Depression Detection through Audio Analysis using Machine Learning Models Ensuring Sustainable Development of Mankind

Prathika Yadav[1], Pooja Jain[1] and Tapan Jain[1]

[1] *Indian Institute of Information Technology, Nagpur, India*

Abstract

Depression is one of the biggest issues of the world today, it affects an individual's quality of life to a considerable extent. In this study we have examined the use of ML (Machine Learning) models and their performance in detection of depression through audio data from a single data source namely the DAIC-WOZ. We extracted the features from audio/voice recordings of patients and trained several different models. The results of our study show that several models can achieve high accuracy in predicting depression levels. Future research could explore the potential of integration of multiple modalities and deep learning approaches to improve the accuracy of depression detection. Overall this study demonstrated that machine learning models have great potential for depression detection using audio data, which requires further research to be validated.

Keywords

Depression Detection, Machine Learning, Sustainable Development, Depression Prediction, Model Comparison

1. Introduction

Depression is a prevalent mental health condition that affects millions of people across the globe, it is marked by a constant feeling of sadness and/or lack of interest in activities that were once pleasurable to a particular person, usually lasting for a long period of time. [1] One of the effective ways to deal with depression is to detect early on in the journey of a person to prevent long-term negative outcomes such as chronic disability and suicide. [2] But in today's world the traditional methods of depression diagnosis, primarily, self-reporting and clinical assessments are very subjective and various factors can lead to or cause depression, this includes social desirability bias and differences in interpretation of the respective symptoms. [3]

The recent advances in machine learning and audio analysis have opened doors to new methods of objective and non-invasive techniques of depression detection. Audio data, including speech and voice patterns, have shown promising capabilities as a modality for detecting depression. This can help with early detection of depression since even individuals without severe symptoms can exhibit acoustic patterns that can be precursors to depression. [4]

Research has shown that Depression is more commonly diagnosed in women than in men, reporting around twice the prevalence in females. Depression ranks among the primary contributors to the burden of disease in women. [5]

According to Economic Times Article [12]. India has 7.5 psychiatrists per million people. Additionally, a survey conducted on mental health across 12 Indian states indicates that there is a significant treatment gap ranging from 70 to 92% for various mental disorders in these regions. Depression is a widespread mental health disorder that impacts millions of people globally. However, conventional methods of diagnosing depression, such as clinical interviews and self-

AISD 2023: First International Workshop on Artificial Intelligence: Empowering Sustainable Development, September 4-5, 2023, co-located with First International Conference on Artificial Intelligence: Towards Sustainable Intelligence (AI4S-2023), Pune, India

dprathikayadav@gmail.com (P. Yadav); poojajain@iiitn.ac.in (P. Jain); tapankumarjain@gmail.com (T. Jain)

https://dblp.org/pid/29/5985.html (P. Jain)

report questionnaires, may have limited accuracy due to their subjective nature. In recent years, audio data, including speech and voice patterns, have been studied as a promising source of information for detecting depression. Distinct patterns in acoustic features such as pitch, intensity, and speech rate have been identified between individuals with depression and healthy controls.

2. Related Work

Depression is a widespread mental health disorder that impacts millions of people globally. However, conventional methods of diagnosing depression, such as clinical interviews and self-report questionnaires, may have limited accuracy due to their subjective nature. In recent years, audio data, including speech and voice patterns, have been studied as a promising source of information for detecting depression. Distinct patterns in acoustic features such as pitch, intensity, and speech rate have been identified between individuals with depression and healthy controls.

Nicholas Cummins et al., [6] is a step towards considering speech as a key objective marker to aid clinical assessment by reviewing the characteristics of patients with depression and suicidal thoughts, including their size, associated clinical scores, and data collection methods, are significant factors in the development of prediction and classification systems for these conditions. The study utilized spectral features, such as Power Spectra Density (PSD) and Mel Cepstral Features (MFCCs), as part of its methodology.

Research in this area was conducted by Laura Verde et al. (2016) [7], who explored the use of speech features extracted from audio recordings to predict depression. The study used features such as pitch, intensity, and spectral entropy and achieved an accuracy of about 85%.

In another study, Emna Rejaibi et al.(2021) [8], the authors of the study examined the potential of Mel-frequency cepstral coefficients (MFCCs), extracted from audio recordings, in predicting depression. They applied various machine learning algorithms, such as support vector machines (SVMs) and random forests, and attained a 72% accuracy rate. Moreover, their MFCC-based recurrent neural network (RNN) achieved an overall validation accuracy of 76.27%.

Ah Young Kim et al. [13], have conducted a study in which they have built a conventional machine-learning model that uses Log-Mel Spectrogram and deep convolutional neural network(CNN). They have focused on using acoustic features and have achieved an accuracy of 78.14% which showed that deep-learned acoustic characteristics can be an approach to automated depression detection.

The study conducted by Hande Kaymaz Keskinpala et al., [14] demonstrated that Mel-cepstral coefficients and energy in frequency bands can serve as a means of distinguishing between depressed and suicidal patients, in which they used a different number of cepstral coefficients and were compared by using unimodal Gaussian modeling. They have performed it on 2 kinds of speech samples, interview sessions, and reading sessions. The conclusion drawn from the study indicates that controlled reading has the potential to offer better results in comparison to interviews.

In a study by Sharifa Alghowinem et al., [15] involved a comparison of the efficacy of various acoustic and prosodic features when used with different classifiers. In the classifiers used, GMM, SVM, MLP, and HFS), they also investigated the classifiers using GMM as input and observed that in the hybrid classifier, the best combination was then GMM was used with SVM. The conclusion about the features that performed best in the detection of depression is Loudness, root mean square, and intensity.

3. Data

Data from DAIC-WOZ Depression database is used for our study, it's a part of a larger corpus, Distress Analysis Interview Corpus(DAIC)(Gratch et al, 2014)[9], that contains clinical interviews designed to support the diagnosis of psychological conditions including depression. The

interviews were collected so as to create a computer agent that interviews people and identifies verbal and non-verbal indicators of mental illness(Devault et al., 2014)[10].

Data collected includes audio along with video recordings and the responses to an extensive questionnaire. The interview was conducted by a virtual interviewer, Ellie; alongside a human interviewer in another room. The dataset consists of a CSV file that contains binary labels on whether the subject is depressed(1) or the subject is not depressed(0). The labels are given based on PHQ score that is calculated based on the overall interview. The Patient Health Questionnaire (PHQ-8) is a widely accepted depression screening tool that consists of 8 questions aligned with the major diagnostic criteria for Major Depressive Disorder as defined in the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). Each question is scored from 0 to 3, for a total possible score ranging from 0 to 24. Studies have found the PHQ-8, which relies on self-reported symptoms, to be both reliable and valid for gauging an individual's depression severity, making it a commonly used depression measure in both clinical and research contexts. In our particular study, we considered the PHQ-8 scores as the gold standard for assessing depression severity within the DAIC-WOZ dataset.

4. Data Exploration

In the process of exploring the data, a correlation matrix was conducted to gain insights into the relationship between different attributes and depression. The findings revealed that gender exhibited a cooler tone in comparison to the other attributes, which had warmer tones. it may be beneficial to explore the potential gender differences in the acoustic features of speech related to depression.

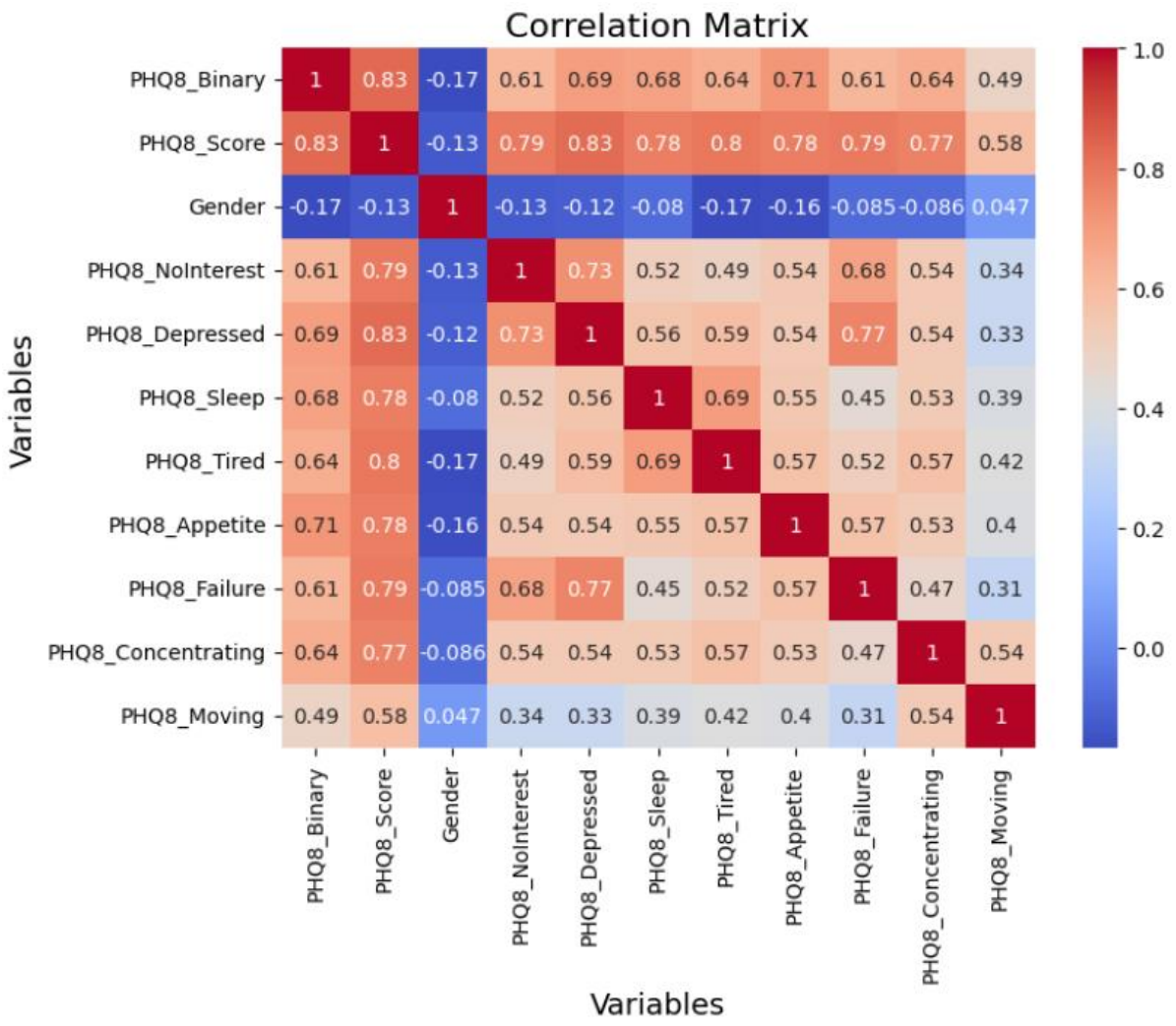

Figure 1: Correlation Matrix

A graph was plotted to understand how the data points are distributed with respect to gender, and whether or not the individual is depressed.

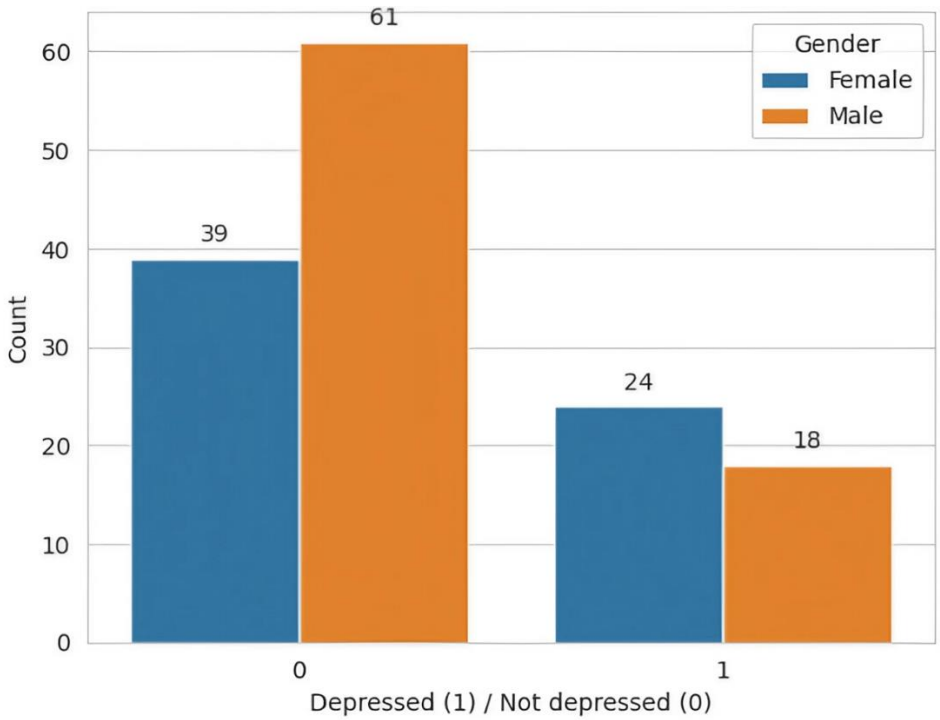

Figure 2: Gender Distribution of Drepressed and Non-Depressed Induvisuals.

The data shows that there are more females with depression than males, with 38.7% of females and 22.8% of males being depressed. These findings suggest that there is a class simbalance and can lead to biased model performance, where the model is more likely to predict the majority class.

5. Data Processing

The Data includes transcript files, which have information about who the speaker is at a given particular time. The transcript includes timestamps and speaker information which can be used for the extraction of the subject audio from the raw audio of the interview. The raw audio file, using the transcript file can be trimmed and concatenated into a single audio file on which the Feature extraction can be performed.

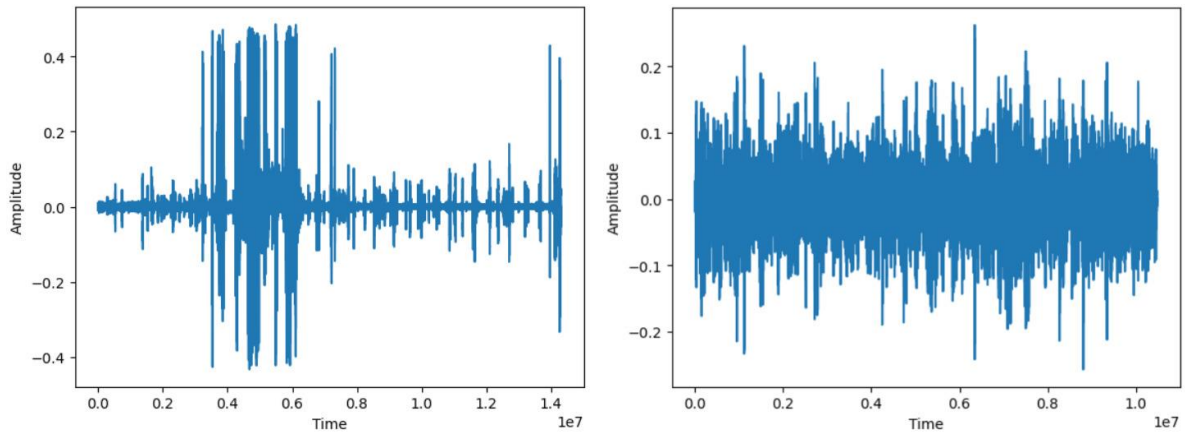

Figure 3: Visual Representation of Raw and Extracted Subject's Audio.

The imbalanced distribution of classes in datasets can significantly affect the performance of machine learning models. In the context of audio data, a lack of depressed individuals in the dataset can result in poor performance when trying to identify and classify depressed speech. To

address this issue, data augmentation techniques have been employed to increase the diversity and quantity of the data. Specifically, time-stretching has been used to randomly adjust the speed rate of the audio recording, creating new, altered versions of the original data. The process involves loading the audio file, applying the time-stretching effect, and saving the new audio file to be used for feature extraction. By increasing the quantity of the depressed audio samples through data augmentation, the accuracy of the machine learning models can be improved.

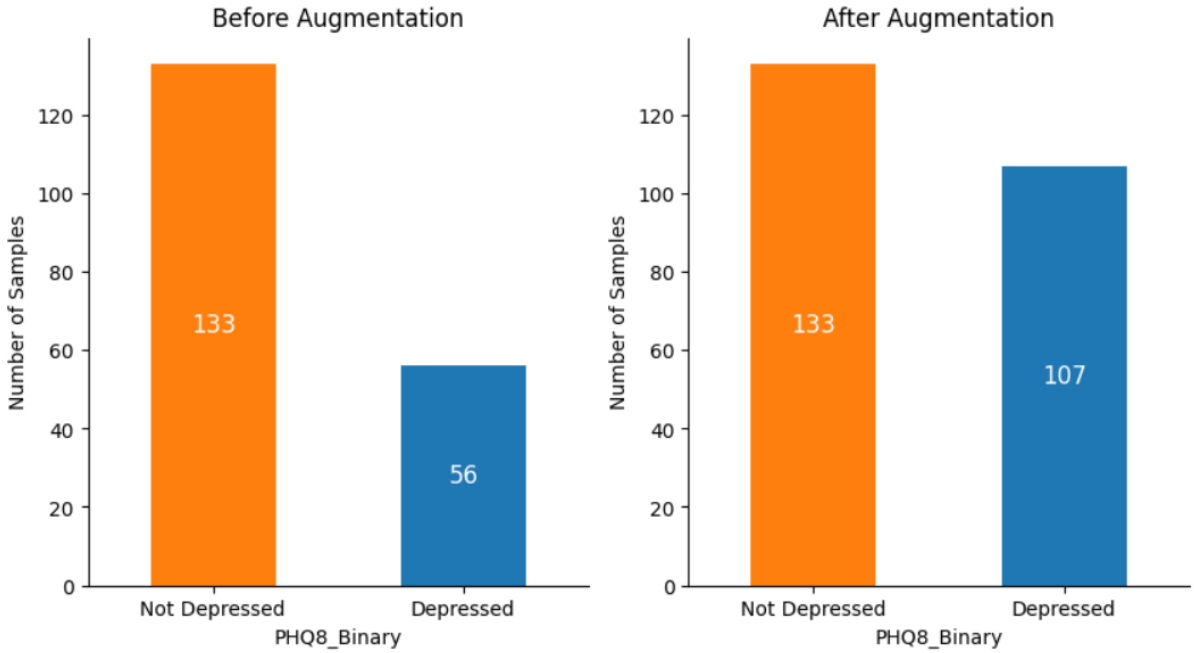

Figure 4: Visual Representation of the depressed and Non-depressed patients before and after data augmentation.

6. Feature Extraction

How features are identified within data is important for machine learning models that identify things like depression. We can extract useful features from the raw data to train their models. This "feature extraction" helps simplify complex data sets into more manageable inputs. For detecting depression, analyzing audio data for meaningful features provides insights into a person's emotions and speech patterns. Certain features in someone's voice may signal depression. By identifying these important characteristics, machine learning models can recognize patterns that accurately predict a person's mental health. By extracting meaningful information from audio recordings using these feature extraction techniques: We can analyze aspects like a person's speech pace, tone and inflection, Look at use of certain words and phrases Identify long pauses and moments of silence. Other vocal qualities that give clues about a person's emotional state, Through spotting relevant details in the audio, the feature extraction helps the machine learning model "learn" what to listen for to identify depression. The right features improve the model's ability to make accurate assessments of a person's mental health. In this study we have leveraged the following feature extraction techniques to obtain meaningful information from the audio data:

1. Mel Spectrogram: Mel spectrogram is a way to represent data in terms of a power spectrum of a signal in the frequency domain. It gives us information about the distribution of the frequency of an audio signal, with the spectral energies mapped to a mel-scale, which is a non-linear scale that is basically an approximation of the auditory response of humans mapped to different frequencies. In the context of depression detection, the changes in frequency distribution of speech can be found to be associated with depression and other mental health conditions. Hence, the mel spectrogram can give us insights into the spectral

characteristics of the speech signal and help us identify relevant features that can be indicative of depression.

2. Log Mel Spectrogram: The log mel spectrogram method consists of finding the logarithm of the mel spectrogram. This algorithm is used to compress the dynamic range of the spectrogram making it easier to understand and analyze.

3. Fundamental Frequency (F0): Fundamental frequency, also known as pitch, refers to the perceived frequency of a periodic waveform. In the domain of speech, it corresponds to the fundamental frequency of the vibration of our vocal cords'. F0 provides us information about the prosodic features of speech, such as the intonation. Rhythm and stress of the speech. These are known to be associated with emotional states and can help us dig deeper for insights into the speaker's emotional state.

4. Spectral Contrast: Spectral contrast is used to measure the spectral shape of a signal that captures the difference between the energy in different frequency bands. It provides insights regarding the spectral characteristics of the audio signal, such as the presence of harmonics and formants. Changes in the spectral contrast of speech have been found to be associated with emotional states and can provide insights into a speaker's emotional state.

5. Recurrence matrix: A recurrence matrix is a binary matrix used in audio processing to identify repeating patterns in the sound. It compares each time point in the signal to all other time points, with a threshold applied to determine whether the points are similar enough to be considered a "recurrence" event. Recurrence matrices are useful in depression detection as they can identify speech patterns indicative of depression, such as repetitive speech patterns and reduced vocabulary.

In summary, these features provide information about the spectral and prosodic characteristics of the speech signal, which have been found to be associated with emotional states such as depression. Extracting these features can provide insight into the speaker's emotional state and aid in depression detection.

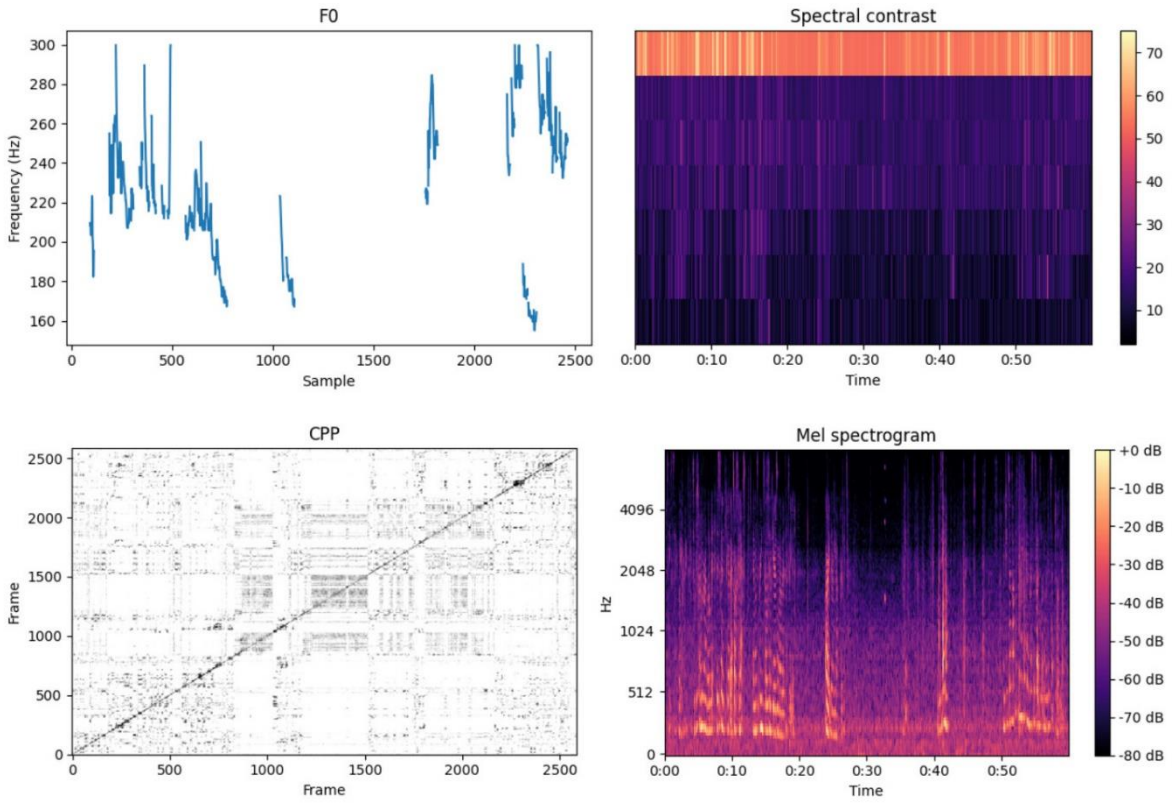

Figure 5: Visual Representation of the features extracted.

7. Implementation Details

The first step in our approach was to obtain the raw audio data and preprocess it to extract the relevant speech segments using start and end times as well as speaker information. Next, we applied data augmentation techniques, specifically time stretching, to increase the size of the dataset and improve the robustness of the model. Feature extraction was then performed on the processed and augmented audio files using a range of techniques, including Mel-frequency cepstral coefficients (MFCCs), fundamental frequency (f0), spectral contrast, and recurrence matrix. This resulted in a set of feature vectors for each audio file, which were then used to train and test various classifiers. In our study, we conducted training and evaluation using a labeled dataset with five classification models: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, and Gradient Boosting. These models were carefully chosen for their relevance and efficacy in addressing the classification task. To evaluate the performance of these models, we utilized the F1 score as the evaluation metric. The F1 score considers both precision and recall, providing a balanced measure of the model's accuracy in correctly identifying positive instances and effectively handling class imbalance. By training and evaluating these models using the F1 score, we gained valuable insights into their classification performance and suitability for the given task.

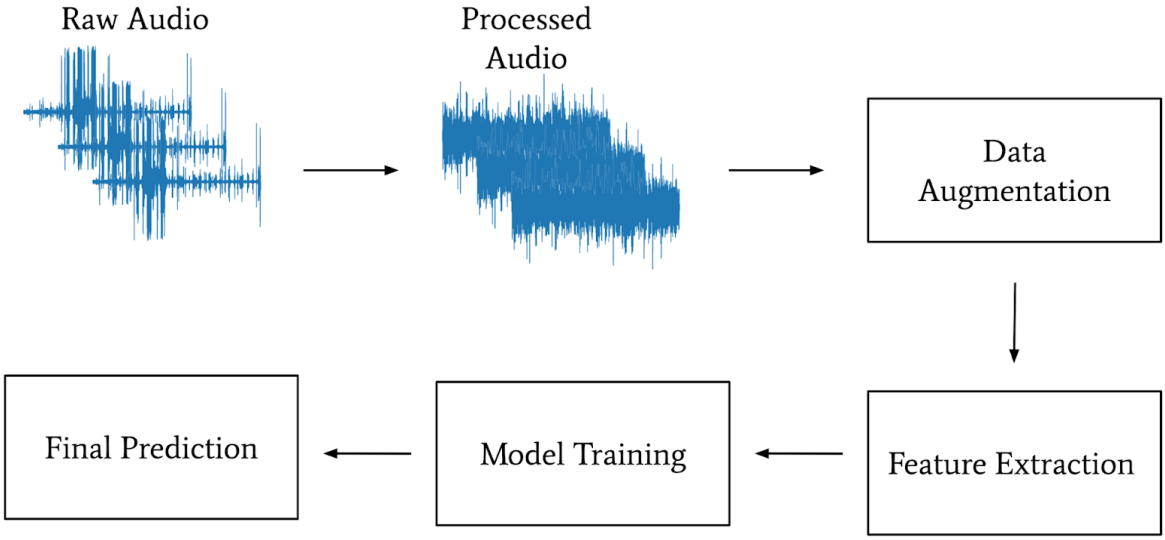

Figure 6: Overall approach for depression detection.

8. Results

The results of the implemented approach for automatic depression classification are presented in this section. The evaluation of the classification models was conducted using the F1 score, which provides a balanced measure of accuracy by considering both precision and recall. The models were trained and evaluated on a labeled dataset consisting of male and female participants. The below table displays the F1 scores achieved by each classification model.

Table 1
F1 Score

Model	F1 Score
SVM	0.71
Random Forest	0.79
Logistic Regression	0.73
Decision Tree	0.82
Gradient Boosting	0.85

Amongst the models we evaluated, the Decision Tree and Gradient Boosting models in particular exhibited the highest of the F1 scores, achieving scores of 0.82 and 0.85, respectively. Additionally, a gender-specific analysis was performed to investigate the impact of gender separation on the classification results. The Gradient Boosting model achieved the highest overall F1 score and it was further evaluated on separate male and female subsets of the dataset. The final F1 scores obtained were 0.886 for females and 0.865 for males. A comparison of gender-specific F1 scores highlights the effects of considering gender-related differences in depression as a condition and its classification. The model resulted in a higher F1 score when trained and tested on the female subset than the male subset. This finding suggests that an approach of gender-specific classification can contribute to the improved accuracy in identifying depression, as it would consider the unique speech patterns and expressions associated with depression in different genders. Overall, the results demonstrate the effectiveness of the Decision Tree and Gradient Boosting models in depression classification. Moreover, the findings underscore the significance of considering gender-related differences, as the gender-specific analysis revealed improved performance when addressing the unique characteristics of each gender.

9. Conclusion

In this study, we proposed an approach for the automatic classification of depression states using speech-based features. The methodology we implemented involved preprocessing the raw audio data to extract relevant speech segments. Data augmentation techniques, specifically time stretching, were applied to increase the dataset size and enhance the robustness of the model. Feature extraction was performed using a combination of techniques, including Mel Frequency cepstral coefficients (MFCCs), fundamental frequency (f0), spectral contrast, and recurrence matrix. To evaluate the classification performance, we trained and tested on different classification models: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision Tree, and Gradient Boosting. The evaluation was conducted using the F1 score, which considers both precision and recall, providing a balanced measure of accuracy. Among the models evaluated, the Decision Tree and Gradient Boosting models demonstrated the highest F1 scores, achieving 0.82 and 0.85, respectively, indicating their effectiveness in accurately identifying instances of depression using speech-based features. Additionally, we have explored the impact of gender separation on the classification performance. The Gradient Boosting model, which achieved the highest overall F1 score, was evaluated further on separate male and female subsets of the original dataset. The results showed that considering gender-specific characteristics and patterns in depression classification led to improved performance of detection. The model exhibited a higher F1 score when it was trained and tested on the female subset compared to the male subset, highlighting the importance of accounting for gender-related differences. In conclusion, this study highlights the potential of speech-based features in the classification of depression. The Decision Tree and Gradient Boosting models showcased very promising results, outperforming most of the other classification models. Moreover, incorporating gender-specific analysis enhanced the classification accuracy of the overall system, emphasizing the significance of tailoring the model training and evaluating it with respect to specific genders present in the dataset.

References

[1] Depressive disorder (depression) (no date) World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/depression

[2] “Practice guideline for the treatment of patients with major depressive disorder (revision). American Psychiatric Association.” Am J Psychiatry. 2000 Apr;157(4 Suppl):1-45. PMID: 10767867.

[3] Latkin CA, Edwards C, Davey-Rothwell MA, Tobin KE. ”The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in Baltimore, Maryland.” Addict Behav. 2017 Oct;73:133-136. doi: 10.1016/j.addbeh.2017.05.005. Epub 2017 May 9. PMID: 28511097; PMCID: PMC5519338.
[4] Albuquerque L, Valente ARS, Teixeira A, Figueiredo D, Sa-Couto P, Oliveira C. ”Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan.” PLoS One. 2021 Apr 8;16(4):e0248842. doi: 10.1371/journal.pone.0248842. PMID: 33831018; PMCID: PMC8031302.
[5] Depression: His versus hers (2021) JHM. Available at: https://www.hopkinsmedicine.org/health/conditions-and-diseases/depression-his-versushers:
:text=Researchers%20have%20known%20for%20years,of%20disease%20burden%20among%20women.
[6] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, Thomas F. Quatieri, ”A review of depression and suicide risk assessment using speech analysis, Speech Communication,” Volume 71, 2015, Pages 10-49, ISSN 0167-6393,
[7] L. Verde et al., ”A Lightweight Machine Learning Approach to Detect Depression from Speech Analysis,” 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), Washington, DC, USA, 2021, pp. 330-335, doi: 10.1109/ICTAI52525.2021.00054.
[8] Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, Said Agrebi, Alice Othmani, ”MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech,” Biomedical Signal Processing and Control, Volume 71, Part A, 2022, 103107, ISSN 1746-8094,
[9] Gratch, Jonathan Arstein, Ron Lucas, Gale Stratou, Giota Scherer, Stefan Nazarian, Angela Wood, Rachel Boberg, Jill DeVault, David Marsella, Stacy Traum, David Rizzo, Albert Morency, L.. (2014). ”The Distress Analysis Interview Corpus of human and computer interviews,”
[10] DeVault, David Artstein, Ron Benn, Grace Dey, Teresa Fast, Ed Gainer, Alesia Georgila, Kallirroi Gratch, Jonathan Hartholt, Arno Lor-Lhommet, Margot Lucas, Gale Marsella, Stacy Morbini, Fabrizio Nazarian, Angela Scherer, Stefan Stratou, Giota Suri, Apar Traum, David Wood, Rachel Morency, Louis-Philippe. (2014). ”SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. 13th International Conference on Autonomous Agents and Multiagent Systems,” AAMAS 2014. 2. 1061-1068.
[11] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B.W. Williams, Joyce T. Berry, Ali H. Mokdad, ”The PHQ-8 as a measure of current depression in the general population, Journal of Affective Disorders,” Volume 114, Issues 1–3, 2009, Pages 163-173, ISSN 0165-0327,
[12] India has 0.75 psychiatrists per 100,000 people. can telepsychiatry bridge the gap between Mental Health Experts & Patients? (no date) The Economic Times. Available at: https://economictimes.indiatimes.com/magazines/panache/india-has-0-75-psychiatristsper-100000-people-can-telepsychiatry-bridge-the-gap-between-mental-health-expertspatients/articleshow/78572684.cms?from=mdr.
[13] Kim A, Jang E, Lee S, Choi K, Park J, Shin H ”Automatic Depression Detection Using Smartphone-Based Text-Dependent Speech Signals: Deep Convolutional Neural Network Approach” J Med Internet Res 2023;25:e34474 URL: https://www.jmir.org/2023/1/e34474 DOI: 10.2196/34474
[14] Keskinpala, Hande Yingthawornsuk, Thaweesak Wilkes, D.M. Shiavi, Richard Salomon, Ronald. (2007). ”Screening for high risk suicidal states using mel-cepstral coefficients and energy in frequency bands. European Signal Processing Conference.”
[15] S. Alghowinem et al., ”A comparative study of different classifiers for detecting depression from spontaneous speech,” 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013, pp. 8022-8026, doi: 10.1109/ICASSP.2013.6639227.
[16] Haihua Jiang, Bin Hu, Zhenyu Liu, Gang Wang, Lan Zhang, Xiaoyu Li, Huanyu Kang, ”Detecting Depression Using an Ensemble Logistic Regression Model Based on Multiple Speech

Features", Computational and Mathematical Methods in Medicine, vol. 2018, Article ID 6508319, 9 pages, 2018. https://doi.org/10.1155/2018/6508319