DEPRESSION SEVERITY ESTIMATION USING LEARNED VOCAL BIOMARKERS

Harrison Costantino
University of California, Berkeley
Department of Electrical Engineering and Computer Sciences
harrisoncostantino@berkeley.edu

DEPRESSION SEVERITY ESTIMATION USING LEARNED VOCAL BIOMARKERS

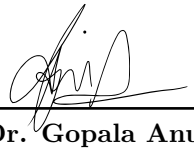
Harrison Costantino

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Dr. Gerald Friedland
Research Advisor
University of California, Berkeley

May 19th, 2022

Date



Dr. Gopala Anumanchipalli
Second Reader
University of California, Berkeley

May 20th, 2022

Date

Acknowledgments

I am incredibly grateful to the many people who have provided guidance and wisdom to me along my academic journey. In particular, I want to express my gratitude to my family for all the support they have given me over the years and to my partner for all the support she has given me during the uncertain times of the pandemic. I would also like to thank my advisor, Professor Friedland, for the invaluable time and assistance he has provided me over the last year. Lastly, I want to voice my appreciation to the people at Kintsugi both for their support and for allowing me to use their dataset.

Contents

List of Figures

List of Tables

Abstract

Human speech contains a rich set of acoustic biomarkers. When properly leveraged, these biomarkers can give powerful insights into the physical and mental health of the speaker. By exploiting these vocal biomarkers, machine learning models can be trained to detect altered speech patterns caused by depression or other mental health disorders. These speech based models serve as powerful, accurate, and non-invasive diagnostic tools. Prior works have explored the potential of these models and proven the feasibility of such systems on toy datasets. To see if these models have potential as a medical device, I re-implement some of these works on a dataset two orders of magnitude larger. Additionally, I introduce a new model that dramatically outperforms the current standard of care. I end with an investigation into this model's behaviour and a discussion of potentially relevant biomarkers.

1 Introduction

Major depression disorder is a serious mood disorder affecting an estimated 8.4% of US adults [26]. Despite the wide prevalence and seriousness of the disorder, general practitioners struggle to identify depression correctly [27], and many cases go undiagnosed. Multiple studies have examined the challenges of diagnosing depression in a general clinical setting and found significant barriers [38]. These barriers include the practitioner's personal biases, the patient's reluctance to speak about their mental health, and poor mental health training for physicians. Physicians have access to additional tools to assist in the diagnosis, but these tools have their own inadequacies. The most common diagnostic tools for depression are surveys such as the Patient Health Questionnaire-9 (PHQ-9) [10]. While the PHQ-9 has strong predictive power, it is not commonly administered and adds significant overhead to the doctor-patient interaction [38].

Because of survey-based diagnostic tools' limitations, vocal-based diagnostic models have been considered as non-intrusive and lightweight alternatives. These models are appealing as they would not require any additional time to complete, provide a deterministic and objective measurement, and easily integrate into the standard healthcare workflow. Speech has long been studied as a way of profiling the speaker [40], and the potential of vocal biomarkers to diagnose depression has been explored previously with great success [13, 22]. However, there are some areas of concern in the field. Several of these studies disagree on whether certain features show positive or negative correlation with depression [22]. The majority of these studies were performed on small datasets without a varied population. With the exception of a few papers, none of these studies exceeded a thousand unique participants, with many failing to reach five hundred. Several of these studies use the same dataset [47, 30, 42, 49, 24, 23, 36, 9, 3, 48]. This dataset, the DAIC-WOZ dataset, contains only 189 unique speakers [16, 46].

In this work, I address the shortcomings of these other papers by working with a new, proprietary dataset that is two orders of magnitude larger than the

		Not at all	Several days	More than half the days	Nearly every day
1.	Little interest or pleasure in doing things	0	1	2	3
2.	Feeling down, depressed, or hopeless	0	1	2	3
3.	Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4.	Feeling tired or having little energy	0	1	2	3
5.	Poor appetite or overeating	0	1	2	3
6.	Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7.	Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8.	Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9.	Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

Figure 1: The PHQ-9. The PHQ-2 is comprised of the first two questions (1 and 2) [41]

DAIC corpus and twice as large as the next largest proprietary dataset [16, 17]. Using this dataset, I re-implement some prior work to determine if their findings generalize to a larger population. I also improve upon the state of the art by introducing a new model that can accurately detect the presence and severity of major depression using only thirty seconds of speech.

2 Background

2.1 Clinical Background

Depression is a serious mental disorder. Depending on the severity, depression can profoundly impact an individual with both physical and mental symptoms. The mental symptoms of depression include persistent sad or empty feelings, irritability, inability to focus, and reduced cognitive ability. Depression can also manifest physically, causing insomnia, decreased energy, and reduced fine motor control [25, 20].

There are several tools to help physicians and psychiatrists identify and diagnose depression. The most common of these tools are the Patient Health Questionnaire (PHQ), the Hamilton Rating Scale for Depression (HAM-D), and

the Structured Clinical Interview for DSM-5 (SCID). The SCID is the gold standard for depression diagnoses. It is a semi-structured interview between a patient and a psychiatrist which takes an hour to complete. While most clinical research is based on the SCID, the labels are costly to obtain. Additionally, the SCID is rarely, if ever, used in clinical practice [14]. Instead, the HAM-D is used. The HAM-D is a 17-question survey filled out by a psychiatrist after interacting with the patient. Each question attempts to score the prevalence of different depressive symptoms [39]. In this way, the PHQ is similar to the HAM-D. Unlike the HAM-D, the PHQ is taken by the patient and the scores are self reported. Because there is no psychiatrist involved in the diagnoses, the PHQ is mainly used as a screening tool. However, it is also the most accessible tool to administer and the most commonly seen in clinical settings. The PHQ comes in three variants: the PHQ-9, PHQ-8, and PHQ-2. The PHQ-9, shown in Figure 1, asks the patient nine questions. Similar to the HAM-D, each question targets a different depressive symptom and asks the patient to rank the severity [21]. The PHQ-8 and PHQ-2 comprise the first eight and the first two questions of the PHQ-9, respectively. The PHQ-8 is used more commonly in academic settings as the ninth question deals with the subject of suicide and is often omitted. The PHQ-2 is sometimes used as a faster alternative to the PHQ-9; however, the PHQ-2 has significantly lower sensitivity than the PHQ-9 as shown in Figure 2.

The PHQ and HAM-D rate individuals on a scale from no depressive symptoms (0) to severe depression (27 or 52, respectively). To convert these scores to a binary label, it is common to use a "cutoff" value where all scores higher are labeled depressed and all scores lower are labeled healthy [39, 21]. For the PHQ-9, that cutoff value is most commonly 10. However, converting the score into 3 or 5 class buckets is also common. This flexibility of labeling allows some freedom when designing a machine learning system, as we can phrase the problem as binary classification, ordinal regression, or regression.

Like any machine learning application, it is vital to obtain a properly labeled dataset. However, this can prove challenging. While many researchers may be drawn to the SCID due to the accuracy of the labels, creating a dataset would be costly and time-consuming. Obtaining HAM-D labels is difficult as well, and if one works with only a few psychiatrists to obtain the labels one risks having personal bias affect the model. The PHQ, being self-reported, is the most obvious choice for collecting data. However, these labels lack any official diagnosis or psychiatrist oversight. Attempting to labeling data using multiple scales is also problematic since the different scales often disagree with one another, as shown in Figure 2. Additionally, these disagreements are mainly one-sided, with the HAM-D labels reporting lower rates of depression than the self-reported PHQ-9. The different versions of the PHQ behave similarly, with the PHQ-2 also reporting lower rates of depression when compared to the PHQ-9. These discrepancies can make it difficult to determine the proper labels for a dataset. To add to the issue, a person's PHQ or HAM-D score can fluctuate depending on the day, the patient's recent experiences, or the doctor's mood and biases. Unlike many classical machine learning problems, these data points

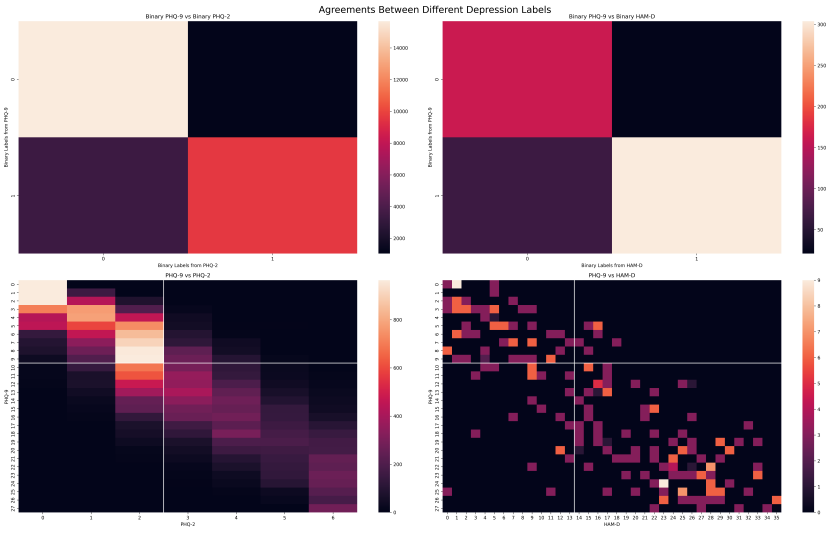


Figure 2: A comparison of HAM-D and PHQ-2 labels against the PHQ-9.

have no true "ground truth".

2.2 Classical Audio Features

The use of speech and audio as input to machine learning systems predates neural models by decades. The study of automatic speech recognition (ASR) and related problems has led to a plethora of different audio features suited for various tasks. Libraries such as openSMILE [12] make extracting common feature sets from audio files trivial.

Perhaps the most well-known audio feature is the spectrogram, a more powerful representation of audio than the raw waveform. Spectrograms show how the frequencies in the signal change over time. A common variant of the spectrogram is the Mel spectrogram. Mel spectrograms are unique as they portray the frequency bands in logarithmic instead of linear space. Depending on the application, this can be advantageous, as this closely matches how humans perceive changes in frequencies.

Another commonly used feature is Mel frequency cepstrum coefficients (MFCCs). MFCCs are commonly used as the main feature in classical ASR systems. They are an alternate representation of the audio signal showing how the cepstrum (as opposed to spectrum) changes over time. MFCCs are the results of applying additional transforms to a Mel spectrogram and show the underlying structures and harmonic nature of the spectrogram itself. The harmonics found in the spectrogram are closely related to the physical production of speech and the acoustics of the vocal tract.

Spectrograms and MFCCs are powerful, but being time series they pose ad-

ditional challenges to work with and incorporate into a model. Because of this, it is common for audio feature sets (such as the sets extracted by openSMILE) to instead be comprised of a large collection of statistics extracted from these signals. Some examples of these features are the mean of the fifth frequency band or the interquartile range of a specific MFCC coefficient. Furthermore, in addition to MFCCs and spectrograms, it is common to extract the same statistics of the derivative or deltas of these signals. Including the deltas helps to capture information about how the signal changes over time.

2.3 Neural Audio Features

In stark contrast to the classical feature sets, which result from purposeful transforms and well studied acoustic properties, neurally learned representations are powerful but uninterpretable acoustic features. Representation learning and self supervised training is a recent but powerful trend in machine learning, and the past two years have seen several neural speech feature extractors be made available.

Useful speech representations are often learned during the training of end-to-end models. One such example is Pyannote [7]. Pyannote is a high-performance speaker diarization system. Speaker diarization is the problem of separating a given audio file with multiple speakers into several audio files each containing only a single speaker. In order to do so, the model must learn an internal representation capturing only the most relevant information for the task of speaker diarization. Pyannote includes an audio embedder as part of its pipeline. This embedding is a fixed length 512-dimensional vector that can be used for other speaker analysis tasks.

Representations are not always simply the byproduct of solving some other task. Self supervised learning has been a growing movement recently in deep learning. The goal of these algorithms is to learn powerful representations first and foremost. The resulting models are then later tuned for downstream tasks. One popular self supervised algorithm is SimCLR [8]. Using similar ideas, Meta AI has developed several different speech feature extractors [2, 18, 1]. Meta AI's newest model, data2vec, uses a domain agnostic algorithm similar to SimCLR to train a large, 1 billion parameter model directly on speech. The performance of data2vec with minimal amounts of ASR finetuning is competitive with state of the art supervised ASR systems, a testament to the power of the learned representations.

2.4 Prior Work on Voice Biomarkers and Depression

The production of speech is the result of a complicated neuromotor pathway involving both cognitive and physical processes [45, 34]. As such, a person's ability to form sounds and words is majorly impacted by their physical and mental condition. Being tired, emotional, or otherwise cognitively impaired leaves a discernible effect on the speaker. Suffering from depression can affect nearly every part of the speech neuromotor pathway. The adverse effects of depression

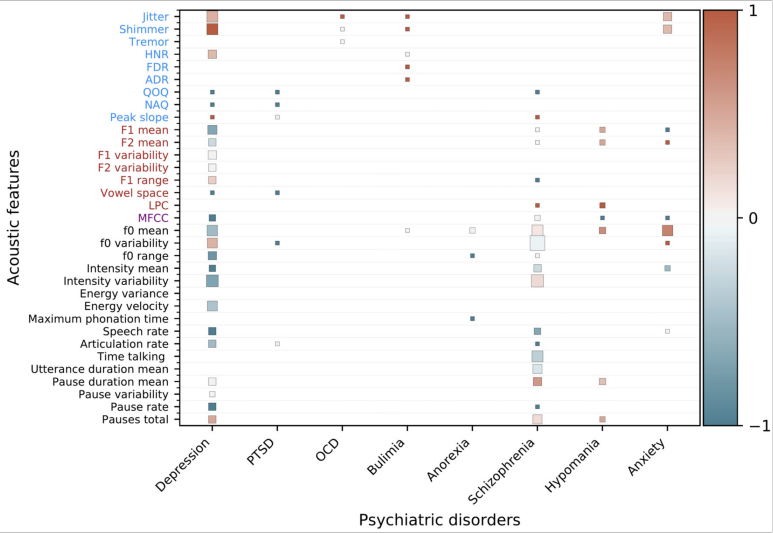


Figure 3: This figure was taken from [22]. The table consolidates biomarkers found in different studies and compares them across conditions. Red indicates positive correlation between the feature and the condition, blue indicates negative correlation, and white indicates contradictory findings between studies.

on an individual's cognitive abilities can cause slower, more simplistic speech. Fatigue, insomnia, and lack of energy make it difficult to speak at loud volumes or with energy and gusto. The decreased fine motor control adversely affects a speaker's ability to form sounds precisely.

The effects of depression on an individual's voice are well understood. Two older studies that tracked the voices of patients undergoing treatment found an extremely high correlation between their treatment progress and the speakers' pause rate and pause lengths [29, 43]. However, human speech is incredibly varied. While within a speaker it may be easy to correlate set features with condition improvement or regression, it is much harder to develop a system that can recognize depression across age groups, genders, and speaking styles.

There has been significant prior work exploring the potential of voice biomarkers as a diagnostic tool [13, 22]. In addition to depression, voice biomarkers have been applied with varying degrees of success to other afflictions such as respiratory and cardiovascular diseases, arthritis, Parkinson's, Alzheimer's, and other mental disorders [13]. While most research in the space has been focused on Parkinson's disease [13], there is a growing body of work focused on depression [22]. One such study modeling depression as a binary classification problem reported performance metrics as high as 87.5% accuracy, 0.91 sensitivity, and 0.83 specificity [11]. Similar studies have also reported high metrics, with one paper

on Chinese-speaking females claiming 72% accuracy and an F1 score of 0.81 [31]. Likewise, studies modeling the problem as regression are also reporting strong numbers. For example, one recent paper using transfer learning reported an MAE of 3.56/24 and a Pearson correlation of 0.49 [17].

Despite the recent increase of research in the area, there is a surprising lack of consensus on what specific biomarkers have the best diagnostic power. Some studies have reported contradictory results [22]; Figure 3 shows the results of a meta-analysis of vocal diagnostic tools. The lack of agreement on useful features is concerning and makes the results reported by these contradictory papers suspect.

There are some other concerning patterns among these papers as well. Many of these papers do not take the proper care to make clean data splits. Often, the same speakers will be in the train and test datasets. In one particularly egregious case, long audio files were split into thirty second chunks and randomly shuffled into train and validation sets. I believe the tendency to have overlapping users is due to the difficulty of procuring the needed datasets and the relatively small size of the available publicly available datasets.

Yet another issue in this space is the overuse of the DAIC-WOZ dataset [16]. A brief literature search yielded eleven different papers using the DAIC-WOZ dataset as their primary training and evaluation data and even more using the dataset as auxiliary data. In addition to the data being recycled, the underlying algorithm often is as well. One example is with a 2016 paper that introduced a model called DepAudioNet [23]. Since 2016, several papers have built off the DepAudioNet algorithm or used it as a benchmark when working with the DAIC-WOZ dataset [47, 35, 19, 5, 6]. While using performance on common datasets as a leader-board is somewhat common practice, these datasets tend to be massive and representative of the real world. The DAIC-WOZ dataset is under 200 unique speakers and suffers from both class and gender imbalance. It is unlikely that lines of work such as these would generalize to novel speakers and data. A recent paper investigating the overuse of the DAIC-WOZ dataset showed that the gender imbalance in the dataset has led to misleadingly high metrics [3].

In addition to potential gender bias, the DAIC-WOZ dataset suffers from an extreme location bias. While not explicitly stated in the metadata, most if not all participants are from the LA area. The participants were sourced using Craigslist and came to USC to perform the interviews [16]. Listening to these files the lack of diversity in the participants' background becomes evident. The homogeneity is concerning for a few reasons. Since the participants are all sourced from one micro-culture, differences in regional accents have the potential to affect these systems. Additionally, different areas may have different prevalence or types of depression in the population. While major depression or PTSD may be high in LA, seasonal affective disorder is likely not as big of a factor as somewhere like Michigan.

While not unique to the DAIC-WOZ dataset, or to the problem of depression classification, data imbalance can also cause issues in reported metrics. Depression only affects an estimated 9% [26] of American adults and severe depression

even fewer. The low prevalence causes data collected from a general population to heavily skew towards the lower end of the scale. When dealing with severity estimation and regression, this can cause misleading metrics. For example, if I were to simply always return the mean of the labels in my dataset I could achieve an MAE of 5.5 out of 28 despite the model having no predictive power.

3 Dataset

3.1 DAIC-WOZ

The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) [16] dataset is one of the few publicly available datasets containing paired speech samples and PHQ labels. The dataset has 193 unique speakers and contains one audio file per speaker. All of the participants were sourced from around the LA area using Craigslist. Each speaker was interviewed for five to twenty minutes by a "Wizard of Oz" (WOZ) system in which the patient talked with a human-controlled digital therapist. It's worth noting that I removed four files from the original dataset for my analysis due to lower audio quality and that the files were manually diarized. The DAIC-WOZ data is equipped with PHQ-8 labels instead of the full PHQ-9. The only demographic data available is the participant's gender. Of the 189 files in this dataset, 133 are healthy and 60 are depressed, resulting in an overall prevalence of 30%. The data imbalance is more severe than it first appears, especially for severity or regression based tasks. Despite there being more depressed classes than healthy classes, there is twice as much healthy data. This results in the individual depressed classes having little to no data. For example, the average amount of data points per healthy class is 13 files compared to 3 files for the depressed classes. The overall gender demographics of the dataset are fairly balanced, with 102 male participants and 87 female participants. However, the gender split is not as balanced as it appears. Despite being the minority of the dataset, the female participants make up the majority of the positive class examples. As shown in [3], many results using the DAIC-WOZ dataset are unwittingly using this gender imbalance to their advantage.

3.2 Proprietary Dataset

In addition to the DAIC-WOZ dataset, I have access to a large proprietary dataset. This dataset contains over 300 hours of voice samples from 15,941 unique speakers. Each speech sample is paired with a PHQ-9 score; a small subset of data is also paired with HAM-D scores. The dataset contains 13256 female and 5792 male participants from ages 16 to 93. There are some imbalances present in the data worth bringing to attention to. Demographically, the dataset has roughly twice as many female participants as male. The ages of the participants are also heavily skewed towards younger or middle-aged adults. A strong class imbalance is also present. While the prevalence of depression in

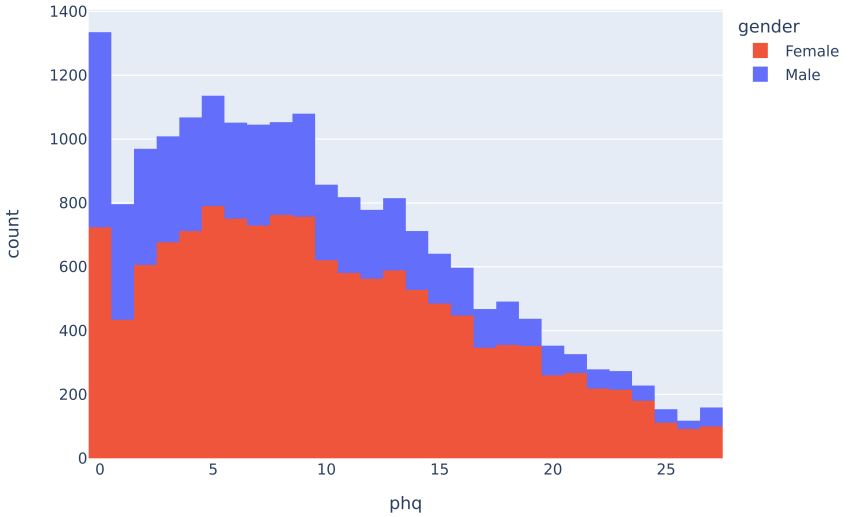


Figure 4: Histogram of PHQ-9 scores in proprietary dataset

the dataset is 44.6%, there are far more healthy participants than severely depressed ones. Figures 4 and 5 show a breakdown of the dataset by age, gender, and PHQ score.

The majority of participants were sourced remotely through a series of Reddit, Facebook, Instagram, and other social media ads. In addition to sourcing participants through social media, a small subset of the dataset was collected through Mechanical Turk. Participants were asked to record a sixty-second response to different open-ended prompts, such as "How was your day?". After recording themselves, participants were directed to a survey that collected demographic information and their responses to the PHQ-9.

As the audio files were recorded on a variety of devices, each audio file was converted to 16 kHz linear PCM for consistency. A sampling rate of 16 kHz was chosen as it is the standard for speech processing. In addition to transcoding every file, a voice activity detector was applied to remove beginning and end silences [44]. To ensure the overall quality of the data, an automated speech quality tool was used to remove any files containing a poor speech sample [28]. The homogeneous nature of the dataset is an important distinction between this and other datasets (such as the one used in [31] or the DAIC-WOZ itself). When researchers have complete control over both the recording equipment and recording environment the resulting dataset is much cleaner and homogeneous making the modeling effort easier.

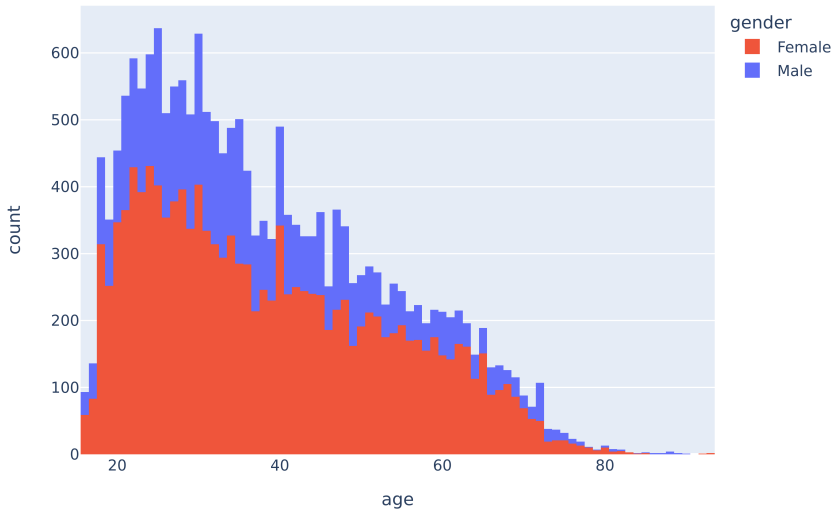


Figure 5: Histogram of ages in proprietary dataset

4 Reexamining Prior Work on Large Datasets

Using this proprietary dataset, I re-implemented two papers that I felt were representative of the work in the field and potential shortcomings of the current research. The first paper, "Re-Examining the Robustness of Voice Features in Predicting Depression: Compared with Baseline of Confounders" [31], is a relatively recent paper from 2019. This paper uses a proprietary dataset of 1000 Chinese-speaking women, avoiding any potential gender bias in their results. It is worth pointing out that my reimplementation is done in with English instead of Chinese. The techniques used are purely classical in their nature, relying on the feature set EMOBASE included in the openSMILE library [12] and utilizing a random forest as the underlying model.

In stark contrast to [31], the other paper I implemented is a purely deep learning based approach. The paper, "DepAudioNet: An Efficient Deep Model for Audio based Depression Classification" [23], was proposed as part of the AVEC 2016 challenge [46] and utilizes the DAIC-WOZ corpus [16]. Since 2016, there have been several responses to this paper [3, 5, 6, 19, 35, 47]. These follow-up papers either attempt to improve upon DepAudioNet or use it as a baseline to benchmark against. One paper reproduces the original algorithm to examine the effect of gender bias [3].

When re-implementing a paper, it is essential to pay attention to the differences in the underlying datasets. Since metrics like precision (and by extension

Description	Accuracy	Precision	Recall	F1
Results from [31]	0.71	0.77	0.84	0.80
Female Only Model Using Features From [31]	0.548	0.535	0.572	0.553
Female Only Model Using Selected Features	0.577	0.557	0.644	0.597
Mixed Gender Model Using Features From [31]	0.561	0.554	0.577	0.565
Mixed Gender Model Using Selected Features	0.578	0.564	0.651	0.604

Table 1: Results of my implementation of [31]

F1 score) are dramatically affected by the prevalence of the positive class, I ensured that for each of my reproductions my validation set prevalence matched the test set in the paper. This ended up being a prevalence of 51% for [31] and 31% for [23, 16]. To balance the training set for both implementations, I under-sampled the healthy class. As the training set in [31] was close to balanced this is a faithful reimplementation. The training set in [23] was far from balanced; however, the authors addressed this using clever sub-sampling at the audio level. Since I am working with a much larger dataset, this step seemed unnecessary.

4.1 Classical Model

The method proposed in [31] is similar to many other papers in the area [11, 50]. In the proposed method, audio is fed through the openSMILE toolkit to extract a 988-dimensional feature set known as EMOBASE. From this feature set, the authors perform feature selection to derive a small subset of 37 features to use with modeling. It is worth mentioning that there were only 36 voice features and that the speaker's age was included as a feature. Finally, they perform binary classification using a random forest and these 37 features.

For my reimplementation, I tried two different approaches. I began by extracting the same EMOBASE feature set the authors did. For my first attempt, I used the exact same 37 feature subset as the authors and trained a new random forest. For the second model, I chose to re-implement the feature selection process instead of reusing the features that worked for the authors. Using a similar method to what was reported, I arrived at a different subset of 37 features that better fit my data. Interestingly, the subset of features I derived had no features in common with the feature set proposed by [31]. In addition, I also reran these models on an exclusively female subset of the larger dataset to ensure a fair comparison.

As seen in Table 1, neither of my implementations performed well. The best performance I was able to achieve on my dataset was an accuracy of 57.8% and an F1 score of 0.604. Somewhat surprisingly, the mixed gender model narrowly outperformed the female-only model. The performance I achieved was significantly worse than reported by [31]. I believe there is one primary reason for this. While the train and test sets of [31] contained mutually exclusive audio,

approximately half the speakers in the test set of [31] were also in the training set. I ensured my train and test sets contained mutually exclusive speakers when reimplementing this paper.

4.2 DepAudioNet

DepAudioNet was proposed in [23] during the AVEC 2016 challenge [46]. In DepAudioNet, the audio is first transformed into a Mel spectrogram. The spectrograms are then normalized to zero mean and unit variance before being fed into a one-dimensional CNN. The output of the CNN undergoes a max pooling operation. The normalized CNN representation is fed into a 3-layer LSTM and then ultimately into a fully connected layer which makes a prediction for each segment of the audio. The final prediction is simply a majority vote of the segment-level predictions.

I based my reimplementation off of the reimplementation code made available by [3]. Similarly to my reimplementation of [31], I created female only models in addition to utilizing the entire dataset. Despite being trained on a larger, more diverse dataset, the performance of my replication fell short of the results claimed by [23] and replicated by [3]. One potential reason for this could be the more homogeneous nature of the DAIC-WOZ dataset in terms of both the speakers' demographics and the characteristics of the recorded audio. In [3] it is suggested that the performance of DepAudioNet was over reported due to the gender bias inherent in DAIC-WOZ, which could also explain the relatively high performance compared to my implementation.

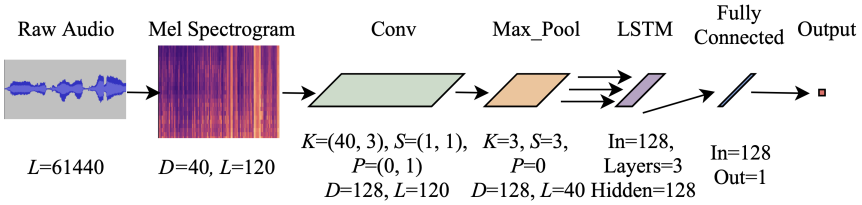


Figure 6: This figure was taken from [3] and shows the architecture of DepAudioNet

Description	Sensitivity	Specificity	PPV	NPV	F1 (+)	F1 (-)
Results from [23]	1.00	0.54	0.35	1.00	0.52	0.70
Reimplementation (Female Only)	0.512	0.567	0.348	0.722	0.415	0.635
Reimplementation (Mixed Gender)	0.649	0.483	0.360	0.753	0.462	0.588

Table 2: Results of my implementation of [23]. (+) refers to the results on the positive/depressed class while (-) refers to the results on the negative/healthy class

5 Method

My proposed method is a hybrid approach combining the strengths of classical speech classifiers and modern neural representation learning. Unlike modern end-to-end systems which work with natural data, the crux of my method is the feature extraction process prior to the data being input to the model. The model itself is a lightweight feed-forward neural net. In this way, my method is similar to other classical works such as [31, 50] that emphasize precisely selected feature sets over complicated models. Where my proposed method differs from these works is the feature extraction process. Instead of utilizing existing feature sets from libraries such as openSMILE, I created a unique feature set combining classical and neural speech representations.

In the first stage of the feature extraction process, the raw audio is converted into three different representations. Using torchaudio [32] I extract a 40-band Mel spectrogram and 20 MFCCs. I additionally feed the audio into Meta AI's data2vec large speech feature extractor [1]. Data2vec is a large, pretrained transformer model that extracts powerful, 1024-dimensional representations from speech. Including the original waveform, the audio is now represented in four different ways. As each representation is a collection of time series, I additionally extract each sequence's deltas to serve as additional features.

After computing the temporally based features, I use the Time Series Feature Extraction Library (TSFEL) [4] to extract a collection of functionals for

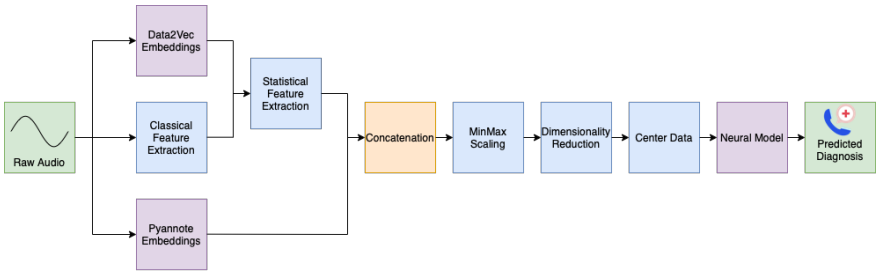


Figure 7: Visualization of model pipeline. Purple boxes correspond to deep learning based methods and blue boxes correspond to classical methods.

each sequence. Similar to the features in EMOBASE, these functionals are a collection of statistics such as mean, standard deviation, and zero crossing rate. Lastly, I extract 512-dimensional speaker embeddings using Pyannote [7]. Lastly, I concatenate the results of the time series feature extraction and the Pyannote embeddings to form the final feature space.

As this feature extraction process results in a large feature space prone to overfitting, I first perform dimensionality reduction on the data. Using sklearn [33], I scale the data to lie in the range (-1, 1) and then use kernel principal component analysis [37] to reduce the dimension. I use the rbf kernel and retain 1024 components. Lastly, I set the data to have zero mean by subtracting the mean of each feature.

While the feature extraction process outlined above is computationally intensive, it results in a descriptive representation of the audio, making it straightforward to train the model. The model itself is a four-layer fully connected neural net with a single output. The model uses the GELU activation function and dropout in every layer. The architecture is identical for both the binary and regression settings. A complete description of the model pipeline can be found in Figure 7.

The only difference between the binary and regression models is the training objective. In both cases, the model is trained for fifty epochs with a batch size of 32, a learning rate of $7.5e-4$, and using the AdamW optimizer. For the binary model, the training objective is binary cross entropy weighted so that the negative and positive classes carry the same weight. The regression model uses mean squared error for an objective and is trained to predict the raw PHQ-9 score.

While the performance of this model was strong when evaluated on the proprietary dataset, the metrics degraded when I ran inference on the DAIC-WOZ dataset. To address this issue, I made a 75/25 train/val split of the DAIC-WOZ data. Then, using the embedding space learned by the model as input, I trained a support vector regressor on the newly created training split. When the SVR was validated using the remaining DAIC-WOZ data, performance dramatically improved compared to the model without any finetuning.

6 Results

In the binary setting, my method achieves both a sensitivity and a specificity of 0.70. While some works claim higher numbers [11, 31], these results are competitive and are validated on a dataset orders of magnitude larger. In the regression setting, my model surpasses the start of the art and achieves a Pearson correlation coefficient of 0.56, seven points higher than the previously best reported 0.49 [17]. The confusion matrix in Figure 8 showing the accuracy of my model is visually similar to the confusion matrix in Figure 2 showing agreement between PHQ-9 and HAM-D scores. This suggests that the performance of my model is competitive with other depression diagnostic tools when

Feature Set Used	Sensitivity	Specificity	PPV	NPV	AUROC	MAE	PCC
Complete Feature Set	0.70	0.70	0.66	0.74	0.77	4.59	0.56
Neural Features	0.71	0.69	0.66	0.74	0.77	4.57	0.55
Classical Features	0.56	0.57	0.53	0.62	0.60	5.38	0.25
Zero Order Signals	0.70	0.69	0.66	0.74	0.76	4.59	0.54
First Order Signals	0.69	0.70	0.66	0.73	0.75	4.64	0.53
Data2Vec Embeddings	0.68	0.71	0.66	0.73	0.76	4.56	0.55
Pyannote Embeddings	0.65	0.58	0.56	0.67	0.66	5.10	0.38
MFCCs	0.57	0.57	0.53	0.62	0.59	5.37	0.24
Mel Spectrogram	0.56	0.57	0.52	0.61	0.58	5.40	0.24
Waveform	0.63	0.46	0.59	0.60	0.56	6.01	0.16

Table 3: Comparison of model performance by feature set evaluated on the proprietary dataset. With the exception of the Pyannote embeddings (which are not temporal in nature) every feature set includes both zeroth and first order signals unless otherwise stated.

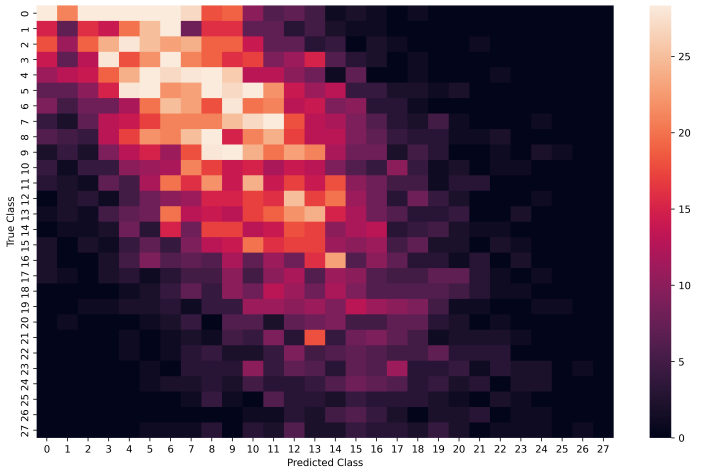


Figure 8: Validation 28-class confusion matrix

measured against the PHQ-9.

In addition to the results of my full method, I also included the results of the model when trained on different subsets of the feature space in Table 3. From these results it is clear that my proposed feature set is superior to the classical feature spaces used in prior works. Even when trained solely on Pyannote embeddings (which lack any temporal information) the model outperforms the entirety of the classical feature set. Interestingly, a model trained exclusively on the deltas of the original signals performs nearly as well as the full fledged model. Prior work has discussed the impact of depression on speaking and

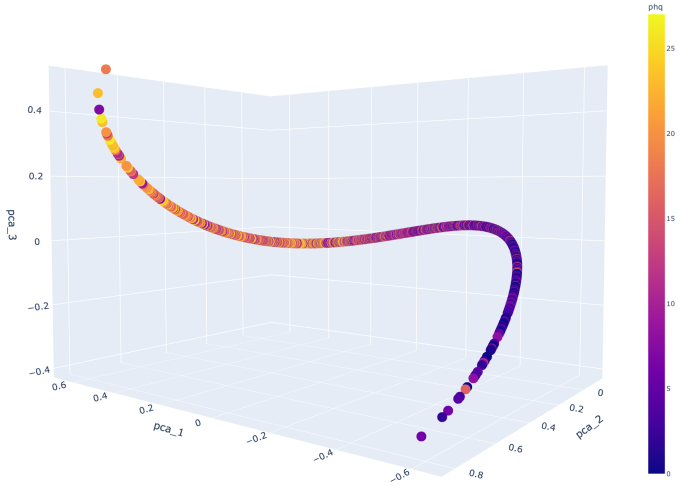


Figure 9: Visualization of validation data in model latent space. Lighter colors correspond to higher PHQ scores.

pause rates [43] so perhaps this result in unsurprising. Even the worst performing model, which was trained solely on statistical descriptions of the waveform itself, achieved better than random performance. This further reinforces the conclusion that a strong correlation between speech and depression exists.

In addition to the performance metrics and confusion matrices, I included a visualization of the model's latent space in Figure 9. This visualization shows the strong relative ordering of speech data my model has learned. The visualization was created using kernel PCA to project the model embedding into three dimensions. While graphed using three principal components, depression severity seems to be the only variable represented.

To test the usefulness of the latent space, I attempted to create a ternary severity model. I did this by training a support vector regressor using the model's latent space as an input. Instead of using raw PHQ-9 labels I bucketed them into a "low, mild, high" ternary label. These labels are similar to the binary labels derived from the PHQ-9 with scores close to the cut-off value instead put into the "mild" category. The ternary model showed strong performance, as seen in Figure 10. While individual class accuracy degraded compared to the binary model, the ternary model achieved a top-2 accuracy of 93% and 89% for class 0 and 2, respectively.

After tuning a support vector regressor, my model performed exceptionally well on the DAIC-WOZ dataset. My model achieved a MAE of 3.45 (on the PHQ-8/24-point scale) and a Pearson correlation of 0.69. Figures 11 and 12 show the confusion matrix and latent space visualizations for the DAIC-WOZ data.

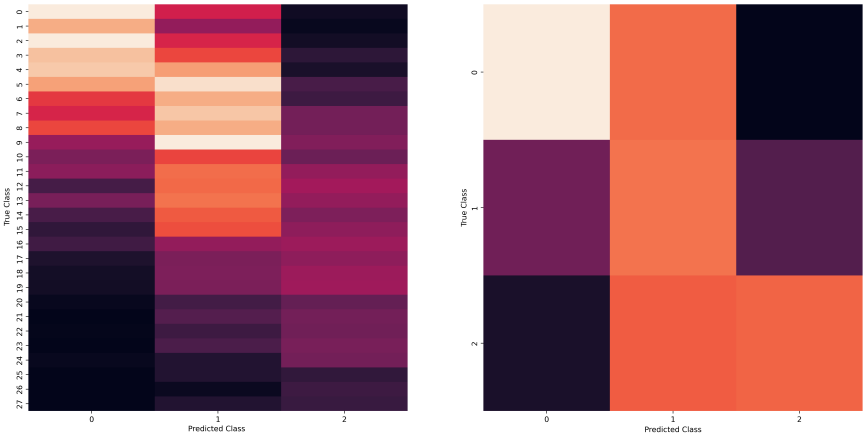


Figure 10: Performance of 3-class regression model trained on model embeddings

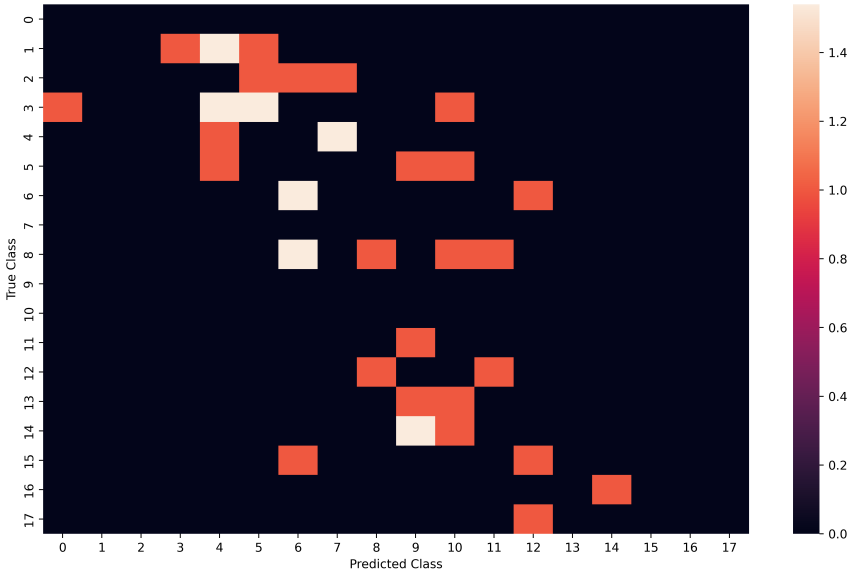


Figure 11: DAIC-WOZ validation set confusion matrix

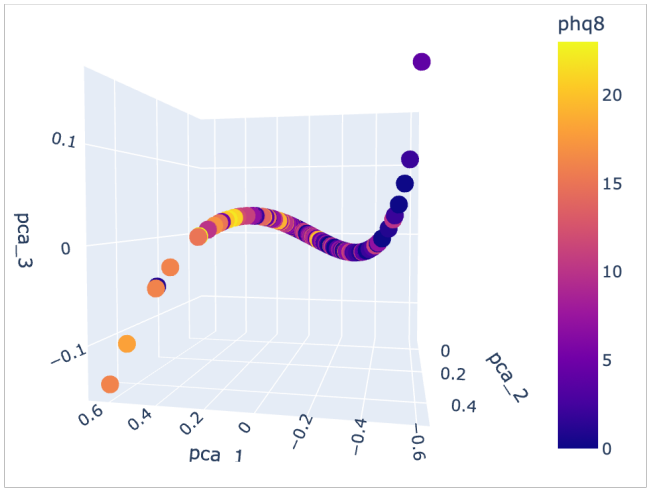


Figure 12: Visualization of DAIC-WOZ data in model latent space. Lighter colors correspond to higher PHQ scores.

7 Biomarker Investigation and Interpretability

One downside of using neural-based models and features is the loss of interpretability. In order to provide some insight into the differences between depressed and healthy speech, I used generative models to compare how "depressed" and "healthy" models spoke. Additionally, to understand how the model was making its decisions, I performed audio perturbation experiments to see how small changes to an audio file affect the prediction.

7.1 Dual VAE Experiments

For this experiment, I split the proprietary dataset into three sub-datasets: a training set containing exclusively healthy speech, a training set containing exclusively depressed speech, and a validation set containing a mixture of both. I then trained two different variational autoencoders. One VAE was trained exclusively on healthy speech while the other was trained exclusively on depressed speech. The code used in this experiment was adapted from [15].

After training the two VAEs, I used each model to reconstruct the entire validation set. This created two new versions of the validation set (files reconstructed by the "healthy" model and files reconstructed by the "depressed" model).

Once I had the three different copies of the validation set I used openSMILE to extract the EMOBASE feature set from every file. To see what features differed the most between reconstructions I normalized the feature range and

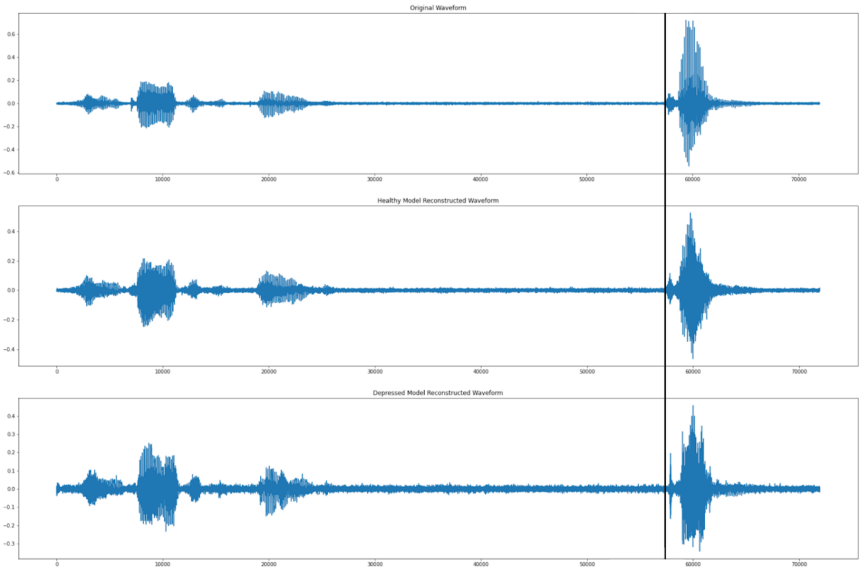


Figure 13: Comparison of an original waveform and the two reconstructions. The black line helps show how the pause length is slightly longer for the depressed reconstruction (bottom) compared to the original (top) or healthy reconstruction (middle).

computed the average difference between each feature across the reconstructions. The results of this analysis are shown in Table 4. I also compared the plots of the reconstructed waveforms and spectrograms against the originals. The most interesting finding in my opinion came from looking at the waveforms. The depressed model consistently reconstructed the speech signal with slightly longer pauses than the healthy model. An example of this can be seen in Figure 13. Lastly, I repeated the entire experiment once more to ensure my findings were robust and not artifacts of those specific models.

Feature	Change
lspFreq_sma[7]_quartile3	0.462
lspFreq_sma[7]_quartile2	0.398
lspFreq_sma[7]_amean	0.379
lspFreq_sma[7]_linregc2	0.314
mfcc_sma[11]_quartile3	-0.311
lspFreq_sma[7]_quartile1	0.284
mfcc_sma[6]_quartile3	0.282
mfcc_sma[6]_iqr1-3	0.252
mfcc_sma[11]_iqr1-3	-0.214
lspFreq_sma[1]_iqr1-3	0.210
mfcc_sma[6]_linregerrA	0.209
lspFreq_sma[7]_iqr1-2	0.199
mfcc_sma[11]_quartile2	-0.192
lspFreq_sma[1]_linregerrQ	0.187
mfcc_sma[6]_iqr2-3	0.186
mfcc_sma[11]_linregerrA	-0.185
mfcc_sma[11]_amean	-0.185
mfcc_sma[11]_stddev	-0.178
mfcc_sma[11]_linregc2	-0.175
mfcc_sma[6]_stddev	0.174
mfcc_sma[6]_linregerrQ	0.172
lspFreq_sma[1]_iqr2-3	0.168
lspFreq_sma[1]_linregerrA	0.167
lspFreq_sma[7]_min	0.167
mfcc_sma[6]_iqr1-2	0.166

(a) Most Changed Features

Feature	Change
F0_sma_kurtosis	0.000568
mfcc_sma_de[5]_iqr1-2	-0.000531
lspFreq_sma_de[7]_amean	-0.000527
pcm_loudness_sma_quartile2	0.000502
lspFreq_sma_de[1]_amean	0.000476
mfcc_sma_de[9]_quartile1	-0.000472
F0_sma_de_skewness	0.000454
lspFreq_sma[4]_min	-0.000407
lspFreq_sma_de[1]_kurtosis	0.000400
lspFreq_sma_de[7]_minPos	-0.000365
lspFreq_sma[2]_iqr2-3	-0.000361
pcm_intensity_sma_quartile1	-0.000272
lspFreq_sma[2]_kurtosis	-0.000224
pcm_loudness_sma_skewness	0.000190
F0env_sma_de_quartile2	-0.000161
mfcc_sma[6]_linregc1	-0.000133
lspFreq_sma_de[3]_maxPos	-0.000105
F0_sma_linregc1	0.000088
mfcc_sma_de[4]_linregc1	0.000078
F0_sma_de_quartile2	0.000067
lspFreq_sma[1]_linregc1	0.000044
mfcc_sma[2]_minPos	0.000039
lspFreq_sma_de[2]_maxPos	-0.000034
F0env_sma_minPos	0.000000
F0_sma_min	0.000000

(b) Least Changed Features

Table 4: Most and least changed features between reconstructions in dual VAE experiment

7.2 Audio Perturbation Experiments

In addition to the experiments with the generative models, I performed interpretability experiments using the severity model proposed in this work. I first created a subset of fifteen files with balanced labels. None of these files were used in training. Using this subset, I created fourteen different versions of the files by applying various transforms to the original speech signal. Each transform belonged to one of four categories: mu law encoding, pitch shifting, temporal masking, and temporal stretching/compression.

The most impactful change to the audio was slowing the playback speed. Slowing the audio by a factor of 0.75X resulted in an average change in prediction of +0.6. Speeding the audio up also had a noticeable impact, with a re-speeding factor of 1.25X resulting in an average change of -0.36. It is unsurprising that changing the audio speed resulted in changed predictions as the correlation between slowed speech and depression has long been known [29, 22, 43]. Modifying the pitch also had a relatively large impact on the model predictions. Strangely, both raising and lowering the pitch resulted in higher predictions. As this behaviour is bizarre, it is possible these changes are simply caused by an artifact of the pitch changing algorithm.

Perturbation	Average Change In Prediction
Re-speed x0.75	0.606
Re-speed x0.90	0.521
Re-speed x1.25	-0.361
Shift Pitch Quarter Octave Up	0.347
Shift Pitch Half Octave Up	0.299
Shift Pitch One Step Up	0.250
Shift Pitch Half Octave Down	0.217
Apply Mask on Middle	-0.143
Shift Pitch One Step Down	0.139
Apply Mu Law with 128 Quantization Levels	0.139
Re-speed x1.10	-0.135
Apply Mu Law with 64 Quantization Levels	0.105
Shift Pitch Quarter Octave Down	-0.073
Apply Mask Every Third	-0.024

Table 5: Effect of each audio perturbation on the model's prediction averaged over fifteen files.

Interestingly, the model seems robust to the other perturbations used. No perturbation resulted in an average prediction change of greater than one on a twenty-seven point scale. In particular, masking sections of the waveform and encoding the audio using mu law seemed to leave the predictions relatively unaffected. The full results can be found in Table 5.

8 Discussion

In this work, I discussed flaws in the current voice biomarker literature and the difficulty these systems have in generalizing to different datasets. I then introduced a new state of the art depression diagnostic tool and showed how it could generalize to novel datasets with minimal work. Lastly, I used the model proposed in this work alongside generative models to explore the vocal biomarkers themselves and gain a greater understanding of the differences between healthy and depressed speech.

Interestingly, the model trained with binary cross entropy loss develops almost as strong a latent ordering as the model trained with mean squared error. The visualizations of their latent spaces look highly similar, and the probabilities produced by the binary model correlate highly (PCC of 0.49) with the raw PHQ-9 score. When examining the points misclassified by the binary model, the majority of them have PHQ scores close to the cutoff value of 10. This implies that the underlying voice biomarkers are the same for both mild and severe depression and that the strength of these biomarkers is highly correlated

with the severity of the condition. These findings are consistent with those of [43, 29].

Similar to modern NLP systems, my proposed method relies on adapting the representations learned by a large, pretrained model. This method seems to surpass both classical and end-to-end systems. The feature set used seemed to be far more important than the quantity of data, as my reimplementations did not improve upon the results reported in the original papers. My work has shown that statistical descriptions of a signal can be just as powerful as the signal itself for diagnosing depression. This result is consistent with prior work showing a strong correlation between speaking rate, pause rate, and frequency variability with the severity of depression. The strength of these statistical methods is higher than previously thought, as evidenced by the success of the delta only model.

There are many potential directions in which to take this research. For this method I used data2vec as-is without any finetuning. Fine-tuning data2vec or using an LSTM head alongside it could prove more performant than my current method. Another potential source of improvement is including natural language processing or computer vision in the model's input to incorporate non-vocal biomarkers as well.

Outside of the modeling, more work should be done to ensure these models are free of any form of bias and achieve the same performance across all demographics. The model proposed in this work has shown robustness to both gender and ethnicity biases. However, age is a demographic that has proven challenging to avoid bias in. In particular, due to their distinct voices, the proposed model fails to generalize to children, early adolescents, or the elderly. With the collection of additional, targeted data it may be possible to adapt the proposed method to these demographics.

This technology also has potential outside of its current proposed use case. As a purely acoustic model, there is no reason a future model should be restricted to a single spoken language as input. Additionally, since there is significant overlap between biomarkers helpful in diagnosing various conditions [22], another direction for future work is the creation of a multi-purpose model. And of course, as these technologies mature, future work may focus on producing more accurate and precise models.

References

[1] Alexei Baevski et al. "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language". In: *arXiv* (2022).

[2] Alexei Baevski et al. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *arXiv* (2021).

[3] Andrew Bailey and Mark D. Plumbley. "Gender Bias in Depression Detection Using Audio Features". In: *EUSIPCO* (2021).

[4] Marilia Barandas et al. "TSFEL: Time Series Feature Extraction Library". In: *SoftwareX* 11 (2020).

[5] Flavio Bertini et al. "An automatic Alzheimer's disease classifier based on spontaneous spoken English". In: *Computer Speech & Language* 72 (2022).

[6] Flavio Bertini et al. "Automatic Speech Classifier for Mild Cognitive Impairment and Early Dementia". In: *Association for Computing Machinery* 3.1 (2021).

[7] Herve Bredin et al. "pyannote.audio: neural building blocks for speaker diarization". In: *arXiv* (2019).

[8] Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *arXiv preprint arXiv:2002.05709* (2020).

[9] Karol Chlasta, Krzystof Wolk, and Isabela Krejtz. "Automated speech-based screening of depression using deep convolutional neural networks". In: *Procedia Computer Science* 164 (2019), pp. 618–628.

[10] Luigi Costantini et al. "Screening for depression in primary care with Patient Health Questionnaire-9 (PHQ-9): A systematic review". In: *Journal of Affective Disorders* 279 (2021), pp. 473–483.

[11] Caroline Espinola et al. "Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study". In: *Research on Biomedical Engineering* 37 (2020), pp. 53–64.

[12] Florian Eyben, Martin Wollmer, and Bjorn Schuller. "Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor". In: *Proceedings of the 18th ACM International Conference on Multimedia*. 2010, pp. 1459–1462.

[13] Guy Fagherazzia et al. "Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice". In: *Digital Biomarkers* 5 (2021), pp. 78–88.

[14] Michael B First. "Structured Clinical Interview for the DSM (SCID)". In: *Encyclopedia of Clinical Psychology* (2015).

[15] Laurent Girin et al. "Dynamical Variational Autoencoders: A Comprehensive Review". In: *Foundations and Trends® in Machine Learning* 15.1–2 (2021), pp. 1–175. ISSN: 1935-8237. DOI: 10.1561/2200000089. URL: http://dx.doi.org/10.1561/2200000089.

[16] J Gratch et al. “The Distress Analysis Interview Corpus of Human and Computer Interviews”. In: *Proceedings of LREC*. 2014, pp. 3123–3128.

[17] Amir Harati et al. “Speech-Based Depression Prediction Using Encoder-Weight-Only Transfer Learning and a Large Corpus”. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 7273–7277.

[18] Wei-Ning Hsu et al. “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”. In: *arXiv* (2021).

[19] N. Janardhan and Nandhini Kumaresh. “Improving Depression Prediction Accuracy Using Fisher Score-Based Feature Selection and Dynamic Ensemble Selection Approach Based on Acoustic Features of Speech”. In: *Traitement du Signal* 39 (2022), pp. 87–107.

[20] Sidney Kennedy. “Core symptoms of major depressive disorder: relevance to diagnosis and treatment”. In: *Dialogues in Clinical Neuroscience* 10.3 (2008), pp. 271–277.

[21] Kurt Kroenke, Robert Spitzer, and Janet Williams. “The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review”. In: *General Hospital Psychiatry* 32.4 (2010), pp. 345–359.

[22] Daniel Low, Kate Bentley, and Satrajit Ghosh. “Automated assessment of psychiatric disorders using speech: A systematic review”. In: *Laryngoscope Investigative Otolaryngology* 5 (2020), pp. 96–116.

[23] Xingchen Ma et al. “DepAudioNet: An Efficient Deep Model for Audio based Depression Classification”. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 2016, pp. 35–42.

[24] Adria Mallol-Ragolta et al. “A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews”. In: *Proc. Interspeech 2019*. 2019, pp. 221–225.

[25] National Institute of Mental Health. *Depression*. 2022. URL: `https://www.nimh.nih.gov/health/topics/depression`.

[26] National Institute of Mental Health. *Major Depression*. 2022. URL: `https://www.nimh.nih.gov/health/statistics/major-depression`.

[27] Alex Mitchell, Amol Vaze, and Sanjay Rao. “Clinical diagnosis of depression in primary care: a meta-analysis”. In: *Lancet* 374 (2009), pp. 609–619.

[28] Gabriel Mittag et al. “NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets”. In: *Interspeech 2021*. 2021.

[29] James Mundt et al. “Vocal Acoustic Biomarkers of Depression Severity and Treatment Response”. In: *Society of Biological Psychiatry* 72 (2012), pp. 580–587.

[30] Muhammed Muzammel et al. "AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis". In: *Machine Learning with Applications* 2 (2020).

[31] Wei Pan et al. "Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders". In: *PLOS One* 14 (2019).

[32] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[33] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[34] Lawrence Raphael, Gloria Borden, and Katherine Harris. *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*. Lippincott Williams and Wilkins, 2007.

[35] Vijay Ravi et al. "Fraug: A Frame Rate Based Data Augmentation Method for Depression Detection from Speech Signals". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 6267–6271.

[36] Afef Saidi, Slim Ben Othman, and Slim Ben Saoud. "Hybrid CNN-SVM Classifier for Efficient Depression Detection System". In: *2020 4th International Conference on Advanced Systems and Emergent Technologies*. 2020, pp. 229–234.

[37] Bernhard Scholkopf, Alexander J. Smola, and Klaus-Robert Muller. "Kernel Principal Component Analysis". In: *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA, USA: MIT Press, 1999, pp. 327–352. ISBN: 0262194163.

[38] Isabelle Schumann et al. "Physicians' attitudes, diagnostic process and barriers regarding depression diagnosis in primary care: a systematic review of qualitative studies". In: *Family Practice* 28.3 (2012), pp. 255–263.

[39] Rachel Sharp. "The Hamilton Rating Scale for Depression". In: *Occupational Medicine* 65.4 (2015), pp. 340–340. ISSN: 0962-7480. DOI: 10.1093/occmed/kqv043. eprint: https://academic.oup.com/occmed/article-pdf/65/4/340/4184312/kqv043.pdf. URL: https://doi.org/10.1093/occmed/kqv043.

[40] Rita Singh. *Profiling Humans from their Voice*. Springer, 2019.

[41] Robert Spitzer and Janet Williams. *Depression — PHQ-9*. URL: https://help.greenspacehealth.com/article/85-depression-phq-9.

[42] NS Srimadhur and S Lalitha. "An End-to-End Model for Detection and Assessment of Depression Levels using Speech". In: *Procedia Computer Science* 171 (2020), pp. 12–21.

[43] E. Szabadi, C. M. Bradshaw, and J. A. O. Besson. "Elongation of Pause-Time in Speech: A Simple, Objective Measure of Motor Retardation in Depression". In: *Britian Journal of Psychiatry* 129 (1976), pp. 592–597.

[44] Silero Team. *Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier*. `https://github.com/snakers4/silero-vad`. 2021.

[45] Pascale Tremblay and Isabelle Deschamps. "Neuromotor Organization of Speech Production". In: *The Oxford Handbook of Neurolinguistics* (2019).

[46] Michel Valstar et al. "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge". In: 2016, pp. 3–10.

[47] Adrian Vazquez-Romero and Ascension Gallardo-Antolin. "Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks". In: *Entropy* 22.6 (2020). ISSN: 1099-4300. DOI: `10.3390/e22060688`. URL: `https://www.mdpi.com/1099-4300/22/6/688`.

[48] Esau Villatoro-Tello, Gabriela Ramirez-de-la-Rosa, and Daniel Gatica-Perez. "Approximating the Mental Lexicon from Clinical Interviews as a Support Tool for Depression Detection". In: *Proceedings of the 2021 International Conference on Multimodal Interaction*. 2021, pp. 557–566.

[49] Le Yang, Jiang Dongmei, and Hichem Sahli. "Feature Augmenting Networks for Improving Depression Severity Estimation From Speech Signals". In: *IEEE Access* 8 (2020), pp. 24033–24045.

[50] Larry Zhang et al. "Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative". In: *Anxiety and Depression Association of America* 37 (2020), pp. 657–669.