IT UNIVERSITY OF COPENHAGEN

Research Project

Decision Tree Models for Audio Feature Classification in Depression Prediction

Gergo Gyori (gegy@itu.dk)

2024-12-15

Abstract

This study explores the utility of vocal biomarkers for depression diagnosis through binary classification methods mapped to binary. Using audio features extracted from speech in the DAIC-WOZ and EATD-Corpus datasets, I employ decision tree algorithms and cnn [what is cnn?] models to evaluate their predictive accuracy. I have reached poor accuracy. My results indicate that using PHQ8 questinoryy provide better results. The study highlights the challenges of using audio data for depression detection and the need for further research to improve model performance and generalizability.

Keywords: depression detection • audio analysis • machine learning • CNN • MFCC

1 Introduction

Depression affects millions globally and represents a significant public health issue[12]. Early detection and intervention are critical for effective management and treatment. Traditionally, depression assessment has relied heavily on clinical interviews and self-reported measures, such as the PHQ-8[7] and PHQ-9 questionnaire[5].

This research aims to advance the field of psychiatric diagnostics by exploring the potential of audio analysis to detect depression. Utilizing machine learning algorithms, specifically Decision Tree and Convolutional Neural Network (CNN) models, this study analyzes vocal biomarkers within audio recordings from two distinct datasets: the Distress Analysis Interview Corpus - Wizard of Oz (DAIC)[11] and the Emotional Audio-Textual Depression (EATD) Corpus [9]. These datasets offer source of vocal expressions aligned with validated depression assessments, providing a foundation for developing predictive models.

The challenge of accurately detecting depression from audio features encompasses several critical issues. These include addressing the class imbalance across different depression severity categories, managing the variability in audio quality, and ensuring the generalizability of models beyond the training data. This study aims to analyze these relationships between audio characteristics and depression and explore the viability of audio-based depression detection as a supplementary tool to traditional methods.

Despite the initial aim of utilizing machine learning to enhance the diagnosis of depression through audio analysis, this study reveals significant limitations. Key lessons include recognizing the efficiency of existing tools like the PHQ-8 and PHQ-9 questionnaires in predicting depression, which often outperform more complex machine learning approaches. Furthermore, this research underscores a critical flaw in many studies: using the same patients' audio samples for both training and testing, which, while improving model accuracy, fails to ensure generalizability to new individuals. These insights highlight the importance of methodological rigor and the need to prioritize practical generalization in future machine learning research.

2 Datasets Description

2.1 DAIC

The DAIC dataset is an essential resource in computational psychiatry, pivotal for developing algorithms to diagnose psychological distress conditions such as depression and anxiety. This publicly available English depression dataset features multimodal data including audio, video, and text transcripts of interviews conducted by an animated virtual agent named Ellie in a simulated clinical setting. The dataset includes 142 participants evaluated with the PHQ-8 score, a popular depression screening tool. A PHQ-8 score of 10 or higher is indicative of depression.

The dataset is divided into three subsets: the training set includes data from 30 depressed and 77 non-depressed participants, the development set consists of 12 depressed and 23 non-depressed participants, and an unlabeled test set. This structure provides a rich, controlled environment for testing and comparing different diagnostic approaches, enhancing the reliability and accuracy of mental health diagnostics.

Interviews are designed to elicit emotional responses through predefined prompts, making the dataset highly suitable for studying vocal characteristics, speech patterns, non-verbal cues, and facial expressions associated with mental health states. The extensive annotations related to behavioral markers allow researchers to explore multimodal integration techniques, further supporting the development of sophisticated diagnostic models. This comprehensive data and detailed annotations are invaluable for advancing methodologies in mental health assessments within artificial intelligence frameworks.

2.2 EATD

The EATD Corpus, created at Tongji University is a dataset that caters to the need for multimodal data in depression research. This dataset includes audio recordings and their corresponding textual transcripts from interviews conducted with both depressed and non-depressed volunteers, making it a vital resource for the development of automated depression detection systems.

The EATD is distinctive as it is the first publicly available Chinese dataset that integrates both audio and text modalities specifically for depression analysis. It comprises contributions from 162 student volunteers who provided informed consent, ensuring the data's authenticity and ethical integrity. Each session in the dataset is annotated according to the Self-Rating Depression Scale (SDS)[14], providing researchers with valuable clinical metrics to correlate with linguistic and acoustic features.

The nature of this dataset allows for research opportunities, including the enhancement of feature extraction methods for depression detection and the development of ML-driven models that utilize multimodal data to assess mental health states more accurately. Moreover, it supports the exploration of computational techniques in identifying depressive symptoms, thereby advancing the field of mental health technology.

3 Literature Review

During the literature review, I primarily focused on the DAIC dataset for several reasons. Firstly, while the EATD dataset is relevant, it is considerably smaller in scale, containing only three single sentences (negative, positive, neutral) from 162 participants. In contrast, the DAIC dataset provides a more comprehensive array of audio and textual data from 189 participants, enhancing the potential to train more robust machine learning models. This choice allows for a deeper exploration of methodologies and outcomes pertinent to the use of vocal biomarkers in depression detection within a larger and more varied participant base, thereby improving the generalization and statistical power of the findings.

For finding the best accuracy, this paper[8] was used, and three three papers were checked which reached the best accuracy. Among these Homsiang et al[2] achieved 95% accuracy using a 1D CNN architecture with data augmentation. Their approach involved converting audio to Mel Cepstral Coefficients (MCC) and implementing various augmentation techniques including noise reduction, pitch shifting, and speed adjustment. Their comparative study of different architectures (1D CNN, 2D CNN, LSTM, and GRU) demonstrated that 1D CNN with augmented data significantly outperformed other approaches, showing strong performance in both depression detection (precision: 0.91, recall: 1.00) and non-depression classification (precision: 1.00, recall: 0.90). This work particularly highlights the importance of data augmentation in improving model performance, as their non-augmented experiments only achieved 71% accuracy with 2D CNN.

Ishmaru et al. [3] achieved 97% accuracy using a Graph Convolutional Neural Network (GCNN) approach that analyzes correlations between audio features. Their model represented the relationships between 65 different audio features (including 24 MCC) as graph structures, allowing it to capture complex interactions between voice characteristics. They conducted two types of experiments: one with overlapping subjects in training and test sets (Setting 1, Speaker-dependent test) and another with completely separated subjects (Setting 2, Speaker-independent test). While Setting 1 achieved 95% accuracy, Setting 2's performance dropped significantly, highlighting a critical challenge in generalizing to new patients. This finding raises important questions about the practical applicability of current depression detection models when applied to previously unseen patients. This research suggests that while high accuracies are achievable in controlled settings, real-world application requires addressing the gap between training and new patient performance.

Yin et al. [13] introduced a novel approach to depression detection from speech by combining transformers with parallel Convolutional Neural Networks (TCC), achieving an accuracy of 94% using 40 band Mel-Frequency Cepstral Coefficients (MFCC). This method of feature extraction was critical in maintaining the fidelity of audio signals, thereby enhancing model accuracy. Importantly, the high accuracy was obtained under experimental conditions similar to "Setting 1" from prior research, where the model was trained and tested on audio samples from the same set of participants. This setup often leads to inflated performance metrics due to the model's limited generalization to new subjects. Their model, which incorporates two CNN streams for local feature extraction alongside a transformer using linear attention mechanisms with kernel functions, reduces computational demands while enhancing the ability to capture temporal dynamics in speech. The results, derived from the DAIC dataset, indicated that their hybrid model outperforms traditional CNN-LSTM architectures. This showcases the effectiveness of parallel processing and advanced attention mechanisms in recognizing depression from long speech sequences, highlighting the importance of robust feature extraction techniques like the 40 band MFCC in achieving high model performance.

In the literature on audio processing for depression detection, various audio preprocessing techniques have been utilized to enhance the quality of the data before analysis. Notably, Homsiang's approach involved some form of audio preprocessing, though specific details are not provided. Other studies have explicitly detailed their methods: for example, Ishmaru et al. described techniques for speech enhancement that include noise estimation and filtration using deep learning models, aiming to improve the clarity and quality of the audio data for better model performance[4]

Conversely, Yin et al. opted not to apply any preprocessing to their audio data. This approach can offer insights into the raw data's effectiveness but may require more sophisticated modeling techniques to deal with potential noise and variability in the audio signals.

This variety in approaches highlights a crucial aspect of audio-based depression detection research: the balance between enhancing data quality through preprocessing and developing models robust enough to handle raw, unfiltered data.

4 Methodology

4.1 Data Preparation

PHQ8 values are organized to multiclass[5]. The values are organized into binary values as well based on[6]. In case of EATD the SDS index the SDS index is categorized by and it is mapped to binary categories.

The table below shows the mapping of the PHQ8 values to binary values.