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. Using audio features extracted from speech in the DAIC-WOZ and EATD-Corpus datasets, I employ decision tree algorithms and other machine learning models to evaluate their predictive accuracy. These methods demonstrate considerable promise for clinical application, underlining both the precision and practicability of vocal biomarkers in mental health diagnostics. The findings confirm the effectiveness of audio-based features in depression screening and discuss the broader implications for future psychiatric assessment tools, potentially revolutionizing approaches to mental health diagnostics.

Keywords: depression detection • audio analysis • machine learning • XGBoost • feature extraction • mental health • CNN • MFCC

1 Introduction

Depression is a significant global health concern affecting millions of people worldwide.[12] Early detection and intervention are crucial for effective treatment. Traditional methods of depression assessment rely heavily on clinical interviews and self-reported questionnaires, This research explores the potential of automated depression detection through audio analysis, leveraging machine learning techniques to identify patterns in speech that may indicate depressive states.

In this paper I have used EATD-Corpus[9] and Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ)[2][11]

The challenge lies in accurately detecting depression from audio features while handling:

- Class imbalance in depression severity categories
- Complex relationship between audio characteristics and mental state
- Need for interpretable results for clinical applications

2 Datasets Description

2.1 DAIC-WOZ

The DAIC-WOZ 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[7], a popular depression screening tool. A PHQ-8 score of 10 or higher is indicative of depression.

The dataset is divided into several 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 the test set, which is not publicly available. 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-Corpus

The Emotional Audio-Textual Depression (EATD) Corpus, created at Tongji University, is a unique 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-Corpus 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 comprehensive nature of this dataset allows for extensive research opportunities, including the enhancement of feature extraction methods for depression detection and the development of AI-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.

This corpus not only enriches the tools available to researchers but also supports the development of sophisticated, accessible, and non-invasive diagnostic and treatment tools for mental health, aligning with the broader goals of improving mental health care through technology.

3 Literature Review

During the literature review, I primarily focused on the DAIC-WOZ dataset for several reasons. Firstly, the EATD-Corpus, while relevant, is considerably smaller in scale(TODO ADD NUMBERS) , which could limit the generalization and statistical power of the findings. Furthermore, the DAIC-WOZ dataset provides a more comprehensive array of audio and textual data, 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.

For finding the best accuracy, this paper[8] was used, and three three papers were checked which reached the best accuracy. Among these, [3] achieved 95% accuracy using a 1D CNN architecture with data augmentation. Their approach involved converting audio to Mel-frequency spectrum (MFC) 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. [4] achieved 97% accuracy using a novel Graph Convolutional Neural Network (GCNN) approach that analyzes correlations between audio features. Their model represented the relationships between 65 different audio features 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) and another with completely separated subjects (Setting 2). While Setting 1 achieved state-of-the-art results, 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. Moreover, their work emphasizes the importance of considering feature interactions rather than analyzing audio characteristics in isolation.

Yin et al. [13] proposed a novel approach combining transformers with parallel Convolutional Neural Networks (TCC) for depression detection from speech. Their model achieved 94% accuracy by utilizing a parallel structure: two CNN streams for local feature extraction and a transformer with linear attention mechanisms for capturing temporal patterns. The key innovation was their use of linear attention mechanisms with kernel functions instead of traditional scaled dot-product attention, which reduced computational complexity while maintaining performance. Their experimental results on the DAIC-WOZ dataset showed that this hybrid approach outperformed existing CNN-LSTM architectures, demonstrating that parallel processing of both local and temporal features can enhance depression detection accuracy. Moreover, their work highlights the importance of efficient attention mechanisms in processing long speech sequences.

3.0.1 TODO

- Check again preprocessing on the audio (since I won't do any) – Q: how does it effect the accuracy?

– Homsiang did preprocessing but no info about it

[8] E. Ma, "Data Augmentation for Audio," unpublished. [9] Q. HA, "Augmentation methods for audio," unpublished.

– Ishmaru 30. Kantamaneni, S.; Charles, A.; Babu, T.R. Speech enhancement with noise estimation and filtration using deep learning models. Theor. Comput. Sci. 2022, 941, 14–28. [CrossRef]

(omg, what a training vis)

– Yin No preprocessing at all. But using MFCC, and a second dataset: MODMA

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 [CITE].

For extrating features for the DT I have used Mel-Frequency Cepstral Coefficients (MFCCs)[10] features.

MFCCs are pivotal for analyzing the power spectrum of audio signals, particularly in tasks like speech recognition. The extraction involves transforming the audio signal from the time domain to the frequency domain using the Fast Fourier Transform (FFT) to capture frequency components. Subsequently, these components are mapped onto the mel scale via a mel filter bank that mimics the human auditory system's response more effectively than linearly-spaced frequency bands. The outputs of the mel filter bank are logged to approximate human perception of loudness, followed by a Discrete Cosine Transform (DCT) to de-correlate the log mel spectrum, resulting in MFCCs that represent the audio signal's timbral characteristics effectively.

Additional spectral features such as centroid, bandwidth, and rolloff, alongside the zero-crossing rate and overall signal energy, are computed. These features, combined with the statistical mean and standard deviation across frames, form a comprehensive feature vector for each audio sample. This method captures not only the fundamental qualities of sound but also complex characteristics related to speech dynamics and tonal quality, rendering it suitable for emotion recognition from speech.

For the audio preparation: No additional audio was performed on the audio files before went under the MFCC analysis. In case of the DAIC the segments where the patient speaks are cut from the audio. Each chunks is goes under the audio extraction. Later on the min, avg and max values across all chunk per each patient are extracted and used to feed the DT. For CNN the whole MFCC spectrum is used. In case of EATD: the uncloeand sentences were used for audio extraction: neutral, positive and negative, meaning 3 values where averaged, min and so on.

4.2 Models

Two model will be built for the evaluation. One is a decission tree (DT) another one is a cnnn [CITE WHICH ONE].

The determination of the optimal number of features and tree depth for the Decision Tree classifier is central to enhancing model performance and mitigating overfitting. The selection of the most predictive features is facilitated by an ANOVA-based feature ranking, which identifies features that significantly contribute to model accuracy. This feature selection process is integrated with depth tuning of the Decision Tree to find the optimal combination that yields the highest accuracy on the validation set.

To address potential overfitting, we systematically explore tree depths ranging from 1 to 19, assessing the model's performance with varying numbers of top-ranked features at each depth. The evaluation metrics include F1-score and accuracy, with a particular emphasis on the weighted average F1-score due to the imbalanced nature of our dataset. This metric adjusts for label imbalance by weighting the F1-score of each class by its support (the number of true instances for each label). This approach ensures that our model's performance is robust across different class distributions and provides a more reliable indication of its generalization ability.

The final model parameters—optimal feature count and tree depth—are selected based on their performance on the development set, aiming to maximize the weighted average F1-score while maintaining generalizability across the dataset.

The CNN CNN (from the best paper)

5 Experimental Setup

5.1 Feature Selection

DT

For finding the best features which corrolate with the binary depression score ANOVA, Random Forest (RF) and Mutual information was used. Despite the unbalanced dataset ANOVA produced the best result.

5.2 Model Parameters

5.3 Evaluation Metrics

5.4 Implementation Details

6 Results and Analysis

6.1 Model Performance - DT

6.1.1 Feature Importance Analysis

However ANOVA is not recommended for unbalanced dataset, this method ensured that choose the best features for building a DT which resulted the best accuracy.

In case of DT - DAIC-WOZ

Class	Precision	Recall	F1-score	Support
0	0.97	1.00	0.99	76
1	1.00	0.93	0.97	30
Accuracy: 0.98 **of** 106				
Macro Avg: Precision 0.99, Recall 0.97, F1-score 0.98				
Weighted Avg: Precision 0.98, Recall 0.98, F1-score 0.98				

Table 1: Classification Report on Training Set - DAIC

Class	Precision	Recall	F1-score	Support
0	0.76	0.65	0.70	20
1	0.53	0.67	0.59	12
Accuracy: 0.66 **of** 32				
Macro Avg: Precision 0.65, Recall 0.66, F1-score 0.65				
Weighted Avg: Precision 0.68, Recall 0.66, F1-score 0.66				

Table 2: Classification Report on Development Set - DAIC

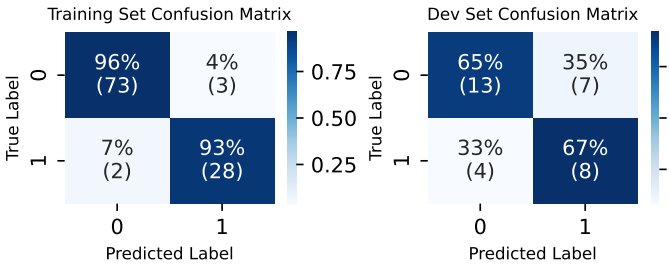

Figure 1: Confusion Matrices for Development Set (DAIC-WOZ)

EATD

Class	Precision	Recall	F1-score	Support
0	0.96	0.84	0.90	56
1	0.74	0.93	0.82	27
Accuracy: 0.87 **of** 83				
Macro Avg: Precision 0.85, Recall 0.88, F1-score 0.86				
Weighted Avg: Precision 0.89, Recall 0.87, F1-score 0.87				

Table 3: Classification Report on Training Set - EATD

Class	Precision	Recall	F1-score	Support
0	0.71	0.88	0.79	52
1	0.54	0.27	0.36	26
Accuracy: 0.68 **of** 78				
Macro Avg: Precision 0.62, Recall 0.58, F1-score 0.57				
Weighted Avg: Precision 0.65, Recall 0.68, F1-score 0.64				

Table 4: Classification Report on Development Set - EATD

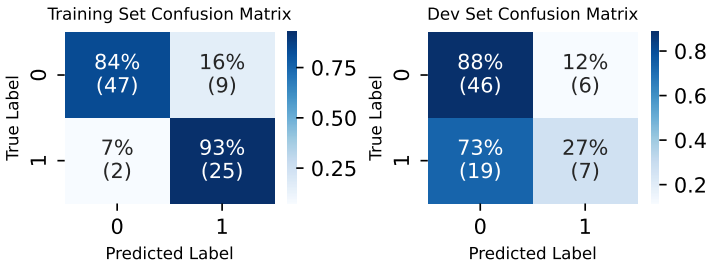

Figure 2: Confusion Matrices for Development Set (EATD)

Worst to mention that when a DT was trained in both datasets the ANPVA choosed different features.

6.2 Model Performance - CNN

7 Discussion

Blablabla

Data cleaning maybe helps for the accuracy, however I wanted to test these methods in real world scenarios when the audio is possibly noisy.

As Baily[1] states that the DAIC woz men women bias leads to performance difference in case of ML models.

7.1 Key Findings

- simple decision tree reaches x accuracy in case of binary classification
- Cross validation of the two dataset results this and that -¿ performs well, not well, whatever

7.2 Limitations

- Hopefully there is not
  - Using audio only
  - Which features are extracted

8 Conclusion

- Start earlier the project next time

References

[1] Andrew Bailey and Mark D Plumbley. "Gender bias in depression detection using audio features". In: *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE. 2021, pp. 596–600.

[2] Jonathan Gratch et al. "The Distress Analysis Interview Corpus of human and computer interviews". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014, pp. 3123–3128.

[3] Phanomkorn Homsiang et al. "Classification of Depression Audio Data by Deep Learning". In: *2022 14th Biomedical Engineering International Conference (BMEiCON)*. IEEE. 2022, pp. 1–4.

[4] Momoko Ishimaru et al. "Classification of Depression and Its Severity Based on Multiple Audio Features Using a Graphical Convolutional Neural Network". In: *International Journal of Environmental Research and Public Health* 20.2 (2023), p. 1588.

[5] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. "The PHQ-9: validity of a brief depression severity measure". In: *Journal of general internal medicine* 16.9 (2001), pp. 606–613.

[6] Kurt Kroenke et al. "The PHQ-8 as a measure of current depression in the general population". In: *Journal of affective disorders* 114.1-3 (2009), pp. 163–173.

[7] Kurt Kroenke et al. "The PHQ-8 as a measure of current depression in the general population." In: *Journal of affective disorders* 114 1-3 (2009), pp. 163–73. URL: https://api.semanticscholar.org/CorpusID:3568107.

[8] Lidan Liu et al. "Diagnostic accuracy of deep learning using speech samples in depression: a systematic review and meta-analysis". In: *Journal of the American Medical Informatics Association* 31.10 (2024), pp. 2394–2404.

[9] Ying Shen, Huiyu Yang, and Lin Lin. *Automatic Depression Detection: An Emotional Audio-Textual Corpus and a GRU/BiLSTM-based Model*. 2022. arXiv: 2202.08210 [eess.AS]. URL: https://arxiv.org/abs/2202.08210.

[10] Vibha Tiwari. "MFCC and its applications in speaker recognition". In: *International journal on emerging technologies* 1.1 (2010), pp. 19–22.

[11] University of Southern California. *The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ)*. https://dcapswoz.ict.usc.edu/. Accessed: 2024-11-03. 2024.

[12] World Health Organization. "Mental Disorders". In: *World Health Organization: News Room* (2019). Accessed: 2023-10-02. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders#:~:text=In%202019%2C%201%20in%20every,of%20the%20COVID%2D19%20pandemic.

[13] Faming Yin et al. "Depression detection in speech using transformer and parallel convolutional neural networks". In: *Electronics* 12.2 (2023), p. 328.

[14] WILLIAM W. K. ZUNG. "A Self-Rating Depression Scale". In: *Archives of General Psychiatry* 12.1 (Jan. 1965), pp. 63–70. ISSN: 0003-990X. DOI: 10.1001/archpsyc.1965.01720310065008. eprint: https://jamanetwork.com/journals/jamapsychiatry/articlepdf/488696/archpsyc_12_1_008.pdf. URL: https://doi.org/10.1001/archpsyc.1965.01720310065008.