IT UNIVERSITY OF COPENHAGEN

Research Project

Decision Tree and Convolutional Neural Network Models for Audio Feature Classification in Depression Prediction

Gergo Gyori (gegy@itu.dk)
Supervised by: Stella Graßhof (stgr@itu.dk)

2024-12-16

Abstract

This study explores the utility of vocal biomarkers for depression diagnosis through binary classification methods. While previous research has reported accuracies exceeding 90% using speaker-dependent approaches, these results have limited real-world applicability due to their experimental setup. Using audio features, specifically Mel Cepstral Coefficients (MCC) and Mel-Frequency Cepstral Coefficients (MFCC) extracted from speech samples from the DAIC-WOZ and EATD-Corpus datasets, I employ decision tree (DT) algorithms and convolutional neural network (CNN) models to evaluate their predictive accuracy, reaching 66% for DT. A critical gap exists between laboratory performance and clinical applicability, particularly in speaker-independent scenarios. Results suggest that traditional PHQ-8 questionnaires remain more reliable for depression screening compared to audio-based detection methods. The study highlights significant challenges in using audio data for depression detection, particularly the difficulty in generalizing to new patients and the impact of feature selection on model performance. These findings emphasize the need for careful consideration of the practical utility of audio-based depression detection systems in clinical applications. Future research should focus on developing robust speaker-independent models that can maintain accuracy across diverse populations and recording conditions.

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

1 Introduction

Depression affects millions globally and represents a significant public health issue [17]. 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 [10] and PHQ-9 questionnaire [8].

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) [14] and the Emotional Audio-Textual Depression (EATD) Corpus [12]. 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 methodological challenges. A fundamental limitation is that current approaches attempt to replicate questionnaire-based assessments (PHQ-8/9) through complex machine learning models, rather than discovering novel audio biomarkers that might provide additional clinical insights. Further-

more, 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. Additionally, the choice and processing of audio features play a crucial role, proving essential for the effective performance and reliability of machine learning models in clinical applications.

2 Datasets Description

2.1 DAIC

The Distress Analysis Interview Corpus - Wizard of Oz (DAIC) [14] dataset is a valuable resource for developing algorithms to diagnose conditions like depression and anxiety. It consists of 189 interview sessions conducted in English with an animated virtual agent named Ellie in a simulated clinical environment. This dataset includes audio recordings, text transcripts, and annotations for verbal and non-verbal cues, such as facial expressions through Action Units [13].

Participants were assessed using the PHQ-8 depression screening tool, derived from a self-reported survey. The PHQ-8 scores, with 10 or higher indicating depression, are the outcome of these surveys. The dataset is structured into three subsets to maintain data integrity: 107 participants in the training set, 35 in the development set, and 47 in the test set.

The dataset is particularly suited for analyzing vocal characteristics, speech patterns, and non-verbal behaviors associated with mental health states. Its detailed annotations support advanced studies into multimodal integration techniques, enhancing AI-driven mental health assessments.

In my research, I will focus exclusively on the audio data to explore vocal characteristics related to mental health.

2.2 EATD

The Emotional Audio-Textual Depression Corpus [12] dataset includes audio recordings and their corresponding textual transcripts from interviews conducted with both depressed and non-depressed volunteers. Each participant has six audio recordings—two each of neutral, positive, and negative sentences—in both cleaned and original formats.

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. Each session in the dataset is annotated according to the Self-Rating Depression Scale (SDS) [19].

3 Literature Review

In 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 heterogenous participant base, thereby improving the generalization and statistical power of the findings.

This study references the comprehensive review by Liu et al. (2024)[11], which lists numerous studies. From this review, I selected the three studies that reported the highest diagnostic accuracy for further analysis. Among these Homsiang et al [5] achieved 95% accuracy using a 1D CNN architecture with data augmentation. Their approach involved converting audio to Mel Cepstral Coefficients (MCC) [16] 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. [6] 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. [18] 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) [16] as an input. 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 fea-

ture 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 [7]

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 [8] based on the depression severity, and these severitities into binary values (non depression, depression) [9]. Table 1 shows how these values are organized. In case of EATD the the SDS index is categorized by and it is mapped to binary categories as it shown in Table 2.

Table 1: Severity levels according to the PHQ-8 score.

PHQ-8 Scores	Severity	Binary
0–4	Non depression	0
5–9	Mild	0
10–14	Moderate	1
15–20	Moderately severe	1
21–	Severe	1

Table 2: Depression severity levels based on SDS index scores with corresponding binary classification.

SDS Index	Severity	Binary
0–49	Normal	0
50–59	Mild	1
60–69	Moderate	1
70–100	Severe	1

4.2 Audio Features

Speech contains various acoustic characteristics that can be analyzed for depression detection. While numerous features exist for audio analysis, this study primarily focuses on two key representations: Mel-Frequency Cepstral Coefficients (MFCC) and Mel Cepstral Coefficients (MCC). Both representations use 40 frequency bands, where each band captures a specific range of frequencies, creating a detailed "fingerprint" of the audio signal. These 40 bands are distributed logarithmically across the frequency spectrum, meaning lower frequencies (which are crucial for human speech) are represented with finer detail than higher frequencies.

Figure 1 shows the MFCC and MCC representation of the word "Depression" [1]. In these visualizations, each of the 40 bands appears as a row, with colors indicating the intensity of the signal at different time points. The MFCC representation captures the spectral features of the audio signal, while the MCC representation emphasizes the cepstral features, providing a more detailed view of the audio signal's temporal dynamics. The choice of feature representation is crucial for model performance, as it determines the information available to the model for classification.

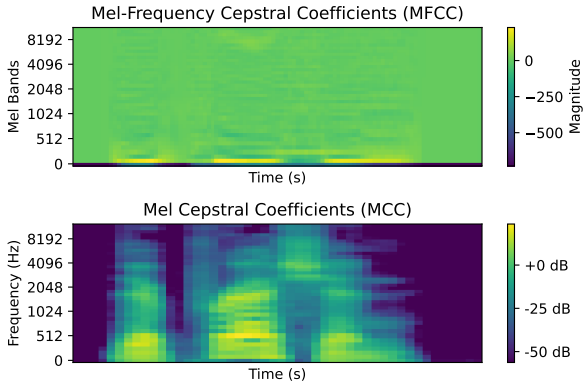

Figure 1: Visual comparison of MFCC and MCC representations for the word "Depression", highlighting their distinct audio feature emphases. Length of the audio signal is 1 seconds, sampling rate is 24Khz, it generates a 100 long MFCC and MCC.

4.2.1 MFCC

MFCCs are pivotal for analyzing the power spectrum of audio signals, particularly in tasks like speech recognition. MFCCs provide a representation of sound that approximates human auditory perception, with the mel scale

[1] Audio record, recorded by me
[2] https://cdn.britannica.com/04/14304-050-1A5E3289/structures-outer-ear.jpg

mapping onto the frequency sensitivity patterns of the human cochlea [2] . The extraction process has several steps: First, the audio signal is transformed from a time-based wave into its frequency components using the Fast Fourier Transform. These frequencies are then mapped onto the mel scale - a special scale that gives more attention to lower frequencies (like human speech) and less to higher frequencies (like high-pitched whistles), just as our ears do naturally. This mapping happens through a mel filter bank, which works like a series of overlapping filters that focus on different frequency ranges, similar to how different parts of our inner ear respond to different pitches. The outputs are then adjusted (logged) to match how we perceive loudness, since humans don't perceive sound intensity linearly - a whisper to normal speech feels like a smaller jump than normal speech to a shout. Finally, these features are mathematically processed (using a Discrete Cosine Transform [15]) to create the final MFCCs that effectively capture the key characteristics of the sound (the result is visualised on the first image 1).

4.2.2 MCC

Similar to MFCCs, MCCs also represent audio characteristics, but with a key difference in how they process the signal. While MFCCs use filter banks to mimic human hearing, MCCs employ a more direct mathematical approach that can capture subtle variations in speech: if MFCCs are like hearing sound through human ears, MCCs are like looking at sound through a precise scientific instrument. This makes MCCs particularly effective at capturing the unique characteristics of human voice, especially in tasks where subtle voice variations matter.

4.3 Feature Extraction Process

For the DAIC dataset, audio segments containing patient speech were isolated and processed. Each segment underwent feature extraction to compute either MFCC or MCC representations. For the Decision Tree model, statistical measures (minimum, average, and maximum values) were calculated across all segments for each patient. For the CNN model, the raw MFCC and MCC spectra were processed in 200-frame chunks.

For the EATD dataset, each participant recorded three different sentences: one with neutral emotion, one with positive emotion, and one with negative emotion. Each of these single-sentence recordings underwent the same feature extraction process as the DAIC dataset, with statistical measures computed for each sentence separately.

4.4 Models

Two model will be built for the evaluation. One is a decission tree (DT) another one is a Transformer-CNN-CNN (TCC)[18].

4.4.1 Decision Tree

The methodology employed for optimizing the DT classifier involves an integrated approach to feature selection and tree depth configuration. The objective is to enhance model performance while preventing overfitting. Feature selection was performed using three techniques to evaluate their effectiveness in identifying the most predictive features:

1. **ANOVA F-value:** Analyzes the variance among classes to identify features that significantly differentiate between them. This method calculates the F-value for each feature to determine its impact on classification accuracy, prioritizing those with higher values for model inclusion.
2. **Mutual Information (MI):** Measures the dependency between the features and the target variable, crucial for capturing nonlinear relationships.
3. **Random Forest Feature Importance (RF):** Utilizes the Random Forest algorithm to estimate the usefulness of each feature based on the impurity reduction it brings to the model.

These methods were chosen to provide a comprehensive analysis of feature relevance from both statistical and machine learning perspectives. The Decision Tree's depth was then tuned to find the optimal balance that yielded the highest accuracy on the validation set, using the most predictive features identified by the feature selection process.

To mitigate the risk of overfitting, I methodically investigated a range of tree depths from 1 to 19, while also varying the number of top-ranked features from 1 to 29. This approach allowed me to assess the model's performance at each depth, using different subsets of top features to determine the optimal combination that enhances model accuracy without overfitting. The evaluation metrics include F1-score [4] 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 my 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.

4.4.2 Convolutional Neural Network

I adapted three CNN models based on the studies highlighted in the literature review. Among these, the TCC model underwent a more detailed analysis. It integrates two parallel CNN streams with a transformer stream. This design effectively combines local and global information processing capabilities. Specifically, the parallel CNN streams are utilized to extract local features from the input, while the transformer stream, employing linear attention mechanisms, captures the temporal dynamics. This configuration is particularly optimized for handling the complexities of 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

Feature selection was critical in determining the best predictors for the binary depression score. Three different methods were evaluated:

- **ANOVA:** Used to identify features that showed significant differences between the two classes of depression scores. It employs the F-value to measure the ratio of variance between groups to variance within groups, with a higher F-value indicating greater discriminatory power of the feature [2].
- **Random Forest (RF):** Provided insight into feature importance based on ensemble learning.
- **Mutual Information:** Assessed each feature's mutual dependency with the target variable.

Despite the dataset's imbalance, the features selected through the ANOVA method demonstrated the most substantial impact on model performance. This method effectively distinguished features that are highly predictive of the binary outcome, thereby facilitating a more focused and effective model training process. The top features identified through ANOVA were then used to train the Decision Tree, leading to the best result in terms of accuracy and generalization on unseen data.

5.2 Model Parameters TCC

The implementation of the TCC model is based on the original paper [18]. 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 0.0005, and a maximum of 100 epochs.

6 Results and Analysis

6.1 Analysis of Speaker Dependency

A critical insight emerged during the analysis of different studies' performance metrics: the distinction between speaker dependent and independent approaches significantly impacts reported results. In *speaker dependent* setups, different segments from the same participant's interview can appear in both training and test sets – for example, various 30-second chunks from a single 5-minute interview might be distributed across both sets. While this ensures no exact duplicate segments exist between sets, the model can still learn speaker-specific characteristics.

In contrast, *speaker independent* approaches maintain strict separation: all segments from a participant's interview are allocated either entirely to training or test sets. This previously unaddressed factor in my methodology explains the substantial performance differences observed in Table 3, where speaker dependent approaches consistently show higher F1-scores compared to speaker independent setups. This finding highlights the importance of clearly specifying speaker dependency when reporting depression detection results.

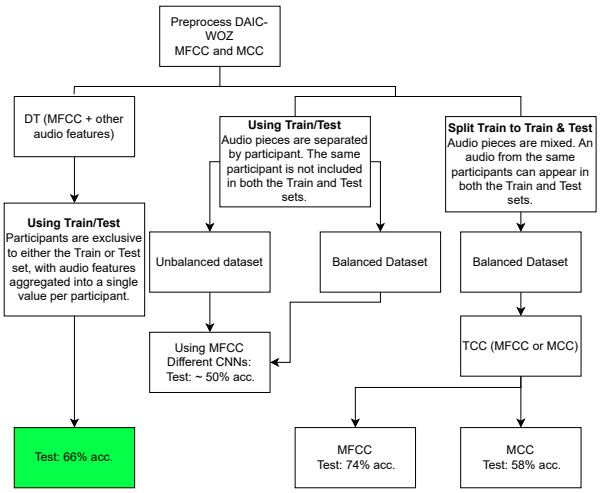

Figure 2: Train/Test split strategies and their results (DAIC)

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 DT models is compared across the DAIC and EATD datasets. In the subsequent TCC section, only the DAIC dataset is used for evaluation. The different strategies are visualized on figure 2.

Table 3 provides a comparison of the models discussed in the literature, highlighting the differences in dataset, model, F1-score, split strategy, speaker dependency, features, and additional features. The models are evaluated based on their performance on the DAIC dataset, with the F1-score serving as the primary metric for comparison.

6.2 Model Performance - DT

While the models showcased performance on the DAIC dataset with an accuracy of 98.11% on the training set (Tables 4 and Figures 3, 5), a significant drop in performance was observed on the development set.

Particularly, the development set for the DAIC dataset displayed only a 66% accuracy (Table 5), 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 6), the development set results were considerably lower, achieving only a 68% accuracy (Table 7). This performance decrement underscores the necessity to consider alternative modeling strategies that might improve generalization across unseen datasets.

Table 3: Comparison of Depression Detection Models

Author	Dataset	Model	F1-score (%)	Speaker Dependency	Features	Additional Features
Gyori	DAIC	DT	65	Independent	MFCC	No
Gyori	EATD	DT	68	Independent	MFCC	No
Gyori	DAIC	TCC	74	Dependent	MFCC	No
Gyori	DAIC	TCC	58	Dependent	MFC	No
Gyori	DAIC	CNN	50	Independent	MFCC	No
Ishimaru et al.	DAIC	CNN	96	Dependent	MFCC	Yes
Ishimaru et al.	DAIC	CNN	49	Independent	MFCC	Yes
Yin et al.	DAIC	TCC	93	Dependent	MFCC	Yes
Homsiang et al.	DAIC	1D CNN	95	Dependent	MFC	No

Table 4: Classification Report on Training Set - DAIC

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 5: Classification Report on Development 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				

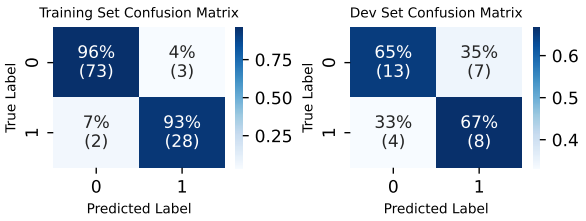

Figure 3: Confusion Matrices - DAIC

Table 6: Classification Report on Training Set - 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 7: Classification Report on Dev. 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				

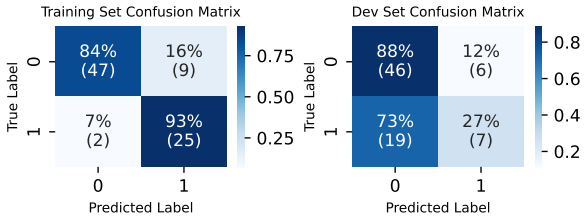

Figure 4: Confusion Matrices - EATD

Training DT on both DAIC and EATD datasets revealed that ANOVA selected different top features for each dataset, indicating potential challenges in generalizing the model across different data conditions.

6.3 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 (see Table 3). A deeper investigation into the literature revealed a critical distinction: while speaker independent approaches (where participants' data is strictly separated between train and test sets) showed similar modest performance, models achieving high accuracy (¿90%) were predominantly using speaker dependent setups.

This performance gap is evident in Table 3, where speaker dependent approaches by Ishimaru et al., Yin et al., and Homsiang et al. achieved F1-scores of 96%, 93%, and 95% respectively, while speaker independent implementations, including our implementation, struggled to exceed 50%. In speaker dependent setups, different segments from the same participant's interview were distributed between train and test sets, allowing models to learn individual speech characteristics rather than generalizable depression indicators.

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 5. The training and validation accuracy and loss are depicted in Figure 6. 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 [18]. The evaluation metrics presented are based on the model's performance at its peak accuracy.

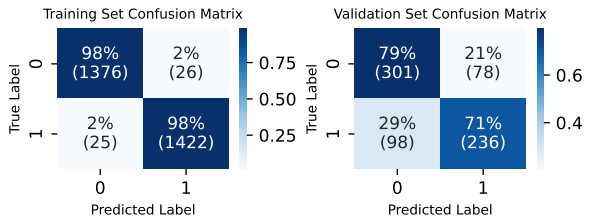

Figure 5: Confusion Matrices (Speaker Dependent - DAIC)

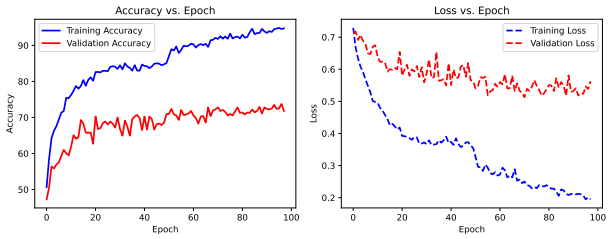

Figure 6: TCC Training (Speaker Dependent - DAIC)

7 Discussion

This study's models are designed to predict depression based on PHQ-8 scores, using them as a binary classifier to distinguish between depressed and non-depressed individuals. While the PHQ-8 is a validated and reliable measure of depressive symptoms [10], this approach raises questions about the necessity of developing audio-based machine learning models.

The PHQ-8 questionnaire assesses clear, observable symptoms including changes in sleep patterns, energy levels, appetite, concentration, and physical activity [3]. While self-reporting through PHQ-8 is straightforward and accessible online, there may be cases where individuals, particularly those with severe depression, are reluctant to actively seek help or complete surveys. In such scenarios, audio-based detection could provide an alternative screening method.

However, several critical limitations affect the development and deployment of audio-based depression detection models. The scarcity of public datasets constrains model robustness and generalizability. Environmental factors such as audio quality and background noise can significantly impact model performance. Our experiments with the TCC model demonstrated how different audio features (MFCC vs. MCC) yield varying results, highlighting the sensitivity of model performance to feature selection and processing techniques.

A significant methodological concern emerges from my literature review: many studies report high accuracies without clearly distinguishing between speaker-dependent and speaker-independent approaches. This distinction is crucial - speaker-dependent approaches, where audio segments from the same individual appear in both training and testing sets, may artificially inflate accuracy metrics. Such models may be merely recognizing individual speech patterns rather than detecting depression-related characteristics, limiting their real-world applicability.

Additional challenges include dataset biases, such as gender imbalances in the DAIC-WOZ dataset noted by Bailey [1], which can affect model performance across different demographic groups. These biases need careful consideration when developing models for clinical applications.

The gap between reported accuracies in research and practical clinical utility remains significant. While speaker-dependent approaches report impressive accuracies exceeding 90%, speaker-independent models - which better reflect real-world scenarios - typically achieve more modest performance. This disparity underscores the importance of rigorous evaluation methodologies that prioritize generalizability over optimistic performance metrics.

8 Conclusion

This study examined the effectiveness of using vocal biomarkers for depression detection through machine learning approaches, specifically comparing Decision Tree and TCC models across the DAIC-WOZ and EATD-Corpus datasets. While the Decision Tree model achieved an accuracy of 66% on unseen data from the DAIC dataset using a speaker-independent approach, the results raise important questions about the practical utility of audio-based depression detection systems.

A critical finding of my research is the significant disparity between speaker-dependent and speaker-independent approaches in depression detection. Many studies report impressive accuracies (90-95%) using speaker-dependent setups, where audio segments from the same participants appear in both training and test sets. However, these results are misleading as they primarily demonstrate the model's ability to recognize individual speech patterns rather than depression indicators. In contrast, my speaker-independent evaluation, which better reflects real-world scenarios by testing on completely unseen participants, achieved more modest but realistic performance metrics.

The challenges extend beyond speaker dependency issues. Traditional survey methods like PHQ-8 and PHQ-9 questionnaires remain more reliable and efficient tools for depression screening compared to complex machine learning models. These surveys provide immediate, validated results without the technical complexities and potential biases inherent in audio-based systems. Additionally, the varying performance observed with different feature extraction methods (MFCC vs. MCC) and model architectures demonstrates the inherent instability in audio-based detection systems.

While this research contributes to my understanding of machine learning applications in mental health, it also highlights the importance of rigorous evaluation methodologies that prioritize real-world applicability. The substantial performance gap between speaker-dependent and speaker-independent approaches, combined with the immediate availability and reliability of standardized questionnaires, suggests that current audio-based depression detection systems may have more academic value than practical clinical utility. Future research in this field should focus on developing robust speaker-independent models and establishing clear evaluation protocols that better reflect deployment conditions.

Acknowledgments

The author acknowledges the use of ChatGPT and Claude.ai for grammar enhancement in this paper. All technical content, analysis, and conclusions remain the author's original work.

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] Minitab Blog. *Understanding Analysis of Variance (ANOVA) and the F-Test*. Accessed: 2024-12-14. 2024. URL: https://blog.minitab.com/en/adventures-in-statistics-2/understanding-analysis-of-variance-anova-and-the-f-test.

[3] Self-Management Resource Center. *Patient Health Questionnaire-8 (PHQ-8)*. https://selfmanagementresource.com/wp-content/uploads/English_-_PHQ-8-1.pdf. Accessed: 2024-12-14.

[4] GeeksforGeeks. *F1 Score in Machine Learning*. https://www.geeksforgeeks.org/f1-score-in-machine-learning/. Accessed: 2024-12-09. 2021.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. "Recognizing action units for facial expression analysis". In: *IEEE Transactions on pattern analysis and machine intelligence* 23.2 (2001), pp. 97–115.

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

[15] Wikipedia. *Discrete cosine transform*. Accessed: 2024-12-14. 2024. URL: %5Curl%7Bhttps://en.wikipedia.org/wiki/Discrete_cosine_transform%7D.

[16] Wikipedia contributors. *Mel-frequency cepstrum — Wikipedia, The Free Encyclopedia*. [Online; accessed 7-December-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Mel-frequency_cepstrum&oldid=1100609773.

[17] 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.

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

[19] 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.