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.

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 Transformer-CNN-CNN (TCC)[13].

4.3 DT

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.

4.4 TCC

I have adapted the TCC model for our application. The model consists of two parallel CNN streams and a transformer stream, integrating both local and global information processing capabilities. In my adaptation, I employ the CNN streams to extract local features from the input while the transformer stream captures the temporal dynamics through linear attention mechanisms, optimized for the dataset.

Each CNN stream processes the input independently to capture diverse aspects of the data, and the transformer stream analyzes the sequence as a whole. The outputs of these streams are then fused, combining their feature spaces to enhance the model's prediction accuracy. This fusion happens in a fully connected layer that integrates learned features before the final classification layer.

Modifications include adjusting the dimensionality of the input features and streamlining the transformer's attention mechanism to reduce computational complexity.

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 [TODO: Cite something about it] Using the top features selected by ANOVA produced the best result.

5.2 Model Parameters TCC

The implementation of the TCC model is based on the original paper [**tcc**]. some model parameters are lowered simply due to the HW limitations (to be able to fit into VRAM wiht limit of 4GB). The model is trained with a batch size of 32, a learning rate of 0.0001, and a maximum of 40 epochs.

6 Results and Analysis

This section presents the results of the experiments conducted in this study, organized into subsections focusing on specific aspects of the project. Initially, the performance of Decision Tree models is compared across the DAIC-WOZ and EATD datasets. In the subsequent TCC section, only the DAIC-WOZ dataset is used for evaluation.

6.1 Model Performance - DT

This subsection evaluates the performance of DT models across two datasets: DAIC-WOZ and EATD. While the models showcased excellent performance on the DAIC-WOZ dataset with an accuracy of 98.11%, precision of 97.00%, recall of 100.00%, and F1-score of 98.46% on the training set (Tables 1 and Figures 1, 2), a significant drop in performance was observed on the development set.

Particularly, the development set for the DAIC-WOZ dataset displayed only a 66% accuracy (Table 2), suggesting issues with the model's ability to generalize to new data.

Similarly, for the EATD dataset, while the training results were promising with an accuracy of 87% (Table 3), the development set results were considerably lower, achieving only a 68% accuracy (Table 4). This performance decrement underscores the necessity to consider alternative modeling strategies that might improve generalization across unseen datasets.

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

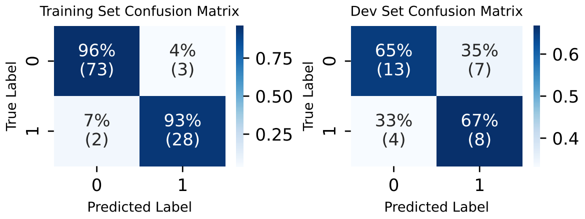

Figure 1: Confusion Matrices for Development Set (DAIC-WOZ)

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

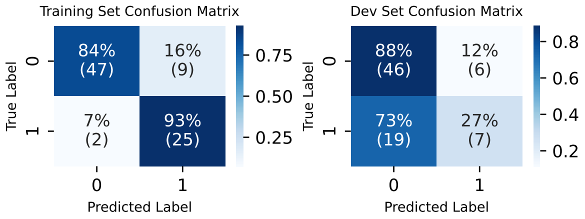

Figure 2: Confusion Matrices for Development Set (EATD)

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

6.2 Model Performance - CNN

In my experiments with various CNN models described in the literature, the models consistently underperformed, achieving only 50% accuracy on the development set. A deeper investigation into the literature revealed that the models achieving high accuracy were trained and tested on data from the same participants, merely split into different sets. This means that the same participant's audio files were divided between the train and test sets. The visualization of this data splitting method is shown in Figure 3.

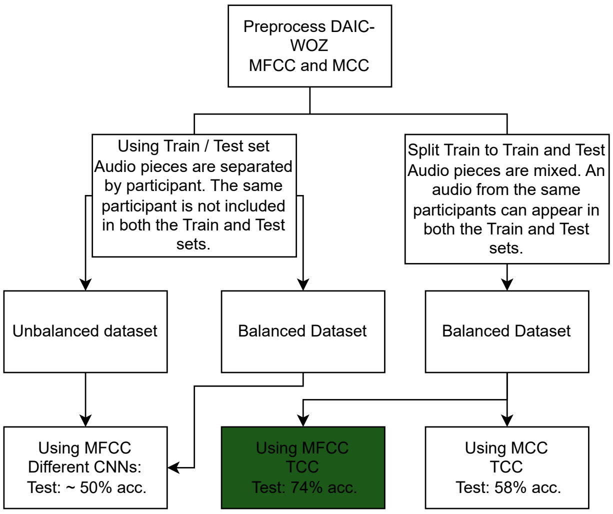
Preprocess DAIC-WOZ MFCC and MCC
Using Train / Test set Audio pieces are separated by participant. The same participant is not included in both the Train and Test sets.
Split Train to Train and Test Audio pieces are mixed. An audio from the same participants can appear in both the Train and Test sets.
Unbalanced dataset
Balanced Dataset
Balanced Dataset
Using MFCC Different CNNs: Test: ~ 50% acc.
Using MFCC TCC Test: 74% acc.
Using MCC TCC Test: 58% acc.

Figure 3: Train/Test split strategies and their results (DAIC)

The CNN models struggled to generalize when tasked with predicting new, unseen participants. The variance in accuracy across different CNN architectures is not discussed in this report. Instead, we focus on the results from the TCC model, detailed in Figure 4. The training and validation accuracy and loss are depicted in Figure 5. The model underwent training for approximately 100 epochs, not to achieve the best possible accuracy but to demonstrate the model's learning capability. The model reached a 74% accuracy on the development set, using a lightweight version of the architecture proposed by Yin et al.[13]. The evaluation metrics presented are based on the model's performance at its peak accuracy.

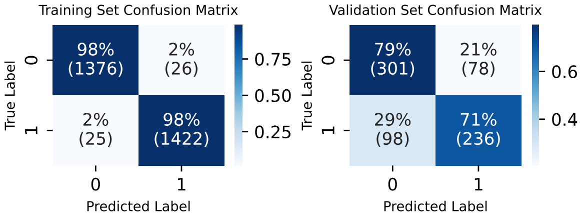

Figure 4: Confusion Matrices for Development Set (DAIC)

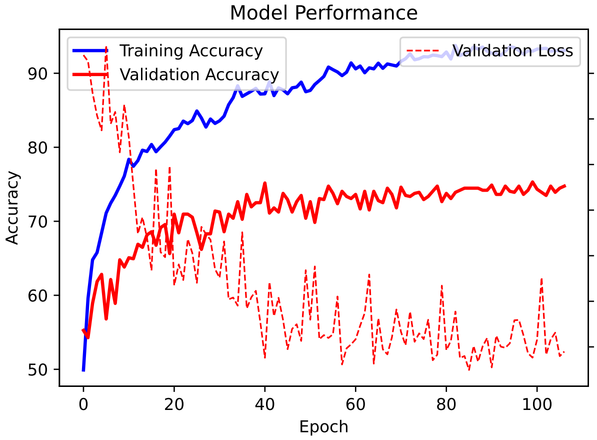

Figure 5: TCC Training (DAIC)

7 Discussion

This study's models are designed to predict PHQ-8 binary scores, which serve as a binary indicator of depression severity. Although the PHQ-8 is a reliable measure of depressive symptoms[7], this reliance raises questions about the necessity and utility of developing machine learning models based on audio data.

The technical feasibility of filling out the PHQ-8 survey, which is available online and considered trustworthy, further challenges the practicality of audio-based models. These models might seem redundant when a simpler and well-established method exists. However, audio-based applications could become relevant in scenarios where individuals are reluctant to complete the PHQ-8 survey. This might include cases where individuals, particularly those with severe or major depression, do not seek medical help.

Nevertheless, the utility of such models is constrained by the limited availability of public datasets, which impacts the robustness and generalizability of the findings. Additionally, factors such as varying audio quality and background noise—dependent on the microphone or the environment—can significantly affect the performance of models trained on audio data.

Furthermore, as highlighted by Bailey[1], biases such as gender discrepancies within the DAIC-WOZ dataset can lead to performance variations across machine learning models. These biases need to be addressed to enhance the fairness and accuracy of predictive modeling in clinical applications.

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] 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.

[3] 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.

[4] Sravanthi Kantamaneni, A Charles, and T Ranga Babu. "Speech enhancement with noise estimation and filtration using deep learning models". In: *Theoretical Computer Science* 941 (2023), pp. 14–28.

[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`.