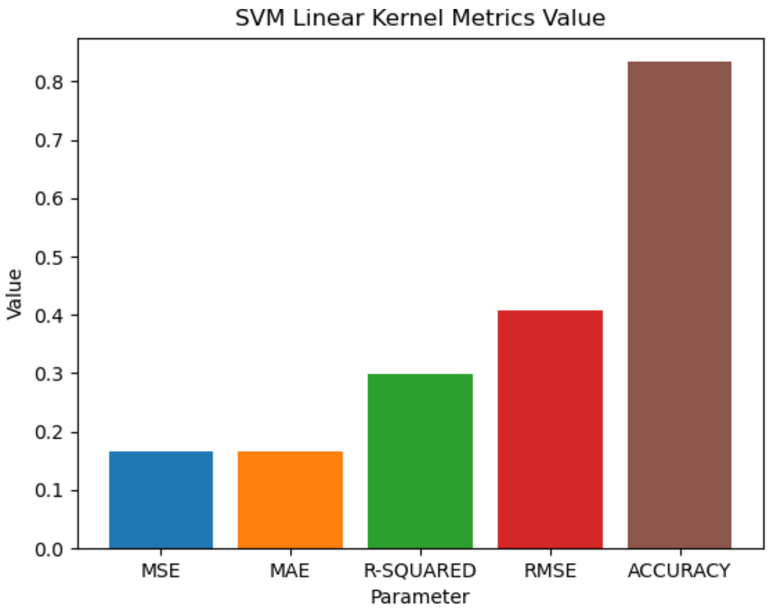
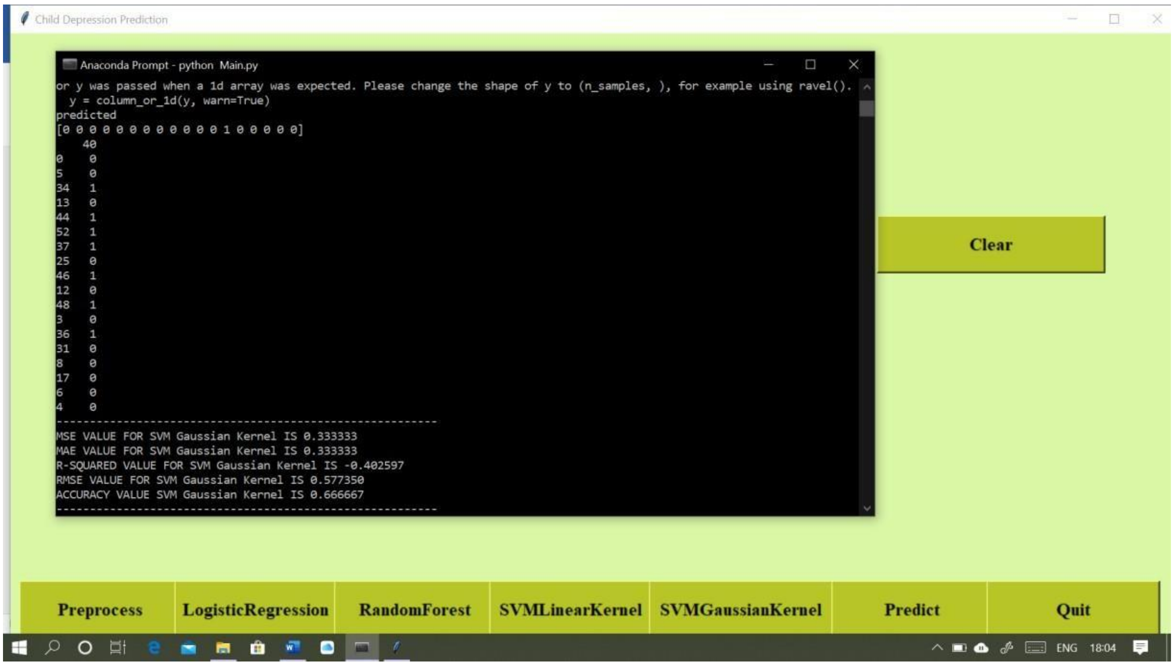


Fig. 5.11 SVM Linear Kernel Prediction Result

SVM Gaussian Kernel:



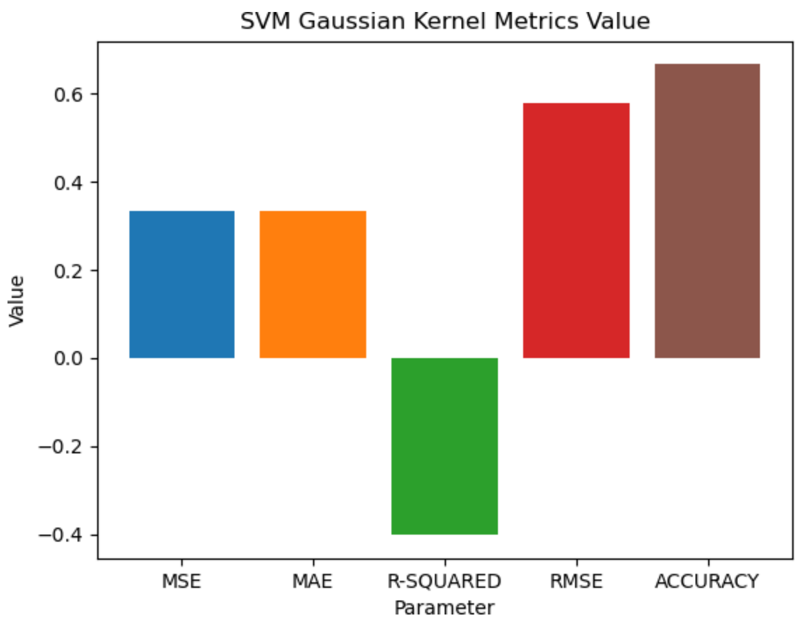
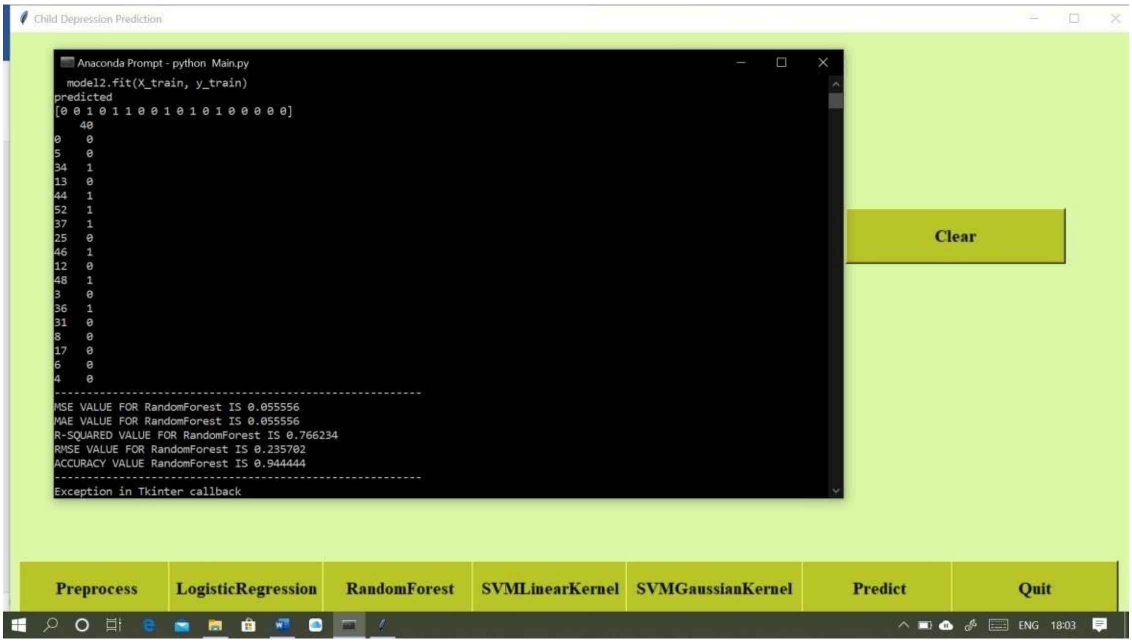


Fig. 5.12 SVM Gaussian Kernel Prediction Result

Random Forest:



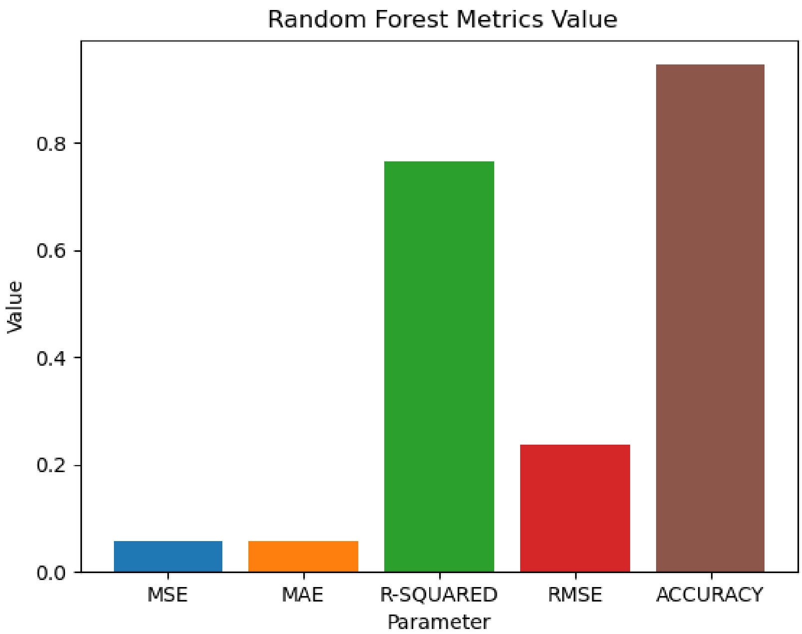


Fig. 5.13 Random Forest Prediction result

CNN with WaveNet :

```
50/50 [==============================] - 0s 3ms/step - loss: 0.5526 - accuracy: 0.6800
Epoch 23/25
50/50 [==============================] - 0s 3ms/step - loss: 0.5709 - accuracy: 0.6600
Epoch 24/25
50/50 [==============================] - 0s 3ms/step - loss: 0.4902 - accuracy: 0.7600
Epoch 25/25
50/50 [==============================] - 0s 3ms/step - loss: 0.4719 - accuracy: 0.7800
<keras.callbacks.callbacks.History at 0x7ff920cd2748>
```

```
[52] model.save_weights("/content/gdrive/My Drive/model78.h5")
```

```
[ ] Xnew = w1

    ynew = model.predict_classes(Xnew)

    print(" Predicted=%s" % (ynew[0]))
```

```
WARNING:tensorflow:From /usr/local/lib/pytho

 Predicted=1
```

```
[ ] ynew1 = model.predict(Xnew)
    ynew1[0][1]
```

```
0.9771748
```

Fig 5.14 Results of CNN with WaveNet

CNN with MFCC:

```
Epoch 4/7
50/50 [==============================] - 0s 3ms/step - loss: 7.9357 - accuracy: 0.4200
Epoch 5/7
50/50 [==============================] - 0s 3ms/step - loss: 3.9741 - accuracy: 0.5200
Epoch 6/7
50/50 [==============================] - 0s 3ms/step - loss: 2.7053 - accuracy: 0.6000
Epoch 7/7
50/50 [==============================] - 0s 3ms/step - loss: 0.8854 - accuracy: 0.7600
<keras.callbacks.callbacks.History at 0x7ff92171dd68>
```

```
[48] Xnew = mfcc1

     ynew = model.predict_classes(Xnew)

     print(" Predicted=%s" % (ynew[0]))

      Predicted=0

[50] ynew1 = model.predict(Xnew)
     ynew1[0][1]*100

     1.6573170613654042e-08
```

Fig 5.15 Results of CNN with MFCC

The comparative analysis carried out between CNN using MFCC features and CNN using WaveNet features show that WaveNet produced more reliable and robust features whereas due to simplicity of the MFCC structure, time taken it to compute the features was much faster than WaveNet with its complex neural network structure. In terms of performance, WaveNet gave slightly better accuracy.

5.5 MODULE 5: OUTPUT

The desktop application built is used to predict depression in young children by analyzing their speech. It gives a graphical representation of the performance results of the four classification models. New imported or recorded audios are detected for depression using the Random Forest model as it gives the most accuracy of 94 % in comparison to Logistic Regression, SVM Gaussian Kernel and SVM Linear Kernel with accuracy of 83.3 %, 66.6 % and 84% respectively.

When MFCC features were used on CNN, it resulted in 76% prediction accuracy whereas the transfer learning model showed an accuracy of 78%. Despite the small differences, theoretically, transfer learning has gained a lot of importance lately due to saving time on training and better performance of neural network. But in our implementation, since small size data set is trained by using large number of features produced using WaveNet resulted in overfitting, thus reducing the predictive accuracy.

CHAPTER 6

TESTING

6.1 MODULE TESTING

Test strategy and approach:

A system must be tested in order to find errors and loopholes in the system. The test objective and features to be tested are defined below

Test objectives:

- All input files must be valid as uploaded by the user
- All input files must undergo feature extraction with the help of right set of python libraries
- The audio must be classified as depressed or normal based on its features

Features to be tested:

- Valid user input path
- Valid user input format
- Feature extraction from the input files as uploaded by the user
- Prediction of correct values to classify the audio file into depressed or normal.
- Valid and compatible python libraries to be imported that is in accordance to the project

6.2 TEST CASE - 1

Test Case#	UTC01
Test Name	User input file format
Test Description	To test user input file as dataset folder which contains audio .wav file
Input	Dataset folder
Expected Output	The file should be read by the program and print path of input folder on console
Actual Output	The file is read and print contents accordingly
Test Result	Success

Table. 6.1 User input file format

Input path check:

When a user inputs the folder, which contains audio files of .wav format, the setup must check for the path of the folder, and store it as the input folder path. The path to the folder should also be printed onto the python console, so that the user can verify if its the same path of the folder that he wishes to give.

6.3 TEST CASE - 2

Test Case#	UTC02
Test Name	User input format
Test Description	To test user input file as dataset folder which contains audio .wav file
Input	folder as null
Expected Output	It Should show the alert Message enter valid input
Actual Output	Shown alert message
Test Result	Success

Table. 6.2 User input format

Validity check:

When a user inputs the format, the expected format is a .wav file. So when a folder is inputted by the user, the whole folder must be checked so that it contains the proper .wav audio files throughout the folder.

If the folder doesn't meet the requirement, i.e., if the folder does not contain all .wav format input audio files, the setup should prompt an alert message to the user asking to input only the .wav format, so that the audio files can be analyzed further for the experiment cause.

6.4 TEST CASE - 3

Test Case#	UTC03
Test Name	Feature Extraction
Test Description	To test whether extracting important features
Input	.wav file
Expected Output	It extracts the important features
Actual Output	Its extracted the important features
Test Result	Success

Table. 6.3 Feature Extraction

Extraction of features:

Every audio file contains the two important features, viz., sample rate and sample data. Sample rate is the number of times per second a sound is sampled. A sample data refers to a unit of audio data that is stored as samples and played at a certain rate so that the audio is heard.

One can perform multiple transformations on sample rate and sample data and hence extract various useful features from it. Features such as Zero Cross Rate, Energy, Entropy of Energy, Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral Flux, Spectral Roll off, MFCC, Chroma Vector, Chroma Deviation can be extracted with certain libraries in python. In our case, we make use of MFCC features to analyze the audio clip.

6.5 TEST CASE – 4

Test Case#	UTC04
Test Name	Test case to determine correct predicted value of depressed or normal
Test Description	To test whether a given statistical model predicts the depression or not of given user input.
Input	A Extracted features
Expected Output	The algorithm should predict the depression stateas per historical data collected
Actual Output	The predicted value by algorithm is closer to the specified value in the historical data.
Test Result	Success

Table 6.4 Correct Prediction Value Determination

Depression detection:

Using the features that are extracted using libraries in python, the audio clip can be analysed and classified into normal or depressed. With the help of MFCC the whole audio signal can be detected. For depression detection, factors such as Fundamental frequency (F0), speech intonation, rate and rhythm, Loudness factor, Formants (the spectrum maxima), Jitter (signal fluctuation) and shimmer (peak variation) are all considered to diagnose depression in the audio and classify the audio into normal and depressed.

6.6 TEST CASE - 5

Test Case#	UTC05
Test Name	Test case for importing valid python libraries
Test Description	To test whether an algorithm to implement congestion nodes works without sklearn and keras models
Input	Import all valid libraries sklearn, flask and keras libraries
Expected Output	An error should be thrown specifying “error importing libraries sklearn, flask and keras libraries”
Actual Output	An error is thrown
Test Result	Success

Table. 6.5 Importing valid python libraries (validity check)

The libraries for the whole process must be imported and it should be compatible with the existing libraries too. The libraries must be valid and in accordance to the program that is being executed. ML libraries such as keras, tensorflow, sklearn, feature extraction libraries such as librosa, UI libraries such as flask,tkinter and exclusive libraries like pydub and wave for audio have been installed in our project. If there is any sort of import error, then the setup must give an import error.

Test Results: All the test cases that are identified and mentioned above have been passed successfully. No errors encountered.

CHAPTER 7

RESULTS

The below table shows the performance evaluation of the four machine learning algorithms implemented. Random forest proved to be the best fit model compared to the other three models with accuracy 94%.

Models	Accuracy
Random Forest	94%
Logistic Regression	83.3%
Support Vector Machine – Linear Kernel	84%
Support Vector Machine – Gaussian Kernel	66.6%

Table 6.6 Performance Evaluation

Deep learning neural networks, WaveNet with CNN gave an accuracy of 78% whereas CNN built with MFCC features gave an accuracy of 76%.

CHAPTER 8

CONCLUSION

The outcomes obtained shows that a machine learning analysis of patterns in the speech of children can identify children with internalizing psychopathology during a short task for induction of anxiety. This model of statistical pattern recognition performs better clinical thresholds on parent-reported child diagnoses collected and indicates its ability within this community as an unbiased diagnostic tool. A detailed analysis of the audio features selected for this classification showed that affected children show low-pitch voices, often inflected and high-pitched content.

FUTURE ENHANCEMENTS

In future we can use the deep learning algorithm to increase the efficiency of the prediction result. The accuracy of Transfer Learning model can be increased by using large data sets. Since WaveNet and CNN are complex neural networks containing many layers, they require very large audio samples to produce accurate predictions without the problem of overfitting. By using very large dataset, transfer learning can promise a very high accuracy percentage.

CHAPTER 9

REFERENCES

[1]. France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). "*Acoustical properties of speech as indicators of depression and suicidal risk". IEEE Transactions on Biomedical Engineering, 47(7), 829–837.* doi:10.1109/10.846676

[2]. Moore, E., Clements, M. A., Peifer, J. W., & Weisser, L. (2008). "*Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech". IEEE Transactions on Biomedical Engineering, 55(1), 96–107.*

[3]. Cummins, Nicholas & Epps, Julien & Breakspear, Michael & Goecke, Roland. (2011). "An Investigation of Depressed Speech Detection: Features and Normalization". Proc. Interspeech. 2997-3000.

[4]. McGinnis, R. S., McGinnis, E. W., Hruschak, J., Lopez-Duran, N. L., Fitzgerald, K., Rosenblum, K. L., & Muzik, M. (2018). Rapid Anxiety and Depression Diagnosis in Young Children Enabled by Wearable Sensors and Machine Learning. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

[5]. McGinnis, E. W., Anderau, S. P., Hruschak, J., Gurchiek, R. D., Lopez-Duran, N. L., Fitzgerald, K., … McGinnis, R. (2019). *Giving Voice to Vulnerable Children: Machine Learning Analysis of Speech Detects Anxiety and Depression in Early Childhood. IEEE Journal of Biomedical and Health Informatics, 1–1.*

[6]. Chu-Xiong Qin, Dan Qu and Lian-Hai Zhang."Towards end-to-end speech recognition with transfer learning". Qin et al. EURASIP Journal on Audio, Speech, and Music Processing (2018) 2018:18

[7]. Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo. “A Study of acoustic features for Depression Detection”, IEEE 2014.

[8]. Ying Yang, Catherine Fairbairn, and Jeffrey F. Cohn, Associate Member, IEEE “Detecting Depression Severity from Vocal Prosody”, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 4, NO. 2

[9]. Le Yang “Multi-Modal Depression Detection and Estimation”. 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)

[10]. Jinming Li,Xiaoyan Fu ,d Zhuhong Shao, Yuanyuan Shang.“Improvement on Speech Depression Recognition Based on Deep Networks”, IEEE 2018.

[11] Harati, S., Crowell, A., Mayberg, H., & Nemati, S. (2018). Depression Severity Classification from Speech Emotion. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

[12] Lam, G., Dongyan, H., & Lin, W. (2019). Context-aware Deep Learning for Multi-modal Depression Detection. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[13] Zhao, Z., Bao, Z., Zhang, Z., Deng, J., Cummins, N., Wang, H., … Schuller, B. (2019). Automatic Assessment of Depression from Speech via a Hierarchical Attention Transfer Network and Attention Autoencoders. IEEE Journal of Selected Topics in Signal Processing, 1–1.

[14] Rani, B. (2016). Detecting depression: A comparison between I-Vector technique and fuzzy membership functions. 2016 International Conference on Inventive Computation Technologies (ICICT).

[15] Huang, Z., Epps, J., Joachim, D., & Sethu, V. (2019). Natural Language Processing Methods for Acoustic and Landmark Event-based Features in Speech-based Depression Detection. IEEE Journal of Selected Topics in Signal Processing, 1–1.

cd

ORIGINALITY REPORT

12%	7%	7%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Ellen W. McGinnis, Steven P. Anderau, Jessica Hruschak, Reed D. Gurchiek et al. "Giving Voice to Vulnerable Children: Machine Learning Analysis of Speech Detects Anxiety and Depression in Early Childhood", IEEE Journal of Biomedical and Health Informatics, 2019 Publication	2%
2	www.kdnuggets.com Internet Source	1%
3	Ryan S. McGinnis, Ellen W. McGinnis, Jessica Hruschak, Nestor L. Lopez-Duran et al. "Rapid Anxiety and Depression Diagnosis in Young Children Enabled by Wearable Sensors and Machine Learning", 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018 Publication	1%
4	www.slideshare.net Internet Source	1%