Deception Analysis Using Deep Learning Based on Voice Stress Detection

MSc Research Project
Data Analytics

Geethu Issac
Student ID: 20210515

School of Computing
National College of Ireland

Supervisor: Mr. Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Geethu Issac
Student ID:	20210515
Programme:	MSc in Data Analytics
Year:	2021-22
Module:	MSc Research Project
Supervisor:	Mr. Jorge Basilio
Submission Due Date:	19/09/2022
Project Title:	Deception Analysis Using Deep Learning Based Voice Stress Detection
Word Count:	6585
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Geethu Issac
Date:	19th September 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Deception Analysis Using Deep Learning Based Voice Stress Detection

Geethu Issac
20210515

Abstract

Voice Stress Detection concurrently is a sorcery that targets to deduce deception calculated by identifying the amount of stress in the voice signal. It becomes conceivable to detect the stressed voice in this century with the significant development, Artificial Intelligence (AI). Voice, being the core for communication is a good source of input signal to an AI model to analyze deception. The demand for healthy mental life of this era is the prime objective tried to be fulfilled with this work. The difference in the fluency of speech of a stressed person from that of an unstressed using the Deep Learning method of Convolutional Neural Network (CNN) is the featured sweep of this work. The dataset used for the implementation of the CNN model for analysing deception is The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). RAVDESS is a combination of unisexual voices of 24 different subjects with six emotions, which was analysed using the neural network and later binary classified into stressed or unstressed. The CNN model is implemented at the beginning on the voice of a single actor followed by 24 actors. A comparison on the existing Machine Learning models with Deep Learning model was also performed. An accuracy of 72.5% was obtained in classifying voice with an acceptable percentage of true positives with the CNN.

1 Introduction

Deception Identification models developed using Deep Learning is a superior trend in analysing the stress of a voice signal. A healthy mental stage is a principal necessity of human beings in this 21st century. The actions of a person entirely depends on the mental health. It is hence predominant to balance this health through sufficient control on the emotions and thereby the actions. Voice with a wide variety of different emotions are a good choice of identification of stress and the actions because can come up with wise as well worst actions in varied activities of a human. Within a span of several emotions, stress is that one particular feeling which can make mental and physical well being in trouble. This research work focus on

- Analysing the emotions in the speeches as they have an salient role in developing the actions of the humans
- Use of source signal other than EMG, EEG and images of face
- Use of voice signal all alone rather a combination of voice and face

- Use of unisexual dataset

- Implementation of Deep Learning technology in deception analysis

- Developing model through recording natural emotions rather than creating emotions with sufficient samples of data

Covid '19 is a pandemic that has affected the entire world so badly that many of them were left jobless. It took an year to be back into a normal life for a large section of people. It was difficult for many to be back into a corporate world due to the stress in the interviews they faced after the pandemic. Even though most of the interviews were conducted online, surveys noted that 85% of the candidates conveyed about the stress they faced before, during and after the interview. As voice was only source to identify stress, it can a good input signal which will be focused on in this research work. In order to bring in a generation with quality mental health for future, it is now paramount to analyze these stressed voices along with the unstressed ones. Although quite a good number of solutions have been developed, most of them fail to provide a feasible content about the mental well being from the literature review conducted for this research. Rather than considering feature engineering for analysis of deception, this work focus on implementing neural networks using voice as source and classifying emotions for stress identification. Although, other source signal like EEG, EMG or combination facial and voice, face images are used for determining the stress, use of voice all alone is a topic of discussion which will be taken forward in this work with atmost novelty. The major purpose in the research modulated into a question is given below.

- **How the deception or depression can be indicated using Stress Detection applying Convolutional Neural Networks?**

Voice analysis for the identification of deception includes a lot of positives and negatives. As discussed earlier, different source signals are used as of now. The use of EEG and EMG are not convenient in the extraction as data since it require sensors or electrodes for the same as well as a large dataset with multiple levels of filtering to remove all noises. The work of(A.Baum; 1990) suggest that neural network may outperform the commonly available techniques in this aspect. The selection of voice as source for this research is due to the classification of emotions as archetypal emotions as discussed in the research performed by(Gedeon and N.Sharma; 2012). A temporal contextual information makes up what we call a voice signal and also the dependency among the frames that lie to each other. A Convolutional Neural Network can take up this contextual information to its hidden layers to perform classification. Windowed frame variation of voice make it a better choice for CNN.

The simple block diagram for the implementation of the research is given in Figure 1. The source signal, here the voice, is passed after several pre-processing methods following the feature extraction. The extracted features can be steered to classifiers to identify emotions, to analyze deception. CNN is used to classification of voice which will be discussed in detail in the following section. The works done till now in accordance with this will also be discussed with this research and its results along with improvements that can be done in future.

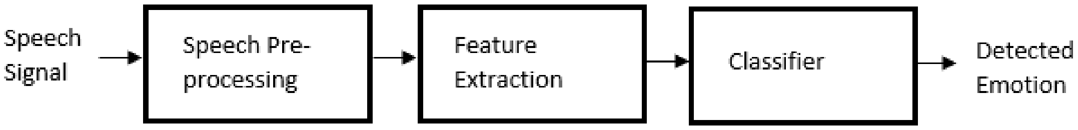


Figure 1: Block Diagram

2 Related Work

Identification of the suitable source signal for the detection of deception is an issue of serious concern. EEG, EMG and face images are common input for the analysis. Due to intervention of the noise signal, EEG and EMG are very difficult to obtain as clean data. The pros and cons in implementing the Artificial Intelligence models such as Machine Learning models and Deep Learning models using the above mentioned input signals as per the previous works will be covered in this section. A new solution is developed in this work to wipe out the cons associated with existing models and input signals. In order to come out with a perfect solution, all the existing models are analysed to improve accuracy and efficiency of the classification. Due to the less interference in developing the data, the voice signals will be used in this research work. Each of the subsections of this particular literature survey discuss with all the existing solutions of this research problem.

2.1 Statistics in Deception Interpretation

Log-Likelihood Ratios as well as Fisher's Linear Discriminant Analysis in the determination of deception in the speech of the Spanish people as discussed in Palacios Alonso (2015) is the major statistical approach proposed till now. Due to the lack of the easily available databank in relation with this, male and female Spanish speeches were recorded for the determination using statistics. It was found in the analysis that the subjects were stressed in recording the speeches although the statistical model performed with accuracy in the voice classification of percent for neutral voices and 56% for the stressed speeches. Also, it is clearly mentioned in that work, the classification accuracy obtained is due to the fact that there existed a bias in the recording of the speeches. It clearly notifies the outcome of using a clearly defined dataset for implementation of statistical techniques. As the work focus on developing a mental health in future, use of the correct dataset and implementation method is a necessity without any chance of convergence. It was also understood that the use of voice signals with less noise are good source signals that can bring better predictions of unknown data. It can also be noticed in the work that the use of unisexual data can give good classification of the stressed and unstressed data. Hence it can be good approach to be taken forward where rather than using either female or male voice , it is great to use combination of both to give accurate values. The software named PRAAT is used in the work of Savita Sondhi and Salhan (2016) for the analysis of stress using the speeches of persons associated with crime. The work discuss the features of the voice signals that can be taken forward in the analysis of the stress content. The dataset was very difficult to interpret as the criminals where discriminated based on the sex. Several vocal parameters were taken into consideration. Some of the voice parameters analysed like shimmer, jitter and pitch were not relevant ones in the analysis. Even though pitch contributed to some extend others not. Pitch was the main feature that was recognized as varying in the analysis. The mean frequency and formant frequency of the speech signal were identified as changing considerably. Another main limitation of the method lies in the use of the particular software for change of features observed and not in all the cases of general application. Similarly, the cost of installation of the software and modelling was not economical as per the researchers and couldn't come up with the features contributing to stress analysis. It was also identified that the speech signal as itself was a good source rather than using the features from them. As the model doesn't provide a feasible solution for the determination of deception, no factors were taken into account to be taken forward in this research work. The features of voice signals and voice signal as raw cannot be connected in a single thread as they don't have any common sink to be connected, hence in this work only raw voice or speech signal will be considered.

2.2 Deception Determination by Transforms

Use of wavelet transforms in the analysis of Malayalam language for the determination of stress was discussed in the work developed by (Anto and Raji; 2009). For the purpose of feature extraction, discrete wavelet transforms of the speech signal used were used rather than the features like loudness, pitch, jitter, frequency. Rather than a bisexual dataset, subjects were both males and females for the work. The bag of corpus consisted of four hundred different Malayalam words in different stress levels. The features were identified from the voice after thirteen different decomposition of the discrete wavelet transform function. For the experimental analysis of feature extraction, Daubechies4 discrete wavelet transform was used. As low frequency components distinguish voice signal much better than high frequency signal, the former is used in the analysis. The use of discrete wavelet transform for the voice feature extraction achieved a good percentage of accuracy. Fast Fourier Transform (FFT) for the analysis of stress is discussed in (Cabrera and Lopez; 2011). How the effect of rate of respiration determines stress was considered. Gold Wave Software and Matlab was used in the work. The analysis of actual deception was failed using FFT making it unsuitable for feature extraction in the voice. In the work discussed by (Baek and Chung; 2020), the instances for developing stress was determined using multiple regression where input is the features of voice. Features were identified using the technique, where some were removed later on for good accuracy. Even though neural network was implemented in the work, R squared value was identified to be quite high with the rise of variables. This in turn affected the accuracy. The use of EEG signal in the mixture of feature extraction and classifiers is the main content in the work discussed by (Prashant; 2019). Teager-Kaiser Energy Operator is used for the feature extraction. TKEO have been used in collation with the classifiers. Even though feature extraction was feasible, use of EEG was not economical and time consuming. The identification of the bands of EEG is tedious, hence EEG is not a good choice in analysis of voice deception. The use of facial recognition is used in the work of (Anna Esposito and Cordasco; 2020) to sum up the combination of emotion recognition in stress analysis with voice. The accuracy of 70.5% was obtained for females and 82.2% for males for this model. As there were in the mismatch of the predictions obtained in true scenario, merging of behavioural and emotional analysis is not a probable approach for the voice stress analysis.

2.3 Machine Learning in Stress Analysis

The incorporation of machine learning in identifying the stress in the tweets of the famous social media app twitter is discussed by (Kinariwala; 2019).The sentiments in each of the tweets are analysed for stress identification using TensiStrength which calculate the strength of deception.The tweets were segregated and Support Vector Machine is used for predictions with ngram. The value of Precision was 65% and recall was 67%. Word Sense Disambiguation is faced in situations of machine learning. Even the use of ngrams failed in some experiments.Also preprocessing was not acceptable and the size of dataset was greatest limitation.

The amount of stress in a voice is identified using the hormone change by machine learning method is implemented in the work of (Vaikole; 2020). The creation of mel frequency coefficients was the preprocessing task accomplished in the work. The features obtained after filtering is used for classification of stress or unstressed situation. The score of stress is calculated to identify the amount of deception. The accuracy of predicting deception was 50%. The dataset used was unisexual which contributes to a major limitation. Also combination of neural networks with filtered coefficients was found to be a good approach which will be considered in this work. Virtual Reality (VR) based game used for detecting deception is discussed in the work done by (Arushi and Teoh; 2021). K Nearest Neighbours, SVM, Random forest classifier were used to identify the amount of deception. As the amount of input features considered were very low, the accuracy of all the models couldn't give an accuracy higher than 57.80%. In such a case neural networks was preferred by the author in case of lesser features.

Prediction of action based on speech is commonly used nowadays. Emotions are mainly used for such analysis. Speech Acquisition Device used for noisy environments with less impact of noise is discussed in (Barlian Henryranu Prasetio and Tanno; 2020). An EDA is performed here to identify the emotions in the speech. The filtering based on Mel frequency coefficients discussed earlier is carried in this research work. Using clustering methods, features in the voice signal were extracted rather than the detection and analysis of deception. Sound continence is a major symptom of a well known neural disorder called Parkinson's disease. Deception and problem of respiration are common issues associated with this. Inorder to detect the disease acoustic elements of speech were considered. Mel Frequency Cepstral Coefficient (MFCC) was used for extracting features from the voice signal. Gaussian mixture model was also taken into account as per the work. Instance based learning which was built on top on neural networks is covered in (Soham Dasgupta and Masunda; 2017) work to analyse the deception. As the criminal scenes are increasing, lie detection can play a good role in investigation where no other method exists till now. VSA can be integrated to analyse the statements of criminals. The sound of the criminals with stress was an issue of concern in this work. A long short term memory structure based deep neural network was discussed in (Felipe Mateus Marcolla and Dazzi; 2017) following all the experimenst listed above. The dataset used was that of an interview as exploratory data to neural network. The neural network developed showed a precision of 72.5% . Under this technique the voices that are in stress were easily identified. It was found that the voice with stress content is highly fluctuated than the other normal unstressed voices.

For the estimation of stress a neural network was implemented by (Hyewon Han and Kang; 2018) using the voice as an input. As the era of automation is quite trustworthy, automated stress detection was an acceptable concept. Hormone level changes are also a notable content in identifying the stress associated with a voice. But it not trustworthy as per the researcher as hormone changes can occur due to mental as well as physical variations. MFCC was adopted to generate the features for analysing the stress in the voice as per the researches available. These MFCC coefficients after preprocessing act as source signals for the model considered. The deep neural networks like long short-term memory and feed-forward networks were used in this work to determine the amount of stress. To get better accuracy and precision, data was obtained with atmost care. Hormone level higher than 10% was considered in generating the dataset with an accuracy of 66.4%. The interviews conducted act as stressed which is discussed in the work of (Kevin Tomba and Hawila; 2018) where the neural network is used to measure stress. MFCC were used as features in the analysis. The datasets used were generating accuracy but was seemed to be biased based on the sex of the subject in the recording which is a factor taken into consideration. For the physically challenged people, automated stress detection was analysed in their own language as discussed in the work done by Dipanshu Someshwar and Chaudhari (2020). Deep Neural networks with gestures of hand were used as input were a text was generated as output. Tensorflow helped in mapping images to language in the work but the stress was failed to be identified.

Emotions in the voice signal of a subject was identified with neural networks in the work discussed by (Lakmal Rupasinghe and Kulathunge; 2021). The accuracy was not as expected even after implementing the AI models due to bias in the dataset. User experience identification was performed by wearable stress detection dataset in the work of (Alexandros Liapis and Voros; 2021). The features considered for analysis was electrodermal activity and skin temperature. The classifier couldn't predict well because the dataset was created using the electrodes for extraction. An emotion detection architecture was created by (Cristina Luna Jimenez and Martinez; 2021a) using bidirectional LSTM. The accuracy of the model was acceptable as only a few emotions were taken into account were most of the emotions analysed were unstressed. An emotion detcetion system was developed by (Cristina Luna Jimenez and Martinez; 2021b) using both face and voice signals. Transfer learning and fine tuning was adopted. Static and sequential models was generated with accuracy of 80% but the dataset was biased in the same which is not of a considerable interest.

2.4 Conclusion

From the assessment of the previously known researches, it was found that using a dataset which is not biased but unisexual can encourage in getting good predictions. This explains that the data considered should equally incorporate male and female voices. If using either male or female dataset, it should only be used for predicting the same type. It was also found that a dataset with a large amount of sample predict well than others. Very few samples in the dataset result in overfitting and makes the model predict worse as input is sensitive. It is also necessary that emotions of a single person is obtained for analysis as it could give better predictions and to classify the vocal characteristics well. Depending on positive or negative emotions, predictions of stressed or unstressed are relevant. All these conclusions are taken as objectives of this work.

3 Methodology

Feature Extraction and Deep Learning Neural Network for prediction and classification are the main concepts that are finely composed in this work from the literature survey conducted. The proposed architecture discussed in the above section will use MFCC for feature extraction and speech processing as well as the Convolutional Neural Network will be used for classification following the Knowledge Discovery in Databases (KDD). The 2 shown above of KDD the steps followed in discovering knowledge in databases.

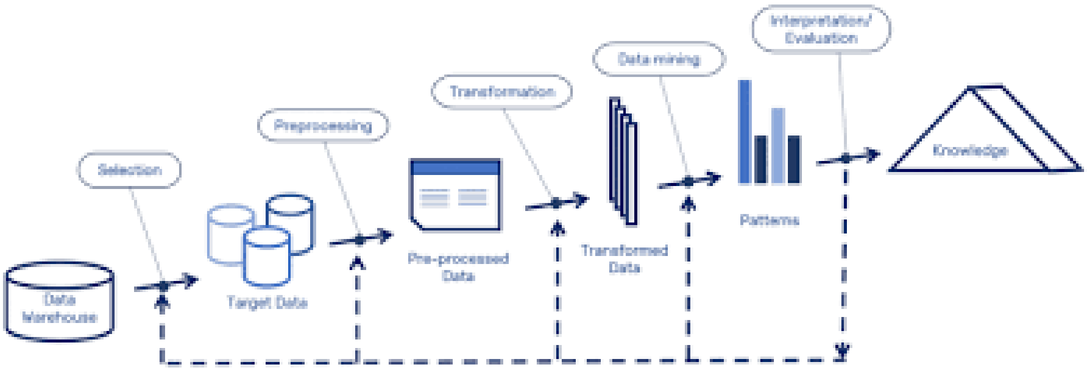


Figure 2: Knowledge Discovery in Databases

These steps were the foundation in implementing the research work and are discussed below.

3.1 Dataset Identification

From the literature survey, it was identified that the selection of the dataset is a supreme task. Ryerson Audio-Visual Database of Emotional Speech and Song dataset was identified as the one which satisfies all those requirements of the work such as unbiased bisexual data, speech data. Five variety of features were incorporated in the acquisition of dataset. 7356 audio files of 24 actors are used in this work as in the dataset of which 12 are male subjects whereas 12 are male subjects. The voice including 24 North American Actors had different emotions such as surprised, happy, sad. fear, anger and neutral base at different intensity. The .wav files having the vocal characteristics of subjects. The auditions were conducted for subjects who are later recruited after assessment with rating from analysers for recording. All the recording space requirements were explained in prior for correct emotion acquisition. Investigators select audio which are undergone pre-processed to adjust intensity and modulation. Proper validation is performed on the evaluation by the investigators. The recordings development block diagram in 3 shows the process involved in the development of the dataset. To explain the integrity of the dataset development the validation process is also included in the block diagram.

The audio signal obtained is passed after giving the labels into the Deep Learning Neural Network. As the primary aim of this research work is the identify the deception in the voice, the neural network classifies the incoming signal as deceptive or non deceptive. It was identified that Convolutional Neural Network can predict well with audio.

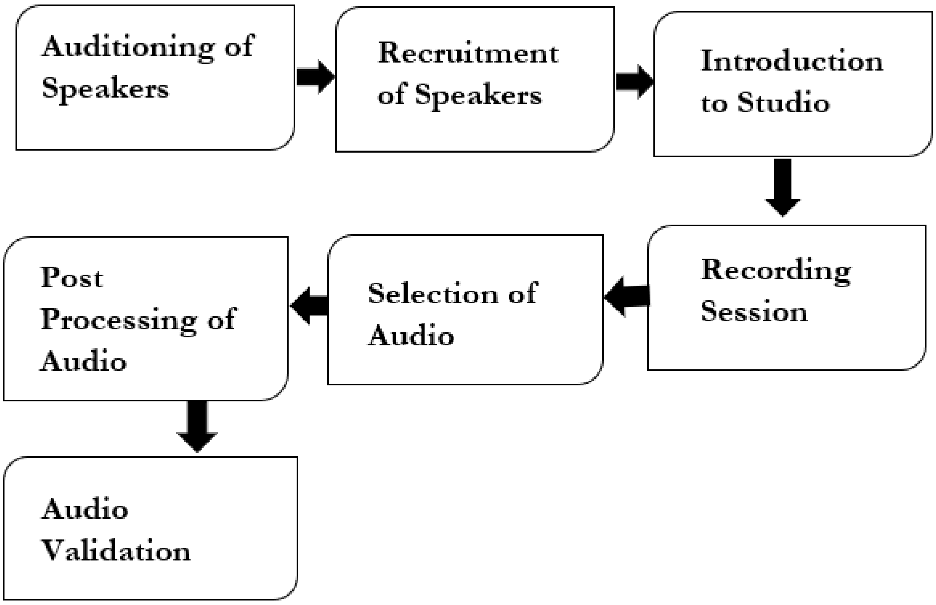


Figure 3: Recording Development

Hence, CNN is introduced into this. The labelled emotions are grouped as positively deceptive or negatively deceptive to identify the amount of deception in different emotions. The discussion on how this is analysed after proper design and implementation will be discussed in the upcoming sections.

3.2 Data Pre processing

After the identification of the dataset, the preprocessing tasks are performed. As the data is audio files in .wav form, the library librosa is used in the preprocessing with python as programming language. The preprocessing steps in the implementation of this work using audio are discussed below as mentioned in Alexandros Liapis and Voros (2021).

3.2.1 Audio Trimming

It was found that there is silence before and after the voice of the actor. It is necessary to trim these silence because the environmental effects in these portions can influence the classification. The library librosa provide trim function from effects package within it. The difference between the trimmed and original voice was noted later.

3.2.2 Spectrogram Plotting

The waveform obtained plotting the audio is called a spectrogram. The librosa allows to create the mel base spectrogram which is built on Mel Frequency Cepstral Coefficients. Rather than bare spectrogram, Mel spectrogram is preferred for classification. The plot will be discussed in the evaluation section.

3.2.3 White noise

In the preprocessing phase, white noise is added into the original audio to avoid the effect of random noises that occur in the surroundings.

3.2.4 Random Shifting

A function is developed to perform random shifting of the audio signal under consideration to allow extraction of synthetic data and thereby better model generalization.

3.2.5 Pitch and Speed Tuning

in order to describe the intensity of the audio, pitch function is implemented as well as for tuning the speed of the speeches of the actors, speed function is developed.

3.3 Data Transformation

The librosa library also gives way for identification of Mel Frequency Cepstral Coefficients (MFCC). This allows the calculation of the peak components in the audio. The delta function within the features of librosa allows an estimate of derivative of any source value.

3.4 Feature Extraction

The dataset used is converted into a dataframe for easy analysis with python. Pandas are used to create the dataframes. Librosa was used for loading the audio files. The load function will generate a floating point series for the incoming input signal. This allows the sampling of input to generate samples of desired values. These samples are used for obtaining MFCC using MFCC function available directly within the librosa. The demonstration of the power spectrum of the audio signal is using the Mel Frequency Cepstrum. The frequency components called Mel Frequency Cepstral Coefficients give a definite potrayal of the audio file which can be obtained theoretically using the 4

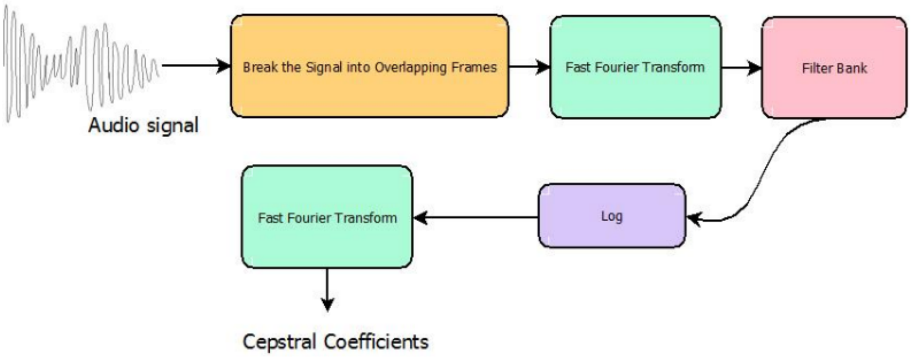


Figure 4: Mel-Frequency Cepstral Coefficients

3.5 Exploratory Data Analysis

The entire dataset taken from the directory initially is read into a dataframe created before with proper labels or headers. The labels were defined as per the emotions as the dataset had six different emotions in that. The labels are given on the basis of the sex of the actor and the emotion in the speech of the actor. The six different emotions are anger, sad, happy, fearful, surprised and neutral emotion or no emotion. The dataset distribution given in 5 shows the distribution of labels. It can be clearly seen that, as per the research objective, a clearly unbiased unisexual dataset is used for analysis of the deception in the audio. These are then labelled after combination to three different categories of interest as positively deceptive, negatively deceptive and neutral audio.

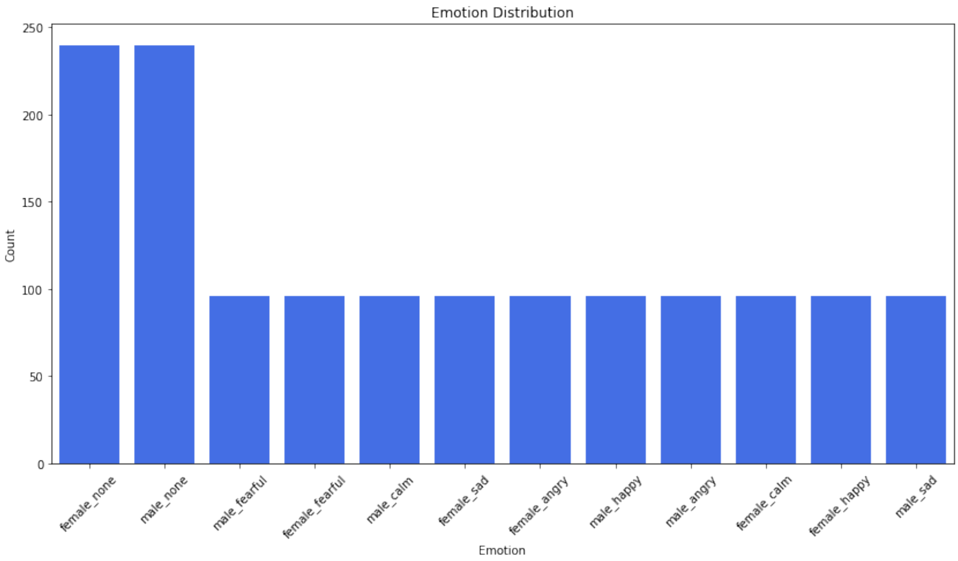


Figure 5: Dataset Distribution

The data labelled as three categories are then used for further evaluation. The categorisation is dependent on the labels in each of the audio files. In the figure6 given below, the waveform of the audio of a single actor is plotted to identify the pitch and other characteristics of the audio. This can be used to identify the frequency and amplitude of the audio signal.

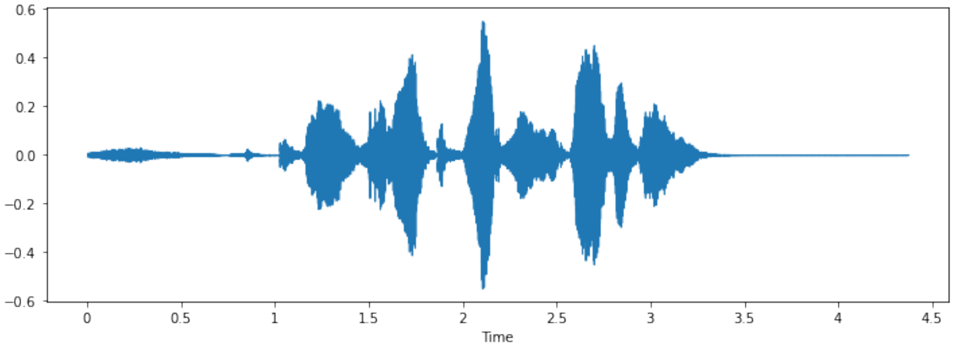


Figure 6: Waveform of Audio

4 Design Specification

The overall implementation of the deception analysis system architecture is divided into three different phases which is specified in detail in this section through 7.

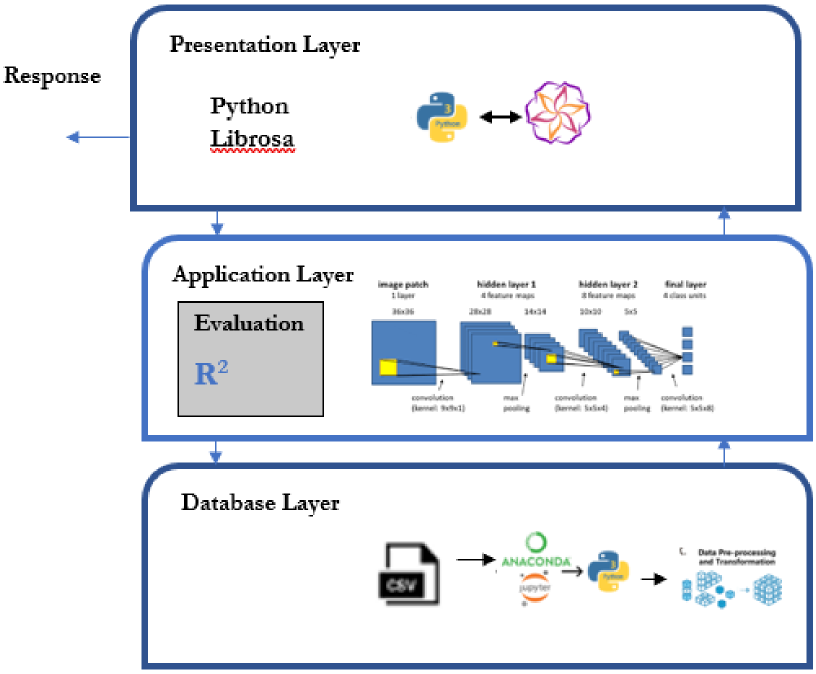


Figure 7: Design Specification

- Database Layer: The dataset from RAVDESS is downloaded from the open source Zenodo official site. The data was in the .wav form which is the common audio file format. A dataframe was created using Pandas of Python to load the data with sufficient headers. These were then preprocessed using Librosa to generate features from audio. The preprocessed dataframe is then further used for classification.

- Application Layer: The preprocessed data is then given to the convolutional neural network for classification. The research work is implemented in two ways. At the beginning the machine learning models like Decision Tree and Random Forest was implemented followed by implementation of neural networks. It was identified that neural networks gave better accuracy which was later implemented on the entire dataset. The confusion matrix is analysed in this work.

- Presentation Layer: The results obtained are evaluated by plotting the loss curve and then the confusion matrix. The classification of audio into deceptive and non deceptive can hence be used with original test audio to analyse the amount of deception in the audio signal.

The entire detailed architecture of the deception detection system model is discussed in the figure 15. The figure gives a wide idea on the steps followed in the analysis.

4.1 Data Interpretation and Analysis

As discussed, the raw audio is in .wav form, which is one of the file format of an audio. An Integrated Development Environment, Jupyter Notebook and libraries of Python such as Pandas and Librosa was used for analysing audio. Pandas convert the audio into dataframes for the efficient analysis. The representation of data into columns using the dataframes is the major highlight in using pandas for data interpretation. This can facilitate the quick access of each and every values in the data. The headers of the dataframe include the gender, the intensity of audio, the emotion and so on. In the filename of the dataset the emotions were specified using numerical values, which were clearly defined in the dataset description of RAVDESS. Hence the model developed will be a supervised learning model. If a model development uses both the input features as well as the specified output labels, then that particular model can be defined as supervised learning model. This particular model can develop predictions from the patterns detected from the explanatory variables and corresponding response variables.

As specified in the above section and in Cabrera and Lopez (2011) the input given to the CNN model will be Mel Frequency Spectral Coefficients with proper peak values of audio. In order to estimate the intensity, spectrograms were plotted in the beginning. Spectrograms are efficient in explaining the frequency spectrum of the audio against time, detailing the power of an audio. As the dataset includes multiple classes of emotions and this work aims on classifying deceptive and non deceptive audio, it is necessary to label the emotions into two or three classes as per the interest. Thus the obtained labels can be positively deceptive, negatively deceptive and neutral. As mentioned earlier,the six emotions in the dataset will be divided among three classes depending on their nature. Hence, a multiclass classification can be considered as binary or three class classification. All the classes are defined for the male and female subjects in the dataset. The strength of the audio signal need to carefully considered in the analysis as this work focus on the mental health of the human beings. To preserve the strength of the audio, it is important to perform certain preprocessing tasks. The preprocessing measures performed in this work include white noise addition, random shift of audio, voice stretch and speed as well as pitch tuning. In every model development, there need to be a training data as well as the test data. The data can be split using the available function in the sklearn machine learning library which allows unbiased split of labels. A Stratified Shuffle Split is performed for this. As per the research conducted and taking novelty into account for this audio dataset CNN model was identified as suitable for model development. But in order to compare this selection, two of the machine learning models discussed below were also implemented. These were implemented on the audio of a single actor to analyse the variation.

5 Implementation

5.1 Machine Learning Models

5.1.1 Decision Tree

DT is a non-parametric supervised learning model, that can be used for classification as well as regression problems. Decision Trees are hierarchical tree structures with leaf, branches and nodes. The hierarchy can be top down or bottom up approach to reduce

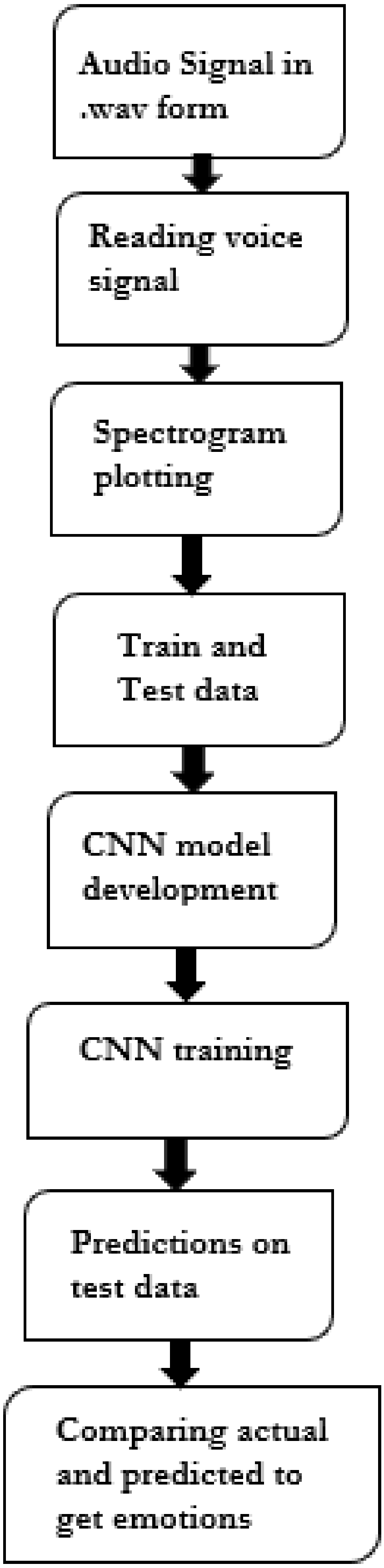
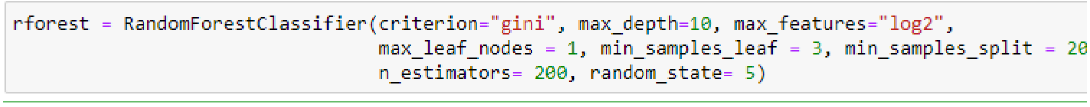


Figure 8: Architecture of the System

the root node to best possible leaf nodes as classes. This work used Sklearn's "DecisionTreeClassifier" library. There are parameters which need to be configured precisely to get the best possible classes.The work focus on implementing, 'max_depth' parameter as "10", allowing the tree to grow to ten branches. The 'max_features' parameter is "auto" to automatically make it number of inputs. The parameters like 'n_estimators', 'min_samples_leaf' and 'min_samples_split' are estimated using GridSearchCV, which is an optimization technique.

5.1.2 Random Forest

Random Forest is a machine learning model which can also be used for both classification and regression tasks. It works based on the concept of ensemble. Several weak models ensemble to form a strong learner in this model. "RandomForestRegressor" of sklearn is used for this work. There are parameters which need to be configured precisely to get the best possible classes in this also.The work focus on implementing, 'max_depth' parameter as "10", allowing the tree to grow to ten branches. The 'max_features' parameter is "auto" to automatically make it number of inputs. The parameters like 'n_estimators', 'min_samples_leaf' and 'min_samples_split' are estimated using GridSearchCV.



```
rforest = RandomForestClassifier(criterion="gini", max_depth=10, max_features="log2",
                                 max_leaf_nodes = 1, min_samples_leaf = 3, min_samples_split = 20
                                 n_estimators= 200, random_state= 5)
```

Figure 9: Random Forest model

5.2 Convolutional Neural Networks

Deep Neural Networks when combined becomes a Convolutional Neural Network. The convolutional neural networks are deep neural networks that works on the concept of convolution. The novelty of the work lies in incorporating Convolutional Neural Network into audio analysis for detecting deception where CNN was mostly used for image processing. This work focus on implementing a CNN using a unisexual dataset for the problem statement as in 10. Kernel is the important element of CNN that can give better convolution outputs with the incoming input audio. The sequential model of deep neural networks was used in this work for deception classification. Convolution Layer, Pooling layer were all added as per the requirement in the development of model meeting the convergence theory as mentioned in Hyewon Han and Kang (2018). The processes like data padding or striding were performed in the convolution layer in order to increase the power of the audio. In the pooling layer either min max or average pooling was performed. The next important aspect related to CNN are the use of activation function. This research incorporates the use of Relu activation function although there are several commonly available activation functions.

6 Evaluation

As per the discussion above after incorporation of sufficient number of layers into the sequential model, predictions can be obtained on any incoming data in acceptable format

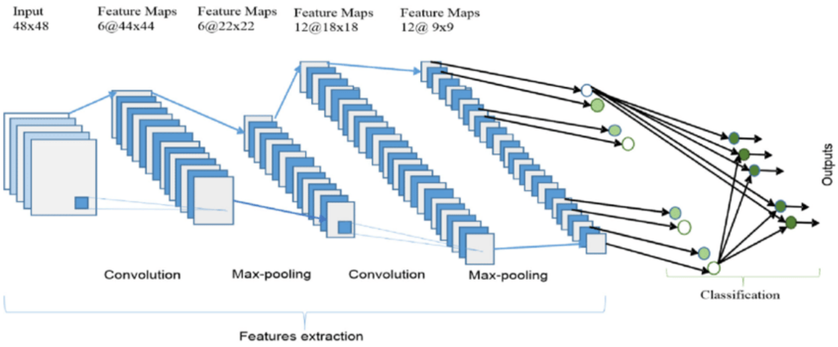


Figure 10: Architecture of CNN

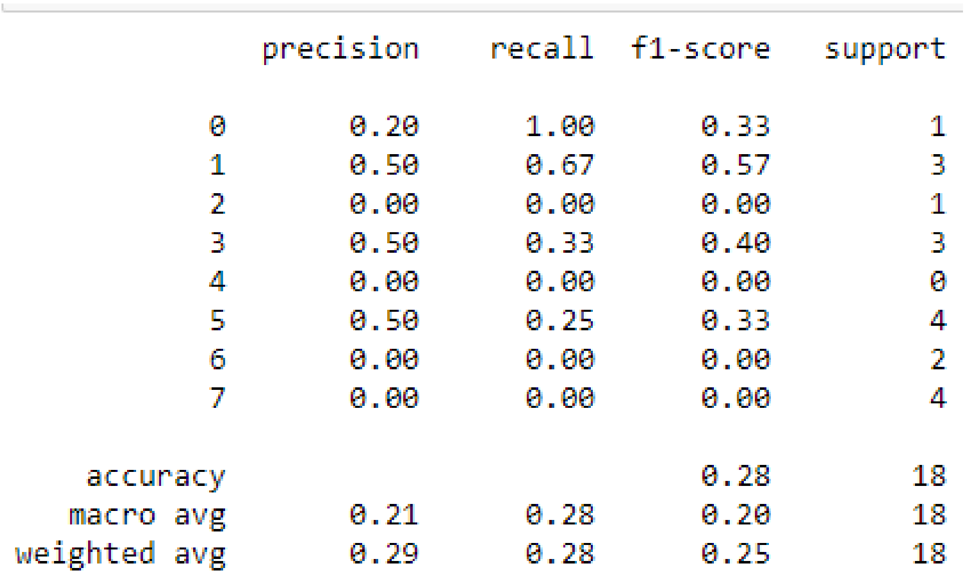
after proper training, testing and validation. Once the model is developed using the training data, test data can be given to classify the emotions as deceptive or non deceptive. In order to find the accuracy and integrity of the model, the true values and predicted values can be compared using the evaluation metrics acceptable by the discontinuous data. As mentioned in the above section the analysis was performed in two ways: using a single actor audio data and using all 24 actor audio. In the first analysis, Decision Tree and Random Forest classifier was implemented as well as the CNN model. But for all the actors data, the neural network was only implemented keeping the accuracy in mind.

6.1 Decision Tree Model

This model was implemented on the data of 6 emotions of a single actor. The classification of the audio into deceptive or not was performed using decision tree classifier model. It is found in figure given below that model performs very poorly for all the classes. The accuracy of the overall model was not acceptable. As the audio signal is discontinuous, classification metrics are the only evaluation method available for the same.

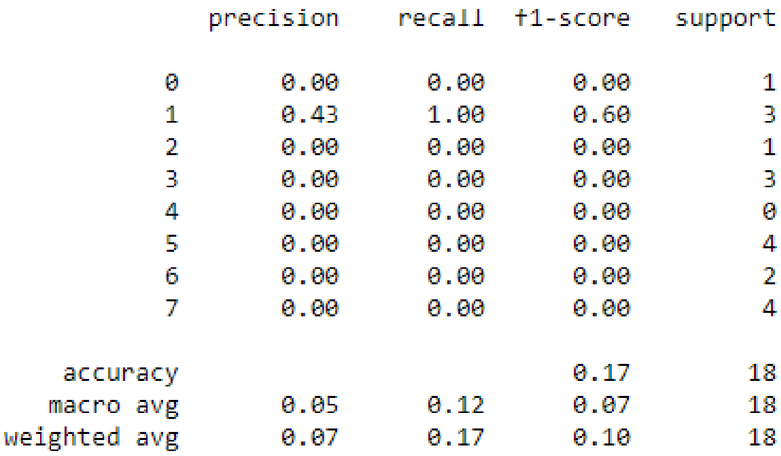
6.2 Random Forest Model

The implementation of the random forest model on the audio data of a single actor performed similar to decision tree classifier. This shown in the figure given below. The model doesn't perform well with this data. So this can be a lead to the development of CNN model in order to check any increase in the overall accuracy of classification. In order to evaluate the model performance, this work uses the performance metrics like accuracy, precision, recall and f1 score. Confusion matrix is the metric that can give all the values for the true positives, true negatives, false positives and false negatives. All the combination of these values form the confusion matrix. A better model will always have low false positives and high true negatives. Another important parameter is the loss curve which gives a comparison between the training loss and test loss. It was observed that the training loss and test loss decreased as the number of epochs increased. The accuracy of a good model is expected to lie above 70% where the classification is expected to be acceptable.



```
              precision    recall  f1-score   support

           0       0.20      1.00      0.33         1
           1       0.50      0.67      0.57         3
           2       0.00      0.00      0.00         1
           3       0.50      0.33      0.40         3
           4       0.00      0.00      0.00         0
           5       0.50      0.25      0.33         4
           6       0.00      0.00      0.00         2
           7       0.00      0.00      0.00         4

    accuracy                           0.28        18
   macro avg       0.21      0.28      0.20        18
weighted avg       0.29      0.28      0.25        18
```

Figure 11: Decision Tree Model Performance Metrics

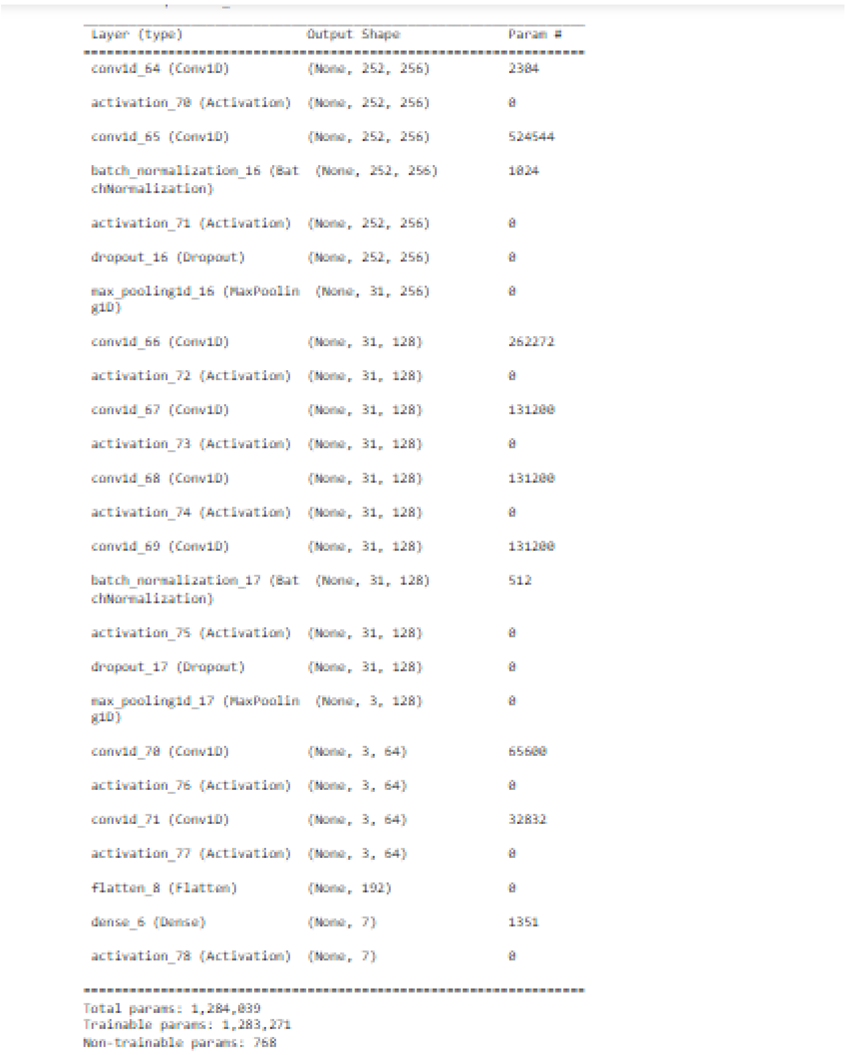


```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         1
           1       0.43      1.00      0.60         3
           2       0.00      0.00      0.00         1
           3       0.00      0.00      0.00         3
           4       0.00      0.00      0.00         0
           5       0.00      0.00      0.00         4
           6       0.00      0.00      0.00         2
           7       0.00      0.00      0.00         4

    accuracy                           0.17        18
   macro avg       0.05      0.12      0.07        18
weighted avg       0.07      0.17      0.10        18
```

Figure 12: Random Forest Model Performance Metrics

6.3 Convolutional Neural Network Model

The convolutional neural network built with the data input as all the audio of 24 actors is given in figure shown in 13. The CNN had different layers including convolution layer, activation layer, dense layer, max pooling layer, dropout layer. Depending on the convergence of the classes as per the multi class classification, these layers are added to neural network to get the classes accurately. As the positive, negative and neutral deception are predicted for male and female, there was 7 different target classes which was allocated in the dense layer.



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d_64 (Conv1D)          (None, 252, 256)          2304

 activation_70 (Activation)  (None, 252, 256)          0

 conv1d_65 (Conv1D)          (None, 252, 256)          524544

 batch_normalization_16 (Bat  (None, 252, 256)         1024
 chNormalization)

 activation_71 (Activation)  (None, 252, 256)          0

 dropout_16 (Dropout)        (None, 252, 256)          0

 max_pooling1d_16 (MaxPoolin  (None, 31, 256)          0
 g1D)

 conv1d_66 (Conv1D)          (None, 31, 128)           262272

 activation_72 (Activation)  (None, 31, 128)           0

 conv1d_67 (Conv1D)          (None, 31, 128)           131200

 activation_73 (Activation)  (None, 31, 128)           0

 conv1d_68 (Conv1D)          (None, 31, 128)           131200

 activation_74 (Activation)  (None, 31, 128)           0

 conv1d_69 (Conv1D)          (None, 31, 128)           131200

 batch_normalization_17 (Bat  (None, 31, 128)          512
 chNormalization)

 activation_75 (Activation)  (None, 31, 128)           0

 dropout_17 (Dropout)        (None, 31, 128)           0

 max_pooling1d_17 (MaxPoolin  (None, 3, 128)           0
 g1D)

 conv1d_70 (Conv1D)          (None, 3, 64)             65600

 activation_76 (Activation)  (None, 3, 64)             0

 conv1d_71 (Conv1D)          (None, 3, 64)             32832

 activation_77 (Activation)  (None, 3, 64)             0

 flatten_8 (Flatten)         (None, 192)               0

 dense_6 (Dense)             (None, 7)                 1351

 activation_78 (Activation)  (None, 7)                 0

=================================================================
Total params: 1,284,039
Trainable params: 1,283,271
Non-trainable params: 768
```

Figure 13: CNN Layers

The figure shown in 14 gives the integrity of the model. It is clearly visible from the plot that the training and test loss are decreasing continuously as the number of epochs are increasing. The loss of training data is seen to be decreasing steeply whereas the loss of the test data is fluctuating but eventually decreasing continuously. As the loss is decreasing, the model can be accepted as a feasible one.

The figure given below shows the confusion matrix developed from the deception analysis of RAVDESS dataset. The parameters such as accuracy, precision, recall and f1 score are obtained from the confusion matrix. The accuracy of the model lies as 73.25%,

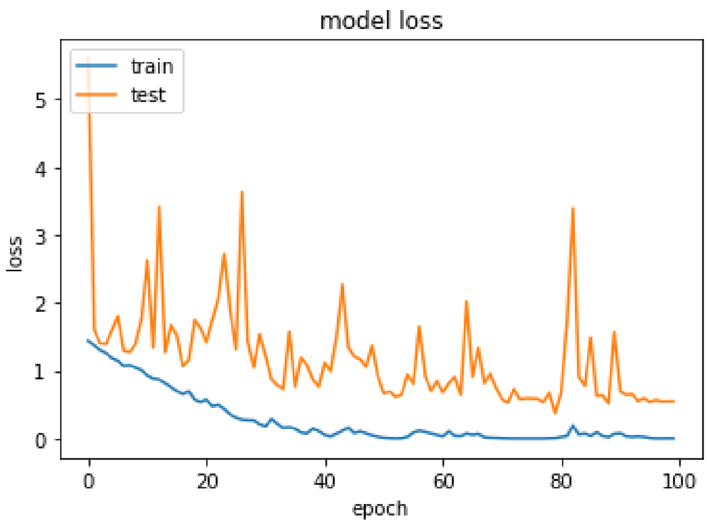


Figure 14: Loss Curve

precision as 87.42%, recall as 73.25 and f1 score as 75.1%. As the work focus on identifying all the deceptive audio, the parameters like precision and recall are taken into account.

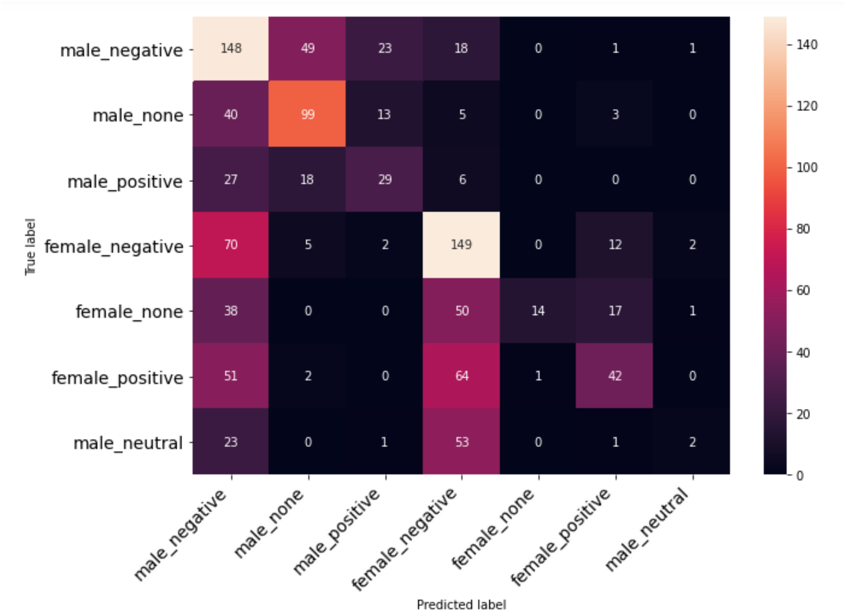


Figure 15: Confusion Matrix

7 Discussion

The research work focused on implementing a novel deep learning model to identify the deception in the audio of human beings. It was identified from the literature survey that no solution exists in classifying the audio as deceptive or not. The novelty lies in the usage of a unisexual unbiased dataset for the classification upon the Convolutional Neural

Network. RAVDESS dataset considered satisfies all the requirements in the objectives followed by CNN implementation. As the audio is discontinuous, the parameters from confusion matrix like precision and recall was taken into consideration. It was observed that the precision of the model was 87.42% which was much better than the machine learning models like Decision Tree and Random Forest developed for the same. Hence the selection of Convolutional Neural Network over Decision Tree and Random Forest can be justified. It was also observed that the use of unisexual unbiased dataset provided a way for the elimination of overfitted biased output. Also analysing six different emotions was a credible concept as all of those could be incorporated together to give classification on deceptive or non deceptive.

As the mental health of this generation is highly getting affected, an Artificial Intelligence model like this can pave a way for healthy mental life in future. Hence this model with the establishment of novelty can be a good approach in classifying upcoming voice signal as deceptive or not. This can hence definitely be made a viable source of mental health in the field of medical science. Therefore relevant insights can be taken forward in analysing the audio signal from this work.

8 Conclusion and Future Work

The aim of this research work was to implement a Deception Analysis model using a Deep Neural Network, the Convolutional Neural Network. The literature survey conducted at the beginning explains the need for some criterion to establish the work with atmost novelty. The criterion include considerably good sample size, audio dependent on gender, distinct emotions of a single subject. The work focused on implementing these criterion and RAVDESS dataset was identified at the initial stages of implementation which satisfied the criterion related to data. The implementation of CNN on such a data is not attempted till now. The dataset included six unrelated emotions of 24 actors, 12 male and 12 female, at two levels of intensity. The emotions include anger, sad, fearful, happy, surprised and neutral. These emotions are then grouped and labelled as positive, negative and neutral deception audio. The emotions like happy, surprised are labelled as positive, fearful angry and sad as negative deception and neutral and neutral deception. Thus three classes are normally predicted. If there is need of gender based deception classification 6 classes will be developed. This can hence state the mental well being of a patient and hence is a great contribution to the field of medical science.

In this twenty first century, each and every action of humans are in hand of machines. Hence, a model like this is expected to have a substantial impact on the health. This model is expected to give better predictions with the disease prediction with either symptoms or visual representations. This model can be extended to work with much more emotions so that a genuine and accurate prediction can be obtained. As the RAVDESS data considered is very unique and feasible, the model developed with Convolutional Neural Network because of its versatility is showing a good precision and recall with low loss.

9 Acknowledgement

The author would like to thank Mr. Jorge Basilio for his immense support and assistance in the implementation of the research work. His proficiency helped in finishing both the

technical part and report fronts.

References

A.Baum (1990). Stress, intrusive imagery, and chronic distress, *Health psychology* .

Alexandros Liapis, Evanthia Faliagka, C. P. A. G. K. and Voros, N. (2021). Advancing stress detection methodology with deep learning techniques targeting ux evaluation in aal scenarios: Applying embeddings for categorical variables, *MDPI, Electronics 2021* .

Anna Esposito, Gennaro Raimo, M. M. C. V. M. C. and Cordasco, G. (2020). Behavioral sentiment analysis of depressive states, *11th IEEE International Conference on Cognitive Infocommunications* .

Anto, B. and Raji, F. S. (2009). Automatic stress detection from speech by using discrete wavelet transforms, *ITBI 09* .

Arushi, R. D. and Teoh, A. N. (2021). Real-time stress detection model and voice analysis: An integrated vr-based game for training public speaking skills, *IEEE Conference on Games* .

Baek, J. W. and Chung, K. (2020). Context deep neural network model for predicting depression risk using multiple regression, *Special Section On Machine Learning Designs, Implementations And Techniques, IEEE ACCESS* .

Barlian Henryranu Prasetio, H. T. and Tanno, K. (2020). Embedded discriminant analysis based speech activity detection for unsupervised stress speech clustering, *IEEE* .

Cabrera, R. and Lopez, D. B. (2011). Voice stress detection: A method for stress analysis detecting fluctuations on lippold microtremor spectrum using fft, *CONIELECOMP 2011, 21st International Conference on Electrical Communications and Computers* .

Cristina Luna Jimenez, David Griol, Z. C. R. K. J. M. M. and Martinez, F. F. (2021a). Multimodal emotion recognition on ravdess dataset using transfer learning, *MDPI, Sensors 2021* .

Cristina Luna Jimenez, Ricardo Kleinlein, D. G. Z. C. J. M. M. and Martinez, F. F. (2021b). A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset, *MDPI, Applied Sciences 2021* .

Dipanshu Someshwar, Dharmik Bhanushali, S. N. and Chaudhari, V. (2020). Implementation of virtual assistant with sign language using deep learning and tensorflow, *Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020)* .

Felipe Mateus Marcolla, R. S. and Dazzi, R. L. S. (2017). Novel lie speech classification by using voice stress, *In Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020)* pp. 742–749.

Gedeon, T. and N.Sharma (2012). Objective measures, sensors and computational techniques for stress recognition and classification: A survey, *Computer methods and programs in biomedicine* .

Hyewon Han, K. B. and Kang, H.-G. (2018). A deep learning-based stress detection algorithm with speech signal, *Association for Computing Machinery* .

Kevin Tomba, Joel Dumoulin, E. M. O. A. K. and Hawila, S. (2018). Stress detection through speech analysis, *In Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE 2018)* pp. 394–398.

Kinariwala, S. (2019). Detection and analysis of stress using machine learning techniques, *International Journal of Engineering and Advanced Technology (IJEAT)* .

Lakmal Rupasinghe, Alahendra, R. P. and Kulathunge (2021). Robust speech analysis framework using cnn, *3rd International Conference on Advancements in Computing (ICAC)* .

Palacios Alonso, Nieto Lluis, R. B. (2015). Analysis of emotional stress in voice for deception detection, *2015 4th International Work Conference on Bioinspired Intelligence (IWOBI)* .

Prashant, M. T. (2019). Human emotion detection and stress analysis using eeg signal, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* **8**.

Savita Sondhi, Ritu Vijay, M. K. and Salhan, A. (2016). Voice analysis for detection of deception, *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)* .

Soham Dasgupta, K. and Masunda, S. (2017). Voiceprint analysis for parkinson's disease using mfcc, gmm, and instance based learning and multilayer perceptron, *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI-2017)* .

Vaikole, Mulajkar, M. J. D. (2020). Stress detection through speech analysis using machine learning, *International Journal of Creative Research Thoughts (IJCRT)* .