2022

The Impact of Emotion Focused Features on SVM and MLR Models for Depression Detection

Alexandria Mulligan

The Impact of Emotion Focused Features on SVM and MLR Models for Depression Detection

Alexandria Mulligan

D17127341

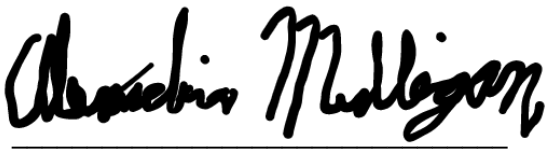
A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
M.Sc. in Computer Science Data Analytics

2022

I certify that this dissertation which I now submit for examination for the award of MSc in Computing Data Analytics, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: 

Date: **08 September 2022**

ABSTRACT

Major depressive disorder (MDD) is a common mental health diagnosis with estimates upwards of 25% of the United States population remain undiagnosed. Psychomotor symptoms of MDD impacts speed of control of the vocal tract, glottal source features and the rhythm of speech. Speech enables people to perceive the emotion of the speaker and MDD decreases the mood magnitudes expressed by an individual. This study asks the questions: “if high level features deigned to combine acoustic features related to emotion detection are added to glottal source features and mean response time in support vector machines and multivariate logistic regression models, would that improve the recall of the MDD class?” To answer this question, a literature review goes through common features in MDD detection, especially features related to emotion recognition. Using feature transformation, emotion recognition composite features are produced and added to glottal source features for model evaluation.

Two emotion recognition based composite features were created and along with the baseline features each was ran in a linear SVM and a MLR model. The MLR models achieved the higher of the two in accuracy with the 51.4% and 51.6% for an emotional load inspired feature and a PCA feature transformed feature. When compared to models ran with just the baseline the difference in recall after 100 iterations generated p scores of 0.071 and 0.056 of the new vitality and PCA emotion composite feature. These values are close to the 0.05 significance threshold indicating further work and research may benefit model performance. In addition, explorative research into gender balance of the DAIC-Woz dataset suggest that further research into gender split, with a balance of MDD score distributions, models would benefit the F0 based features used in the two proposed emotion recognition feature. Additionally further work into targeting smaller sections of audio after an uplifting question would be more conducive to these new features.

Key words: *MDD, classifier, SVM, MLR*

ACKNOWLEDGEMENTS

This endeavour would not have been possible without the guidance and expertise from Sean O'Leary who I would like to express my deepest gratitude for his assistance throughout this process.

I would also like to acknowledge and thank the University of Southern California for access to the DAIC-Woz database that without this research study would not have been possible.

Finally, I would love to express gratitude to my family and friends. Through editing session and unconditional support, they are a pillar on which this research stands.
Most especially I would like to thank my late sister, whose memory pushes me to achieve, and who led me to the research topic within this study.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1. INTRODUCTION

1.1 Background

Major Depressive Disorder (MDD) is a mental disorder defined in the Diagnostic and Statistical Manual (DSM) (American Psychiatric Association [APA], 2013). The DSM contains symptoms and diagnosing criteria of mental disorders for the purpose of having a shared definition to lead psychiatric research. Criteria symptoms for MDD include patients reporting a depressed feeling or hopelessness on a majority of days, a decreased magnitude in pleasure from hobbies, insomnia, loss of energy and a decrease in thinking speed or ability to make decisions (APA, 2013). The current diagnosing procedure for MDD is based on the evaluation of a patient in a psychiatric interview by a trained clinician. There are also scales such as the Hamilton depression scale and the PHQ scale. These scales can be used by an interviewer or be self-reported to measure MDD in research populations or as a screening criterion (Cameron et al., 2008). Despite a shared criteria for diagnosing, 25% of people with MDD remain undiagnosed (Epstein et al., 2010). Barriers stem from a lack of patient self-awareness, social stigma, and physician bias (Epstein et al., 2010; Tlachac et al., 2021).

In order to identify and assist in diagnosing methodology, research in this domain focuses on computer-based modelling methods for MDD detection, including feature extraction and model algorithm development for MDD detection. These models can include facial recognition (Schultebraucks, et al., 2022), audio processing (Shinohara et al., 2021) and language processing (Uddin, Dysthe, Følstad, & Brandtzaeg, 2022). Research focusing on the use of language utilizes data from text sources or audio recordings. In Tlachac et al., (2021) participants were given options to submit scripted audio, unscripted audio, text messages, and social media posts. 90% of participants submitted scripted audio, 78% submitted unscripted audio, and other categories were below 50%. Audio files are a data form people appear comfortable with for depression detection screening applications.

Features from patient acoustic data correlate with the psychomotor DSM symptoms of MDD such as slowed speech and a decrease in pitch variability throughout (Kanter, Busch, Weeks, & Landes, 2008; Low, Bentley, & Ghosh, 2020). Low et al., (2020) investigates mental health detection from audio research; including

63 research papers for MDD detection in the ten years prior. From these reports the most frequent features with the highest correlation to MDD are jitter, shimmer, and F0 variability. Additional low-level features with correlation to MDD are split into the categories of source features, filter features, spectral features, prosodic features, and time-based behavior features. In addition, the features with the highest correlation are mapped to control of the vocal tract and emotion recognition research (Low et al., 2020). Additional features such as phoneme vowel spacing are current leads in research for MDD detection (Muzammel et al., 2020; Yamamoto et al., 2020). Additional research focused on emotional recognition proposes higher level features with the goal of increasing MDD detection accuracy, including vitality, a feature proposed by Shinohara et al., (2021). Vitality is derived from a process that combines low level features into scores for joy, sorrow, calm and excitement that are then combined into the proposed vitality feature that has a reported correlation of $r=-0.33$ with a p of <0.05.

Other research in this domain focuses on model development or selection using historical low level acoustic features for MDD detection. Models used in recent research include decision trees (Chen & Pan, 2021), deep learning Neural Networks (NN) (Schultebrauckset al., 2022), Convolution Neural Networks (CNN) (Vazquez-Romero & Gallardo-Antolın, 2020; Huang, Epps, & Joachim, 2020), logistic regression (Cohn et al., 2009), and Support Vector Machines (SVM) (Liu, Wang, Zhang, & Hu, 2020; Jiang et al., 2017). While deep learning has demonstrated a potential for high levels of accuracy, model selection for clinical diagnosing needs to be driven not only by performance but also by clinical requirements. Due to reported biases in current data, stemming from bias in clinicians and patients, explainable models are critical in the mental health domain (Thieme, Belgrave & Doherty, 2020). This means that applications need to select models that ensure explainability in data pre-processing and algorithm classifications (Itani & Rossignol, 2020). An explainable system allows for biases to be found and leaves the medical decision up to the physician. These works have led to current research utilizing audio for MDD classification. One gap in modern research is the connection between the varied focal points within the domain of supervised mental health machine learning detection. This experiment tests previously discovered acoustic features in combination with historical

values and using traditional model algorithms to determine performance impact as a clinical diagnosing tool for MDD.

1.2 Research Project & Problem

Screening patients and research study participants for MDD currently requires a psychiatric interview for DSM diagnosing criteria or the use of the PHQ-8 questionnaire. Both methods are impacted by reporting bias from the patient and potential bias from the interviewer. Determining a method to screen patients for MDD in order to prioritize or otherwise aid in their connecting to a psychiatric health care provider for treatment rather than waiting for an initial screening would constitute an improvement in mental healthcare efficiency, helping patients and clinicians alike.

Screening using models that are trained with features extracted from audio recordings offers a low cost and informative method for MDD detection. Some of the highest reported preforming models using acoustic features for MDD classification rely on composite features tracking emotional load and phoneme-based features using deep learning models. However, deep learning models violate the need for the explainability of a diagnosis that a clinical tool requires. The clinician must be able to understand the factors in a patient's classification. Research on the accuracy and recall of MDD detection using traditional prosodic, time domain and some spectral features using deep learning and traditional machine learning algorithms indicate that the performance of the deep learning models are only slightly ahead of traditional algorithms. Without rigorous testing, newly developed features may yield unknown impacts on the performance of SVM and MLR models for MDD classification.

Therefore, the research question of this study is: "What is the impact of additional acoustic features measuring emotional load and vowel shape, when combined with traditional features, on the recall of supervised models trained for binary classification of Major Depressive Disorder?"

1.3 Research Objectives

The aim of this research at its inception is to investigate the impact of composite features representing emotional load and phoneme shaping on Major Depressive Disorder (MDD) detection using traditional machine learning models; Multivariate Logistic Regression (MLR) and Support Vector Machines (SVM).

The original null hypothesis was if two supervised models are taught, one using MLR and the other using SVM algorithms, for binary classification of patient audio files to detect Major Depressive Disorder utilizing traditional acoustic low level features[1] and composite features representing emotional load and phoneme spacing, the developed models will not achieve a statistically significant ($p<0.05$) increase in recall of the MDD class compared to the baseline models taught only with the low level features[1].

However, based on the literature review conducted in this research area, it is apparent that modern research into vowel shaping for MDD detection relies on manual linguist phoneme mapping or automatic tools that still require a trained individual to correct. With an available dataset containing 189 audio files with an average duration of 16 minutes, phoneme mapping was outside the available resources of this project. Due to this, the research problem was refined to: "What is the impact of additional acoustic features measuring emotional load to traditional features on the recall of supervised models trained for binary classification of Major Depressive Disorder?"

The, null hypothesis is also updated for the new objective to: if two supervised models are taught, one using MLR and the other using SVM algorithms, for binary classification of patient audio files to detect Major Depressive Disorder using traditional acoustic low level features[1] and a composite feature representing emotional load, then the developed models will not achieve a statistically significant ($p<0.05$) increase in recall of the MDD class compared to the baseline models taught only with the low level features[1].

The high-level objectives for this research of the new null hypothesis are those below:

1. To identify and determine an open-source method to extract composite acoustic features that correlate to emotional load of a patient from patient speech.
2. To determine an open-source method to extract phoneme structure features of a patient from patient speech.
3. To train and test a SVM model using the low-level feature[1] group
4. To train and test a MLR model using the low-level feature[1] group

[1] Low level feature group [1] : jitter, shimmer, F0 variability, mean pause duration

5. To train and test a SVM model using both the low-level feature[1] group and the composite higher level acoustic features.
6. To train and test a MLR model using both the low-level feature[1] group and the composite higher level acoustic features.
7. To determine the impact of adding composite features on recall of the MDD class.

1.4 Research Methodologies

This study uses secondary research methodologies. The data used was gathered by the University of Southern California and makes up the DAIC-Woz benchmark dataset (Gratch et al., 2014). Acoustic features extracted using the Covarep open-source library are provided within this dataset, however this experiment conducts its own extraction process from the patient audio files provided. The outcoming features are all quantitative in nature and are derived from the mathematical representation of the sine waves.

In addition, a synthesizing of previous research into emotion recognition and acoustic feature classifiers for MDD detection is conducted for the new emotional load composite feature. The result of this process is discussed in the literature review chapter. The MLR and SVM models along with their parameter values were determined from this secondary literature review as well. The experiment conducted between the models including the composite feature and that with only traditional feature groups provided empirical data of the recall, precision, accuracy, and F1 scores for MDD detection. The hypothesis can then be rejected or accepted based on the comparison of these results.

1.5 Scope and Limitations

The scope of this research falls within the supervised machine learning domain focused on multivariate logistical regression and support vector machines models for Major Depressive Disorder (MDD) detection. This project aims to develop a model that follows clinical best practices for a screening or diagnosing assistive tool that can be used in clinical settings. Furthermore, but the model would not entail specialized audio equipment and instead use features that could be extracted with microphones of various hardware specifications.

This study uses only the DAIC-Woz benchmark dataset to limit the impact of hardware and more importantly speech scenario differences. Scenario differences include interview, story reading, photo description and other scenarios, which have been shown to impact feature correlation with MDD as will be discussed further in the literature review chapter. The choice of using a single dataset limits the language to only English with a further limiting factor of all accents are of those within the United States of America. The initial data collection took place in Los Angeles California with native born English speakers. Additionally, this study assumes that the participants answered questions honesty in their PHQ-8 survey and did not attempt to consciously alter or otherwise mask their speech or vocal patterns. This assumption was taken since participants were granted pseudonymization and confidentially alongside the ability to consent to their data being shared in the benchmark dataset beyond the initial research project. It is further necessary to assume that a PHQ-8 score of greater than or equal to 10 indicates MDD with the confidence level necessary in the psychology domain for diagnosing criteria (Kroenke et al., 200).

This project also inherits limitations from the original collection procedure of the DAIC-Woz database. No additional measures beyond the MDD PHQ-8 score were taken, though participants were removed if they self-reported any additional mental health disorders, physical disability or intellectual disability that could impact the features gathered. Since no other official screenings were conducted, this experiment cannot determine if depression alone is the cause of acoustic changes. The unknown impact of additional elements would also be present in a possible clinical scenario and as such was determined to be acceptable to the scope of this study.

Programming in the study was conducted in MATLAB using the COVERAP open-source library and Python. Python libraries included librosa 0.8.1, opensmile 2.4.1 and parselmouth 0.4.1 for audio feature extraction. The parselmouth library allows python to run Praat commands through installed Praat software version 6.2.17. Additional python libraries and their version numbers used in this study are listed in the Appendix A.

1.6 Document Outline

This research will review in some depth the state of current research, the methodologies of this experiment and the results achieved. Chapter 2 will focus on additional research in the domain of computational applications for mental illness

detection with a focus on MDD. In addition, Chapter 2 will detail and explore the research in audio features used for MDD detection, the models used in current research and the situational impacts such as age, additional mental health disorders and style of speech audio gathered on MDD classification models.

Chapter 3 details the methodology and processes used in this experiment. Chapter 4 reports the results of the baseline and composite feature models. This chapter will also provide discourse of these results and how they relate to the null hypothesis declared in this chapter.

Chapter 5 contains a summary of state-of-the-art research, the research problem and the results found. This final chapter will also discuss the potential of future work and recommendations to conduct research beyond the scope of this project to further pursue the research problem examined herein.

2. LITERATURE REVIEW

Upwards of 25% of Americans with MDD are undiagnosed (Epstein et al., 2010). MDD has a range of symptoms including depressed mood , decreased pleasure, insomnia, loss of energy and a slowdown of mental speed (APA, 2013). Epstein et al., (2010) conducted focus groups for a total of 146 adults to understand their process in diagnosing MDD . The experiences of the participants in the study were grouped into three themes by Epstein et al., (2010): “knowing”, “naming”, and “explaining.” The first two of these are relevant to diagnosing delays. The knowing phase refers to when symptoms are present, but participants were unaware that they were showing symptoms. Many reported that they became aware when family members, professions or friends commented that something was not right which triggered exploring for a cause, i.e., the naming phase. Depending on the person, this phase ranges in length, and the goal of designing a clinical tool is to assist in screening people for depression that falls into the knowing or naming category, those unaware of their symptoms or searching for the cause of their symptoms.

This literature review goes through the clinical requirements that need to be considered due to the scope of the project targeting a clinical screening or diagnosis assistive tool in the first section. The second section will focus on models that use text or video data and explain why the decision was made to go with an audio based application. Following that, the third section covers common audio features used in MDD detection models with attention paid to those features that appear in emotion recognition research. The fourth section covers possible impacts in acoustic research focusing on age, gender, and speech style. The final section of this literal review focuses on the model algorithms used in MDD classification from both the deep learning and machine learning scope and explain why SVM and MLR were selected for this study.

2.1 MDD Detection Ethical Concerns and Clinical Requirements

When working in the applied computing domain for mental health applications, there are critical ethical concerns and clinical needs required for diagnosing assistive tools. Thieme et al., (2020) conducted a systematic review of research into machine learning and mental health. Their final corpus contains 54 articles regarding psycho-

social functioning mental illness disorders, including MDD, with varying data types and algorithm approaches. While not focused on MDD detection, the clinical and ethical concerns concluded are shared within the broader scope of mental health detection.

One discovery was evidence in multiple studies that clinicians trusted machine learning recommendations, especially newer clinicians, even when unable to explain the reasoning of the model (Theime et al., 2020). Additionally, there are concerns over models learning the biases of their datasets. Explainable models are necessary to prevent bias going unchecked, resulting in prevented access to medical care. (Theime et al., 2020) For these reasons explainability needs to drive decision making in this still developing area of research. Another concern covered by Thieme et al., (2020) regarded that application base research often is driven by technical possibility without concern to the target population. This includes targeting models using data types and methods that the target population is willing to engage with when presented the option. This aspect is what drove the decision of this study to focus on acoustic feature models over hybrid, visual or text models as will be explained in the next section.

2.2 Non-Audio Based MDD Detection Models

MDD detection models have been developed using visual data, audio data, demographic data, and text-based data. Visual data targets facial tracking or body position/posture tracking (Cohn et al., 2009; Lin et al., 2020). Audio data feature models have been developed targeting environmental noise (Di Matteo et al., 2020) or audio from patient speech (Liu et al., 2020). Text based models use natural language processing methods with corpora built from social media posts (Tong et al., 2022) or transcriptions of patient audio files (Xezonaki et al., 2020). Demographic data uses features related to race, height, weight, age, and gender to predict depression and is deployed in research as a baseline model, as in Pan et al., (2019). Applications have also been developed combining different data groups including text and audio data from interviews (Shen, Yang, Lin, 2022) or audio and visual data from interview recordings (Cohn et al., 2009).

2.2.1 Overview of Research in Visual Data Models

Visual based data models in MDD classification are associated with facial expression tracking. Cohen et al., (2009) is a research study that investigated facial expression mapping to classify MDD patients. While older, it is a critical research paper, introducing the of using response time durations and fundamental frequency (F0) to predict MDD (Cohen et al., 2009). The main portion of this study, focused on facial feature tracking and testing models built from manually mapped action units and active appearance model mapping for feature extraction. The first model required trained technicians and research to manually map the action units. The active appearance model required a trained technician to manually label 3% of the keyframes, and the rest of the video was automatically aligned. Accuracies for these models are 88% and 79%, respectively. While the accuracy reported is high while using an explainable SVM model, they require a trained expert to manual map at least 3% of the key frames to achieve these scores. While this is well-suited for a longitudinal research study with MDD patients, the requirement of a trained expert prevents this from being a data approach suitable for an on demand diagnosing assistant tool in a medical practice or an online screening metric.

However, current research has focused on using automatic extraction for facial tracking, utilizing software OpenPose and OpenFace as in the research by Lin et al., (2020). This research added body fidgeting and self-adaptor movements into an MDD classifier. The self-adaptor movement model achieved an f1-score of 83.38% in a linear regression threshold classifier. Lin et al., (2020) mentioned that the facial movement tracking required smoothing due to failed extraction frames; and that their fidget detection process requires additional tunning with more participants. This model shows visual data developed in a way to enable use in more screening purposed applications. However, there is currently no research indicating whether or not the general population would tolerate being video recorded at home or in a medical clinic for general screening purposes. As will be discussed in the text base model section, audio data has demonstrated this aspect.

2.2.2 Overview of Research in Text Data Models

Text based models have shown promise in MDD detection models and require fewer manual annotations compared with visual data models, which require tuning for

each participant. Textual data often occurs in models using features from transcribed interviews with patients or from social media posts. Transcription models suffer from the need for manual transcription, as current research reports the accuracy of automatic transcription services upwards of 45.88% mismatch transcription (i.e., lost meaning compared to manual transcription), Louw, (2021). This percentage decreases in low noise background to 35.71%. In non-interview settings with only a single speaker, this number can drop to 20%. However, as presented by Louw, (2021), current software is unable to be used beyond a first draft for researchers to later fine tune. This would rule out transcription based models within the scope of designing a model that can function with minimal human annotation.

Other text-based models use social media posts or patient text messages to predict or track severity of MDD. These models don't require any human annotation, as they are in text format. Tong et al., (2022) conducted an experiment using twitter posts for depression detection with the aim of proposing a cost-sensitive boosting pruning tree. The experiment was concluded with a f1 score of 0.869 reported, however the MDD class was determined by a user's post history containing a phrase similar to "I'm depressed." (Tong et al., 2022). As argued by Kanter et al., (2008) this will lead to misclassification of the MDD class due to the colloquial meaning of the word depression which may not be indictive of MDD. Burdisso, Errecalde & Montes-y- Gómez, (2019) developed a new model, the SS3. This experiment attempted to avoid this increase in misclassification by only using users with a history of saying they were depressed and mentioning a diagnosis within the post. The f1-score achieved was 0.61 (Burdisso et al., 2019).

While text-based scores in social media post are comparable with audio-based features models in current research participants are shown in an experiment conducted by Tlachac et al., (2021) to prefer submitting audio data for MDD detection. Tlachac et al., (2021) reported that out of 70 participants, 90% completed the optional scripted audio phrase task, and 78.5% completed the unscripted audio prompt for their MDD detection application. Only 44% of participants opted to share their text messages, 15.7% opted to share their twitter post and 0 participants were willing to share their Instagram post. With the goal to reach more people to assist in MDD screening and detection it is important to use methods that participants are willing to be screen by. The number of participants demonstrated as willing to complete audio recordings for

MDD detection voluntarily indicates that audio features are a viable data type for screening applications in this regard.

2.3 Audio Features for MDD Detection

Acoustic features have been shown to be impacted by MDD along with other mental health disorders. Low et al., 2020 presents a systematic review of automatic classifiers for mental health based on acoustic features from speech. Acoustic features for human speech fall into five categories: source, filter, spectral, prosodic and time features. Source features are impacted by the glottis in the larynx and include jitter, shimmer and Harmonic-to-Noise Ratio (HNR). Filter features are those impacted by the vocal and nasal tract shaping the sound. Features in this group are vowel spacing, formants, and formant frequency ranges. The most common spectral feature is the Mell frequency cepstral coefficients (MFCCs) and their first and second derivatives. Lower MFCCs represent the vocal track while higher MFCCS represent aspects of the vocal fold source (Low et al., 2020). Prosodic features are those related to pitch or perceived intonation including F0, intensity and energy change. Time features are often grouped within the prosodic category. These refer to response time, pause time and other features representing aspects of the rhythm of speech (Low et al., 2020).

In the systematic review conducted by Low et al., (2020) the papers regarding MDD showed overlap in low level features selected for models. These included jitter, shimmer, F0 variability and mean pause duration. These features are associated with the psychomotor impact of MDD which increase the time necessary for the brain to signal for the change in the larynx and vocal cords (Low et al., 2020). These features make up the historical feature group for the baseline model in this research project

The rest of this section looks at research in the emotion recognition domain and in features mapped by phonemes to express some of these low level acoustic features as higher level features with stronger correlations to MDD detection

2.3.1 Acoustic Features for Emotion Recognition

One of the DSM defined symptoms for MDD is low mood or a depressed feeling for a majority of days. With the context of this symptom being present for a majority of days it can be assumed that a patient with MDD would have this at the time their speech is recorded. Identifying emotions from acoustic features of recorded

speech is an ongoing area of research that overlaps with MDD detection due to the symptoms of MDD containing low mood and decreased excitement.

One of the primary papers that contributed to the emotional feature aspect of the research problem in this study is Shinohara et al., (2021). This study purposes two new indexes, vitality, and mental activity, for MDD classifications. The study consisted of 44 participants who recorded defined Japanese phrases. From these recordings emotional recognition software ST Ver 3.0 was used to rate the strength of the emotions in the recording. The emotions were: anger, sorrow, joy, calmness, and excitement. The mental activity index did not show correlation with MDD, however, as a longitudinal based index tracking change in vitality this cross-section research study was not designed to adequately test it. Vitality on the other had did demonstrate a negative correlation of -0.33 ($p<0.05$) and is defined in the formula below.

$$vitality = 0.60 \times \frac{joy}{joy + sorrow} + 0.40 \times \frac{calm}{calm + excitement}$$

Figure 2-1. Vitality Equation

However, during this experiment vitality was unable to be calculated due to the proprietary nature of the ST Ver 3.0 software. The software itself is an element in US Patent number 7340393 B2 filed on March 4, 2008. This patent indicates that the software uses features based on change of amplitude, tempo of speech, and the power spectrum, for intonation tracking as features for emotion recognition. While vitality itself cannot be recreated, it does support the notion that emotional recognition features have a correlation with MDD detection.

Schuller, Rigoll & Lang, (2004) conducted a research experiment studying emotion recognition in audio recordings of participants within an automotive environment. Their aim is to propose a novel network model combining acoustic and linguistic features to determine the emotional state of a driver. Audio was gathered in both German and English from 13 participants, one of which was female, acting out different emotions. Actors were asked to label their recordings with the appropriate emotion: anger, disgust, fear, joy, sadness, surprise and neutral. The recordings were then preprocessed, and over 200 features were extracted containing statistical metrics of silences, pitch, energy, spectral energy in frequency bands among them. Linear Discriminant Analysis (LDA) was used to determine the top 33 informative features. Relative pitch and pitch statistical metrics made the top 12 features. Additional notable

feature in the top 33: mean duration of silences, relative energy, the zero-crossing rate, and spectral energy in pitch windows below 250 and below 650 Hertz.

2.3.2 Acoustic Features Phoneme Based

In the original research problem of this study one of the main objectives was to understand the deep learning feature extracted by Muzammel et al., (2020) using a CNN to map vowel and constant spacing. This objective aimed to propose a vowel spacing based phoneme feature that could be used in traditional machine learning models. Higher level deep learning acoustic features in MDD detection are high informative as they capture stresses within each sound and word which through deep learning can be taught to a model. Muzammel et al., (2020) extracted the phoneme spacing feature through Praat software by first labelling the voiced and unvoiced portions of audio using a built in Praat segmentation function with the recommended default parameters for human speech. The second step involved phoneme level alignment to the audio using the transcripts from the DAIC-Woz dataset and the Carnegie Mellon University pronouncing dictionary. However, when replicating this process there was no indication on which tool kit or Praat function was used for this step which is traditionally done through manual mapping. An open-source automatic phoneme transcriber was located and applied to the audio in the DAIC-Woz database (Corretge, 2022). The results required extensive manual fine tuning and was then determined to be outside the resources of this study.

While it is no longer a target feature the phoneme space mapping in Muzammel et al., (2020) illustrates how beneficial higher-level phoneme based features are in MDD detection. Three proposed models were developed, one targeting vowel phonemes, the second targeting consonants phonemes and the final targeting the combination of both. In order the achieved accuracies were 78.77%, 80.98% and 86.06%. These are in line will other deep learning models that use over twenty features. The recall scores for the MDD class were also of note and scored 66%, 64% and 73% respectively which as will be seen in the deep learning machine learning section below is higher than other deep learning approaches. While this study was unable to replicate this feature, it demonstrates a performance that would benefit from further research into possible applications outside of the targeted deep learning approach or with additional features.

2.4 Impacts on Audio Features for MDD Detection

Working with acoustic speech features requires consideration for other factors impacting patient speech. Recent research within the domain of emotion recognition or MDD detection have aimed to study unique cases. These situations include studying the impact of younger age groups, gender, and the speech exercise taken during patient speech recordings. This section of the literature review evaluates these recent studies to determine common features effective for these situations for feature selection in the aim of a suitable generalized model.

2.4.1 Age Range

Research focusing on MDD detection is often preformed with data from adult participants due to data security concerns and population availability. DAIC-Woz contains 189 adult patients of an unspecified range. However, depression disorders also occur in children and the elderly. McGinnis et al., (2019) reports on research in MDD detection for children from 3 to 8 years old. Audio was recorded in an altered Trier Social Stress Task to induce anxiety and stress in the children conducted during a home visit. The features reported to have high impact on model classification for internalizing disorders (for children this includes symptoms of depression and anxiety) were the zero-crossing rate, Mel frequency cepstral coefficients, dominant frequency, mean frequency, perceptual spectral centroid, spectral flatness, and signal energy in 5 frequency bandwidths (McGinnis et al., 2019). These features are reported as correlated to MDD in adult studies as discussed in the acoustic feature section of this literature review.

Lortie et al., (2015) is another study that investigated how age impacts the perceived quality of a voice through amplitude and frequency. Participants ranged from 18 to 75 years old. Age did impact features in sustained vowel sounds, but in the context of continuous speech Lortie et al., (2015) reported no statistically significant change in participant age and feature correlation to MDD.

These two studies suggest that classification models for children, adults and older adult populations regarding MDD, or depression disorders in general, share informative features.

2.4.2 Gender

Beyond age the next demographic in question for impact is patient gender. While many features such as the range of the absolute values of pitch in a patient's audio aim to capture data that is independent of the starting fundamental frequency, which differs on patient sex, other features such as the Mel spectrogram may give pitch or vocal track elements higher weight and therefore act as indicators of gender (Bailey & Plumbley, 2021). Bailey & Plumbley, (2021) examined the DAIC-Woz dataset for gender bias and purposed methods to counter the bias found. The DAIC_Woz dataset has 44 female participants with a ratio of 5:8 for MDD participants to control participants. The 63 male participants ratio of MDD to controls is 2:7 (Bailey & Plumbley, 2021). To adjust for possible bias, researchers split the gendered participants into two datasets for two different models. This was the case for Jiang et al., (2017) whose work is further explored in the next section. In situations where this is not desirable due to limited data or the objective of a single model for generalized screening, Bailey & Plumbley, (2021) recommend machine learning models to limit possible bias through down sampling to quadrants (MDD male, control male, MDD female, control female) rather than the binary classification label alone. While their experiment did demonstrate a difference in performance on gender balanced data for the deep learning model, it was not statistically significant enough that to conclude the Mel spectrum added gender by proxy into the model. However, the difference does illustrate gender to be a possible concern in the domain that needs to be examined further.

2.4.3 Audio Recordings' Speech Exercise

Jiang et al., (2017) investigated the impact of speech activity types in MDD detection of various classifiers. The study had 85 healthy controls and 85 participants with MDD, the ratio of women to men was 51:34 and 53:32 respectively. The participants were recorded during a psychiatric interview with 3 positive focused questions, 3 neutral questions and 3 negative questions. The second speech task was to describe in their own words four photos of faces expressing positive, neutral, negative, and crying emotions. Finally, participants recited a reading passage with sections targeting each emotional undertone. Acoustic features were extracted from all these recordings using openSmile software with a total feature count of over 1500. The

feature values were normalized and went through PCA before being passed to the classifiers.

While a novel classifier was proposed, the critical insight provided by Jiang et al., (2017), for this research study, is the performance of the SVM classifier for each speech task. Jiang et al., (2017) reports for the male SVM classifier the picture task achieved the highest accuracy at 70.83% with a mean recall of 67.65%. The interview task in comparison achieved 65.74% accuracy and a recall of 69.93%. For the female SVM classifier, the highest accuracy was in the interviews at 67.31% with recall at 68.63%. For both the male and female SVM classifiers, the highest recall was found in the reading task, near 75% for both, and the interview providing the second highest at around 69%. The interview style task is the most commonly available and frequently used in other research papers. Jiang et al., (2017) suggest in their results that predetermined phrase reciting may prove to be more reliable going forward. However, interview data has reported values that show promise worthy of further development.

2.5 Machine Learning Model Selection

MDD detection has been approached with both machine learning and deep learning models. This section of the literature review looks at a selection of papers using different model types. The first section investigates deep learning approaches, specifically work done in Convolution Neural Networks (CNN) application research for MDD detection. The second section delves into Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT) models.

2.5.1 Deep Learning MDD Detection Models

Deep learning approaches within MDD detection are often limited in performance metrics due to limited data. However, recent research into CNN approaches have demonstrated successful models that achieve comparable scores in accuracy. The work of Huang, Epps and Joachim (2020) and the work of Srimadhur and Lalitha (2020) both studied aspects of this area of research. Huang et al., (2020) present a novel framework evaluated with naturalistic and clinically gathered data. Four models with this framework were developed, trained and tested for each dataset (DAIC-Woz and SH2-FS) and each feature group (formants, spectral centroid frequencies, MFCCs, and delta MFCCS) (Huang et al., 2020). Srimadhur and Lalitha

(2020) present research testing a spectrogram-based CNN model against an end-to-end CNN model for MDD detection using the DAIC-Woz dataset with spectrograms and MFCC coefficients as features. Two spectrogram CNN models and six end to end CNN models were developed, with the models containing larger kernel sizes reporting higher accuracies. The performance metrics for the DAIC_Woz trained models developed in Huang et al., (2020) and the identified highest preforming model for each type in Srimadhur and Lalitha (2020) are reported in Table 2-1.

Table 2-1. Deep Learning CNN Performance Metrics

Paper	Model	MDD F1-score (%)	MDD class Recall (%)	Accuracy (%)
Srimadhur & Lalitha (2020)	Spectrogram CNN	66	77	61.32
Srimadhur & Lalitha (2020)	End-to-End CNN	78	80	74.64
Huang et al., (2020)	FVTC-CNN Formants	46	Not Reported	73.5
Huang et al., (2020)	FVTC-CNN Spectral Centroids	42	Not Reported	69.6
Huang et al., (2020)	FVTC-CNN MFCCs	40	Not Reported	74.8
Huang et al., (2020)	FVTC-CNN delta MFCCs	37	Not Reported	75.2

2.5.2 Machine Learning MDD Detection Models

Recent research into MDD detection using machine learning algorithms target investigation into situational variables regarding data collection, model algorithms including fusion models, and identifying performance increasing subsets of features or new composite features. Lie, Wang, Shang & Hu, (2020) investigate a new fusion model approach and situation variables in speech type. Jiang et al, (2017) as mentioned previously looks into the impact of the speech type gathered as well as impacts on emotional undertone of the speech task. Jiang et al., (2017) reports performance metrics for three machine learning classifiers: K-nearest Neighbours (KNN), Gaussian Mixture Model (GMM), Support Vector Machine (SVM). The performance metrics of all three classifiers under the interview speech style for each gendered model is reported in Table 2-2

Liu et al., (2020) purposed a novel machine learning approach to MDD detection where a binary tree was constructed where the nodes are SVM models. The

voice segments are tested with the SVM model and with their specificity and sensitivity split at the node. It was reported in the study that more than half of the test participants were male (37 to 16 female) and with gender specific models this led to a discrepancy between the perspective performance metrics. These metrics are shown in Table 2-2 alongside those from Jiang et al., (2017)

Table 2-2. Machine Learning Model Performance Metrics

Paper	Model	MDD Specificity (%)	MDD class Recall (%)	Accuracy (%)
Liu et al., (2020)	Interview SVM Baseline Male: Female	Not Reported	Not Reported	59.2:54.2
Liu et al., (2020)	Binary tree Fusion Male: Female	Not Reported	Not Reported	70.5: 63.2
Liu et al., (2020)	New Binary Tree Fusion Male: Female	Not Reported	Not Reported	75.8: 68.5
Jiang et al., (2017)	KNN Interview Male: Female	63.73:60.68	60.07:62.79	61.95:61.75
Jiang et al., (2017)	GMM Interview Male: Female	59.15:60.89	61.81:66.35	60.44:63.68
Jiang et al., (2017)	SVM Interview Male: Female	69.93:68.63	61.28:66.04	65.74:37.61

2.5.3 Comparing Machine Learning and Deep Learning

The accuracy scores of the deep learning CNN models, as shown in Table 2-1, were higher than the reported accuracy in the machine learning section, Table 2-2. However, the MDD class recall values of Huang et al., (2020), were lower than the recalls reported in Jiang et al., (2017). Currently there is a lack of research evaluating both deep learning and machine learning approaches on the same dataset and in the same context. As such, no definite conclusion can be made, however, with reported performance metrics both deep learning and machine learning warrant further research. At present the objective of the model as a screening tool warrants preference to models with higher degrees of explainability and higher recall performance. For these reasons the scope of this study remains in the machine learning algorithm approach.

2.6 Conclusion

This literature review covered a range of topics within the scope of MDD detection. First the ethical concerns and clinical requirements were discussed covering the need for explainability and the need to keep the target population in mind during the decision-making process of this research study. The second section covered the range of data types including visual, text and audio-based features MDD classifiers are built with. Research in each feature type was present and the conclusion was that audio models while not the highest in performance metrics allowed for no manual annotation and were shown in Tlachac et al., (2021) to be preferred when the test population was present with the choice of what features to provide.

Following the research leading to the decision to go with an audio-based model, research was presented covering the historical acoustic features used in MDD detection. In addition, features associated in the emotional recognition domain and the concept of higher order features mapping to phoneme were presented. The next section presented current research into possible impact of age, gender and the recorded speech exercise has on model performance. While these areas of research are still in their preliminary stage, they present cases and points of interest to keep in mind during the experiment phase of this study. Finally, research in deep learning and machine learning applications of MDD detection were presented and the deep learning models showed higher accuracy the machine learning approaches showed higher recall. For this study recall of the MDD class is the targeted performance metric due to the cost of not recommending treatment for someone with MDD being higher then falsely recommending treatment for a healthy individual. For this reason and the increase in explainability machine learning models, SVM and LR were selected for this study.

The next chapter with further expand on the feature and model selections while explain the steps and methodology of the experiment.

3. DESIGN AND METHODOLOGY

This chapter contains information regarding the dataset chosen for this experiment, the software and programming languages used for feature extraction and model building and defines the method of evaluation for this experiment.

3.1 Dataset

The DAIC_Woz benchmark dataset from the University of Southern California (Gratch et a., 2014) was selected for this experiment. This section is going to cover an overview of the context in which the dataset was original gathered, the pre-processing sets for the audio files, the feature extraction methods chosen, and the steps for assembling the feature dataset for this experiment.

3.1.1 DAIC-Woz Dataset

Gratch et al., (2014) reports on the creation procedures of the Distress Analysis Interview Corpus (DAIC) dataset. The DAIC dataset was assembled in a greater project regarding computer agent interviewers for psychiatric disorder detection. The original participants in this study are from the Los Angeles area and are members of the general public or veterans of the United States armed forces. Metrics included screening for MDD, PTSD and anxiety (Gratch et al., 2014). Participants were involved in face-to-face or teleconference interviews, Wizard-of-Oz interviews in which a human agent drove the virtual agent talking with the participant, and an interview with an autonomous driven computer agent. The video and audio data from the Wizard-of-Oz interviews make up the DAIC-Woz benchmark dataset that has been selected for this experiment.

The DAIC-Woz dataset has 189 participants. There are 5 participants whose data is incomplete or was gathered with technical issues (patient ids 373, 444, 451, 458, 480), which were removed from the dataset for this experiment. Participant 402 also had technical issues, but it was related to the facial camera and does not impact the audio file used in this experiment.

Within the provided data for this experiment timestamps from the annotated transcript, MDD scores, MDD binary classification label, gender and the raw audio files are used.

3.1.2 Audio Feature Extraction

In the feature extraction phase of this experiment, multiple software and programs are involved to create partial datasets, all using the patient id as the primary key. The datasets are joined on the patient id to create the final feature dataset.

From the original data the MDD scores, binary classifications, gender, and patient ID are kept for the demographic subset of data used for balancing purposes. These features are not given to the models.

The second dataset extraction is the duration-based features. Ellie, the avatar interviewer, can be heard in some of the audio files, but not in others. In order to ensure Ellie was completely isolated, the transcript time stamps are used to remove the periods of audio in which Ellie can be heard. These timestamps are also utilized to accurately determine the duration and response pause time between the interviewer and speaker. The preference would have been to use energy detection in the signal to determine this value, but the feature was not accurate enough on some of the participants' audio to define when Ellie stopped speaking. The mean, standard deviation and median of the pause duration, and overlap duration were calculated and added to the final feature set.

COVAREP is an open-source MATLAB library and is used to extract 52 acoustic features for the final set. These include the normalised amplitude quotient (NAQ), quasi open quotient (QoQ), differential of first two harmonics (H1H2), parabolic spectral parameter (PSP), peak Slope, Mel-Cepstral coefficients (MCEP) 1-24, phase distortion mean (HMPDM) 0-9, and phase distortion deviation (HMPDD) 1–12. The statistical measures, mean, median and standard deviation are included in the final feature set for these features.

Next is the data extracted using the parselmouth python library that extracts features using the Praat software. Prior to the extraction of the novel features a parselmouth script created by Feinberg, D. (2018) entitled "Measure Pitch, HNR, Jitter, Shimer and Formants" was adapted for classical feature extraction. These

features are the statistical measures, mean, standard deviation, maximum and minimum of local jitter, local shimmer, formants 1-5, and fundamental frequency (F0).

Finally, the novel features are extracted within the parselmouth script using both parselmouth and the librosa python library. These features are included in Table 3-1.

Parameters were selected using a baseline of the recommended default values for human speech in each software's literature. Parameters were adjusted and tested to on 10 randomly selected patient audio files for comparison to generate the final values. The parameter tuned the most involved the windowing frames overlap percentage and window length for feature extraction. Features from the glottal source were extracted with a 10ms window length based on the minimal time for muscle movement to impact feature values (Feinberg, D., 2018). Formant based features were extracted with a 40ms window length. This was necessary due to the formants being wider in bandwidth frequency. The extended window length by comparison allows for the necessary resolve based on lowest expected pitch and the decrease in resolution is acceptable.

Table 3-1. Extracted Features and Relative Features

Feature	Statistical Measures	Description
Spectral Centroid	Mean, Standard Deviation (SD), Median	Shows were the central mass of the spectrum of the sine wave is
Power Spectral Density – Welch method	Mean, SD, Median	This indicates the spread of power within frequency
Mel Frequency Cepstral Coefficients (MFCC) 1-13	Mean, SD, Median	Lower MFCCs are related to the pitch and higher are related to vocal folds
Relative Spectral Cent to F0 mean and its derivative	Mean, SD,	As the Spectral Centroid was gathering in a window it was taken relative to the F0 mean
Relative Pitch Range to F0 mean	Mean, mean of the Maximums, Mean of the minimums, SD	Windows of absolute pitch range taken over the F0 mean
Relative Pitch gradient to F0 mean	Mean, mean of the Maximums, Mean of the minimums, SD	Windows of absolute pitch gradient taken over the F0 mean

All of these features, as float numerical values, are then combined into one dataset as the feature dataset used in the MDD classifiers of this experiment. The full extraction process of the features is illustrated in Figure 3-1 below, with the flow within the green section repeating for each individual patient file.

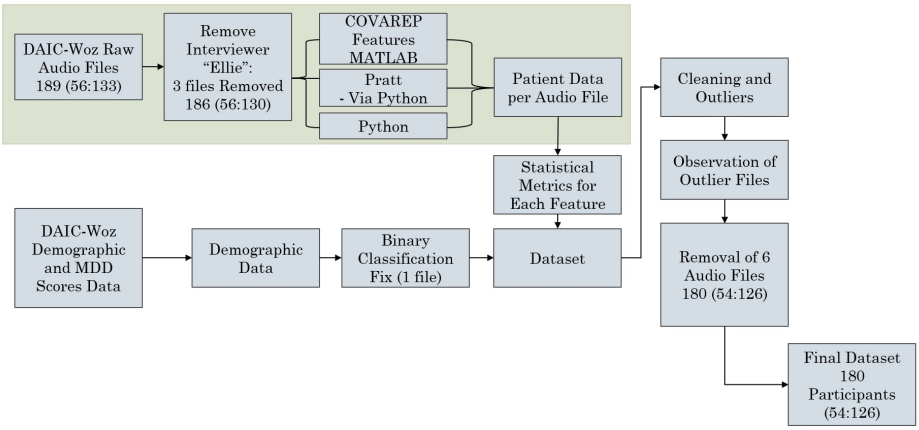


Figure 3-1: Feature Extraction Process

3.2 Machine Learning Models

3.2.1 Support Vector Machine Models

Support vector machines (SVM) have a goal to plot a decision plane for classification in the feature space with the largest margins possible. This is shown in Figure 3-2 (Kelleher et al., 2020) where the models transition from the left image to the right image in training to optimize the distance of the solid decision line (Kelleher et al., 2020). Two SVM models are built in this experiment using a linear kernel and a poly (d=3) kernel to test which performs best in the dimensions of the large feature space.

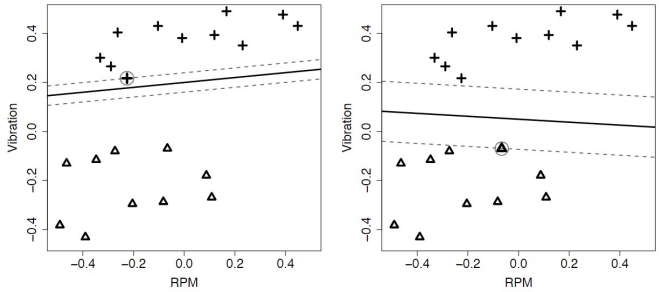


Figure 3-2: Kelleher et al., 2020

3.2.2 Multivariate Logistic Regression Model

Logistic regression (LR) assume that the prediction class is a binary value with numerical features. A LR model predicts the classification by mapping the relationship between the training features using a defined error function to improve the weights of each feature. The equation representing logistical regression is below. (Kelleher et al., 2020)

$$\log\left(\frac{p(x)}{1-p(x)}\right) = w_0 + w_1X_1 + w_2X_2 + w_3X_3 + \cdots + w_nX_n + \varepsilon$$

Figure 3-3. MLR Base Equation

The LR model in this experiment is given a random state to start at using a generate random state numbers function and uses the limited-memory Broyden-Fletcher-Goldfarb-Shanno (lbfgs) algorithm for its solver parameter. Through a grid search, 5 iterations were determined to be the best fit for both this solver and the Newton algorithm solver. The benefit of the lbfgs solver is that its better at handling saddle points in the error gradient plane. The largest con is that there is a possibility of the model not converging, however testing indicated that this was not an issue in this experiment.

3.3 Evaluation Method

To evaluate the null hypothesis, two SVM models and two MLR models must be ran. An 80% training data and 20% test data split will be used for this data. The majority class of healthy controls totals 126 and the minority class totals 57. As such, during the random data training and test split, data will be stratified to maintain the portion of controls to MDD participants. One hundred random states have been pre-generated through a number randomiser and will be used for all models evaluated to ensure the data splits are equal.

One SVM and one MLR model will be trained with only the group 1 features: shimmer, jitter, mean pause duration, F0 mean, F0 standard deviation, F0 minimum, and F0 maximum. This model will be run 100 times with the 100 random states pre-generated. The recall values and accuracy values will be stored. These values will be plotted and compared to the results of the second model.

The second SVM and MLR model will be trained with the features from the first in addition to a PCA generated composite feature from low level features with

strong emotion recognition correlation: Bands 1 – 3 spectral energy, formant frequency means, relative F0 max to F0 mean, relative spectral centroid to F0, average Formant, response duration standard deviation and F0 absolute range mean. The same model parameters will be used to train the models, storing the recall and accuracy values for comparison. A third model will be trained using features identified through a lasso regression feature selector and compared. A fourth model will be trained using the group 1 features and the additional novel features listed in section 1 of this chapter. A final model, the fifth, with be ran for both MLE and SVM using the base group 1 features and a second proposed emotional feature. This feature will use the base line of the vitality equation in the original Shinohara et al., 2021, as depicted in Figure 2-1. Without the property software the joy over sorrow component and the calm over excitement component cannot be directly recreated. Instead, the 0.6 weight will be given to the mean of the spectral centroid relative to F0 standard deviation. Spectral centroids represent the centre of mass of the spectrum, which to a listener effects the brightness of a sound. The calmness to excitement component, with a 0.4 weight, will be represented with the harmonics to noise ratio (HNR). From a listener perspective HNR impacts voice quality, vocal fry or breathiness. These final tests will also be evaluated with the 100 random states for comparison with the other models.

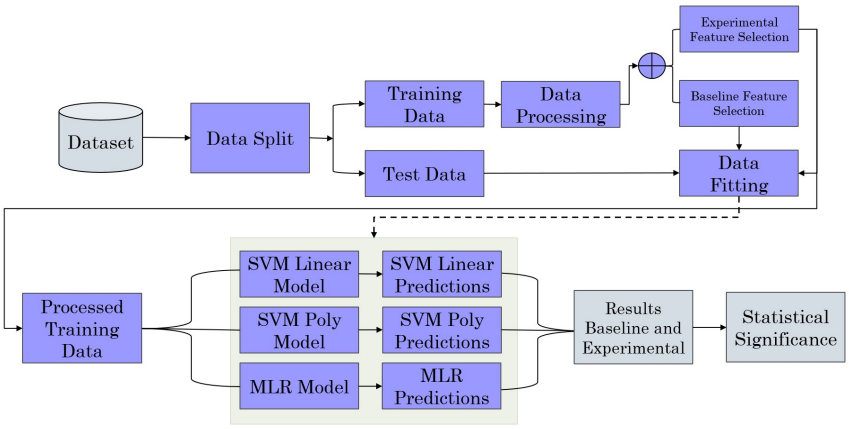


Figure 3-4: Experimental and Evaluation Process

Once all of the models are executed for reporting purposes, they will be run again with a single official training set of the DAIC-Woz dataset for reporting purposes.

The recall distributions and the accuracy distributions of the model iterations for each training set will be graphed and compared using the Wilcoxon-test to observe

any statistically significant changes- that is, observed with a p-value < 0.05. This will determine if the null hypothesis can be rejected or is failed to be rejected in this study. Figure 3-4 above illustrates the high level steps described in this section with the purple steps being repeated to generate the 100 instances of the performance metrics for comparison purposes.

4. RESULTS, EVALUATION AND DISCUSSION

This chapter will cover the results of this study's investigation. The first section will briefly discuss the spread of the MDD binary class and the decision to down sample the data before testing. The acoustic features presented in this paper will then be discussed along with presenting how well they correlate to the MDD class. The remaining sections cover direct model to model comparisons to test discussion points and the research problem of determining if a composite emotion acoustic feature can improve the recall of SVM and MLR models when ran in combination with historically frequently used low level features: jitter, shimmer, F0 mean, F0 standard deviation, and mean pause duration.

4.1 MDD PHQ8 Score Spread

Before testing of the models, the specific PHQ8 scores were investigated. The spread of the participant scores can be found in Figure 4-1. A majority of participants fall to the left to the MDD cut-off score of 10 and the average score rounds to 6 overall. Participant 409 was changed to the MDD class due to a PHQ8 score of 10 being recorded but a non MDD binary classification. Another critical point regarding the PHQ8 scale is that from 0 to 10 are the healthy controls. However, the MDD participants are spread on a scale ranging from 10 to 25 and there is a gap in this data set of scores between 17 and 24. Also, while 10 is the cut-off as mentioned in Kanter et al., (2008), the difference between 9 and 10 is not as significant as a PHQ8 score difference from 10 to 11, though one has MDD and the other does not. With a significant portion of the MDD class being near this cut-off, more noise is likely to be introduced.

To balance out the MDD and control class, down sampling the training data through random sampling was performed. The training and test data were split pre-down sampling and stratified to maintain the class proportions for the test data.

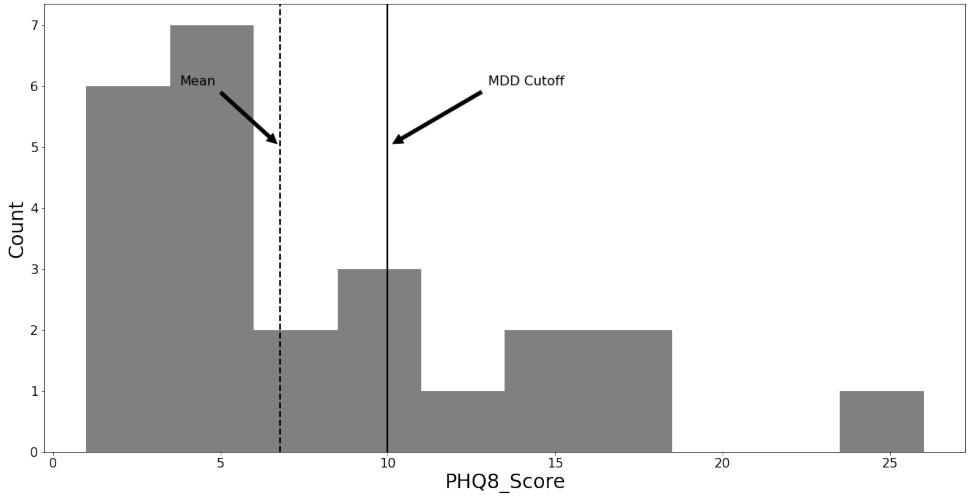


Figure 4-1. PHQ8 Scoring Distribution DAIC-Woz

4.2 Acoustic Features Extraction and Correlation

After feature extraction, the correlation between the MDD class and the feature set was explored. Table 4-1 shows the top 20 features in absolute magnitude of correlation to the MDD binary category. A full feature heatmap can be found in Appendix A Figure 1. The correlation values were determined using the Spearmen correlation test .

The Pearson correlation test was not used due to failing the assumption of normality in the data. Spearman's first assumption is met with the continuous nature of the test features. The second assumption prefers features with linear relations; however, Spearman's correlation test can be used regardless of its limit of not being able to detect nonlinear relationships. During testing a Lasso regression feature selector was developed in addition. When all of the features were evaluated with the lasso feature selector, only MCEP 23 and the response time removing overlap with the interviewer duration were selected. As such, for comparison testing, a model was developed using these two features alone as will be reported in the following sections.

The raw python based script feature extraction and model set up can be found at the follow repository https://github.com/amulligan12/Disertation-. A section of the process was also completed using the COVAREP library in MATLAB which is not sharable to the repository.

Table 4-1 Top 20 Correlated Features with PHQ Binary

Top 20 Features	Correlation with PHQ8 Binary
Mean Response Duration	0.218455092
Mean Response without Overlap	0.199915447
HMPDM 7 Median	0.163973399
Standard Deviation of Overlap Time	-0.158386688
HMPDM 10 Standard Deviation	-0.158368773
MCEP 14 Median	0.152114435
Standard Deviation of Response without Overlap	0.150327481
MCEP 14 Mean	0.146083466
HMPDM 22 Standard Deviation	-0.141616082
Max F0 Range Relative to F0 Standard Deviation	0.140945974
HMPDM 6 Median	0.140833627
HMPDM 22 Mean	0.137818805
Max F0 Relative to F0 Mean	0.135585113
Max F0 Gradient Relative to F0 Mean	0.135585113
HMPDD 10 Median	0.134468267
MCEP 10 Median	-0.133128052
HMPDM 4 Standard Deviation	-0.132085493
HMPDM 8 Mean	0.13089436
HMPDD 10 Mean	0.130000883

4.3 Baseline and Emotion Composite Feature Results

The first test in this experiment is to compare a SVM and MLR model trained on the baseline features covering shimmer, jitter, F0 variability and mean pause duration. The 100 recall results from these models can be found in Figure 4-2, and the average performance metrics are in Table 4-2. The liner SVM model achieved the highest average recall at 54.5% with the baseline features. As can be seen, the polynomial SVM model achieved the lowest recall and accuracy at 49.5% and 46.8%. This was counter the relationship of performance of the linear and polynomial SVM in Tlachac et al., (2021).

The next models developed used training data combining the baseline features with the composite feature targeting the combining of features through PCA. The features used included the relative spectral centroid to F0 mean, the response duration, the Bane 1-4 spectral energy, the means for Formant 2 and 3, the standard deviation of Formant 1 and 2, and finally the relative F0 gradient to F0 mean (Schuller et al., 2004). Figure 4-3 show the distribution of the recall performance of the linear SVM and the MLR model, and Table 4-2 contains the accuracy and recall averages.

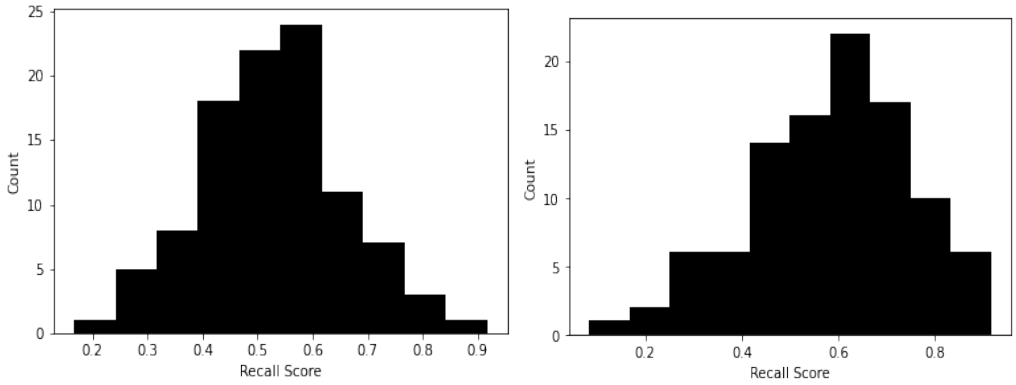


Figure 4-2. MLR (left) SVM linear (right) Baseline Models, Recall Distributions

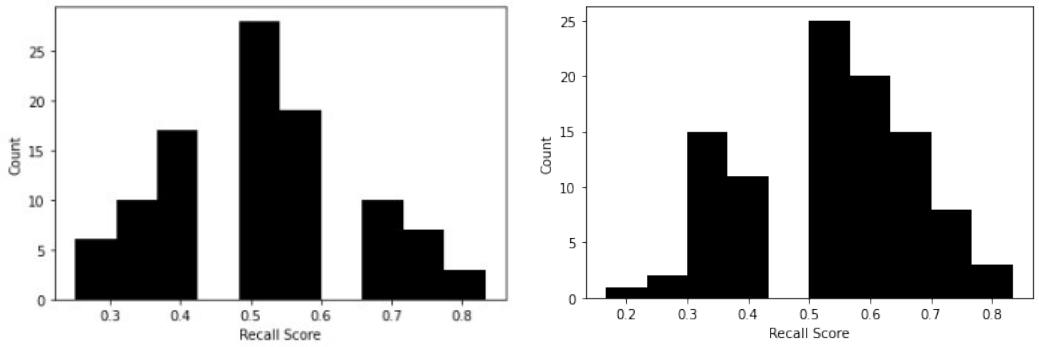


Figure 4-3: MLR (left) and SVM linear (right) Baseline + PCA Emotion Feature, Recall Distributions

4.4 SVM and MLR Emotion Composite Feature on Vitality

The original work that led to the proposed research problem of this study was by Shinohara et al., (2020) and the presented index of vitality to measure emotional load. Vitality itself was unable to be recreated, however the objective of the index was to combine features representing the sum of joy over sorrow and calmness over excitement with a 0.6 and 0.4 multiplier respectively. While it is unclear what features specifically map to each emotion, the harmonic to noise ratio is perceived as the quality of voice in breathlessness, cracking, and vocal fray. As such, in an attempt to purpose a substitute vitality, the HNR feature is used to represent calmness over excitement.

While HNR represent aspects of voice quality the relative spectral centroid with respect to the standard deviation of the fundamental frequency is used for the joy over sorrow element. This feature was chosen as the spectral centroid represents the

central mass of the spectrum or where the most energy would be concentrated. Levels of energy are mapped to happiness in emotion recognition work (Ke et al., 2018).

With these secondary features combining low level aspects to track emotion, the SVM and MLR models are trained with the new vitality and baseline features. The linear SVM and MLR recall distributions are shown in Figure 4-4, and performance metrics are provided in Table 4-2.

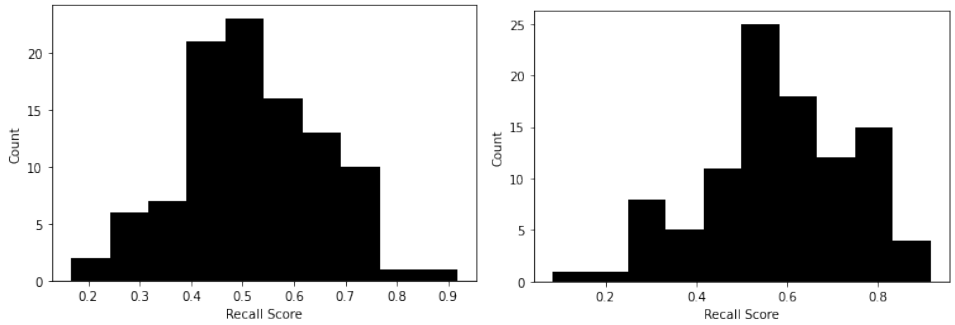


Figure 4-4. MLR (left) and SVM linear (right) Baseline + New Vitality, Recall Distributions

4.5 SVM and MLR Full Acoustic Dataset

While the main objective of this paper is to compare the SVM and MLR models including the proposed emotional features with the baseline model, many features in the Spearmens correlation test showed to be more impactful to MDD detection. Therefore, a lasso regression feature selector was created, which identified the median of MCEP 23 and the mean response time without overlap as the only features necessary for MDD detection. For comparison purposes, a linear SVM model and MLR model were developed, and the recall distributions can be found in Figure 4-5. The linear SVM model and MLR model both achieved their highest model accuracy in this model variation at 62.9% and 60%. This indicates that the emotional composite features might benefit from these two features, and it warrants further research though was not developed within this study.

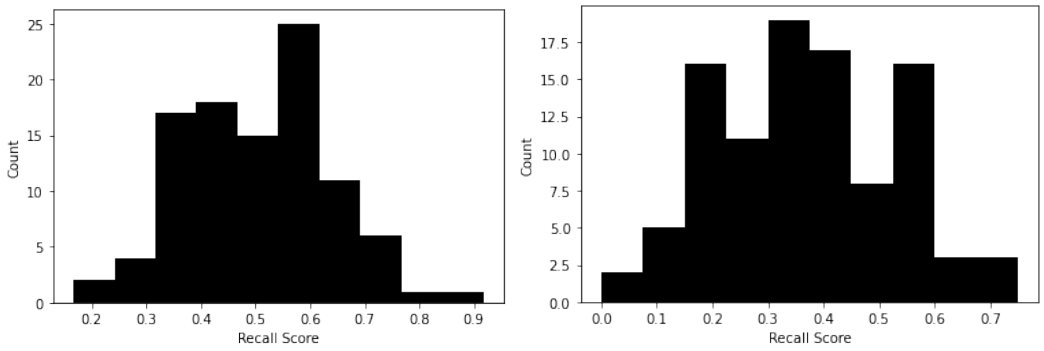


Figure 4-5. MLR (left) and linear SVM (right) Lasso Selection Feature Model, Recall Distributions

4.6 SVM and MLR Historical and all Novel Features

Within this research study, different acoustic features not typically extracted during MDD detection research were evaluated. The features are listed in chapter 3 but entail taking spectral and pitch-based features over windows relative to the windows mean F0. In the spearmen correlation test, 3 of these features made it into the top 20. The recall distributions shown in Figure 4-6 are from SVM linear and MLR models developed using the baseline features plus the top 20 relative features extracted (max F0 relative to F0 standard deviation, max F0 relative to F0 mean, max F0 gradient relative to F0 mean).

Due to the historical data already containing mean pitch-based features (Low et al., 2020) and despite the high correlation the three new features had to MDD detection, these models performed the lowest in accuracy with the linear SVM at 40.8% and the MLR at 42.6%. Recall of the MDD class was also low at 47% and 51.1% respectively.

While the novel features do represent vocal tract movements and are mapped to MDD, further research is warranted on nonlinear relationships between these features and others to increase both recall and accuracy of the model.

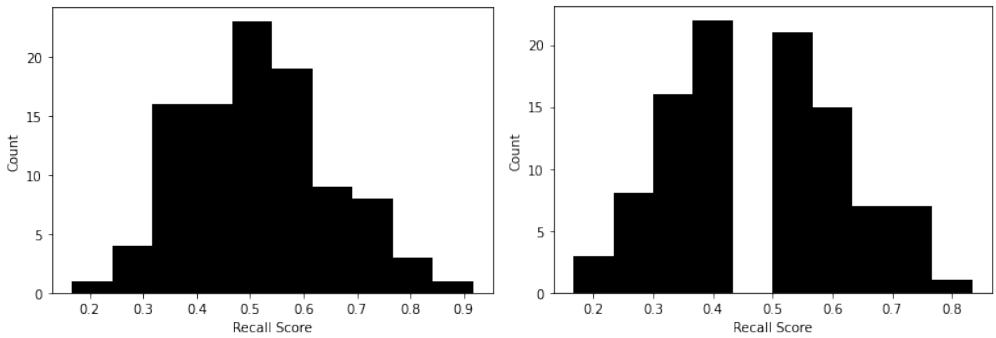


Figure 4-6. MLR (left), linear SVM (right) baseline and Novel Features in Top 20

4.7 Discussion

The experiment models in this study investigated the impact of adding new features to baseline historically used features of jitter, shimmer, F0 variably and mean pause duration. Table 4-2 lists the mean accuracy and recall each of model achieved over 100 iterations. The distribution of the achieved recalls is shown in their respective figures prior in this chapter.

An interesting note about the mean accuracy and recall values are in comparison to past work. The mean MDD recall ranges for 0.369 to 0.545 are in line with Huant et al., (2020), although both the average recall and accuracy was significantly lower than those achieved by Jiang et al., (2017) and Liu et a., (2020). However, in both of these cases, models were divided by gender of the speaker. In this study all speakers are present in the same model. Due to the number of features taken relative to the F0, including half of the proposed new vitality feature, models may benefit from taking gender into account.

In order to test if any of the models had a statistically significant difference in recall compared to the baseline a Wilcoxon test was performed with a p value of 0.05. The test statistic and p value for each model in comparison to the baseline is reported in Table 4-2. No model had a p score < 0.05, though the MLR with the PCA emotion feature did achieve a p value of 0.0568. While this is currently not significant, it does indicate viability for future testing and work. As no model demonstrated a recall distribution with a statistically significance difference this study fails to reject the null hypothesis that adding selected emotion based composite features would have no impact on recall performance of SVM and MLR models compared to models with the baseline features alone.

Table 4-2 Performance Metrics of Experiment Models and Wilcoxon Test Results to Baseline Model

Features	Model	Mean Accuracy	Mean Recall	Wilcoxon test statistic	Wilcoxon test p-value
Baseline	SVM Linear	0.51	0.545	N/A	N/A
	MLR	0.52	0.525	N/A	N/A
Baseline + PCA Emotion	SVM Linear	0.526	0.529	531	p = 0.442
	MLR	0.526	0.514	386.5	p = 0.0568
Lasso Selector	SVM Linear	0.629	0.369	390	p = 1.29
	MLR	0.6	0.501	1683	p = 0.251
Baseline and F0 Relative Features	SVM Linear	0.567	0.47	751	p = 8.084
	MLR	0.56	0.511	1225	p = 0.21
Baseline + new vitality	SVM Linear	0.511	0.541	177	p = 0.77
	MLR	0.513	0.516	193	p = 0.0717

5. CONCLUSION

5.1 Research Overview

Research into applied computing for MDD detection focuses on feature extraction, selection, and model algorithm development. Since MDD detection deals with mental health, there is a strong need for explainable models with high levels of accuracy and MDD class recall for screening purposes. Audio data was demonstrated in Tlachac et al., (2021) as a screening method participants would choose to engage with. The literature review covered impacts such as gender and age into acoustic features, features common in emotion recognition such as vitality (Shinohara et al., 2021), and current research into both deep learning and machine learning algorithms for MDD detection.

5.2 Problem Definition

The research problem investigated in this study stemmed from a gap in research testing identified features with correlation to MDD detection outside of isolation as the only feature in a model. This was the case in Shinohara et al., where the vitality index was tested for correlation with MDD but was not tested in combination with any other features. Building upon this, the research question of this study is: "What is the impact of additional acoustic features measuring emotional load and vowel shape, when combined with traditional features, on the recall of supervised models trained for binary classification of Major Depressive Disorder?"

5.3 Design/Experimentation, Evaluation & Results

The null hypothesis at the core of this experiment was that adding composite features from low level acoustic features associated with emotion recognition would not impact the recall performance of MLR and SVM utilizing the baseline features.

To test this, linear SVM and MLR models were developed using the baseline features of shimmer, jitter, max F0, F0 range, the mean of F0 and the mean pause duration from participant audio files in the DAIC-Woz dataset.

Each model was run 100 times with the same random state inputs and data splits. A new vitality based feature was evaluated using HNR for the calmness and

excitement component and the relative spectral centroid with respect to the standard deviation of the fundamental frequency for the happiness and sorrow components. Another model tested a PCA feature created through PCA selection of common emotion recognition features, including spectral band energy and formant frequencies.

The vitality based feature model achieved an average accuracy of 51.1% and 51.3% for the linear SVM and MLR models. The recall of the MDD class averaged 54.1% and 51.6%, respectfully. The PCA emotional feature SVM model achieved a 52.6% accuracy with an MDD recall of 52.9%. The MLR model achieved 52.6% accuracy with a recall of 51.4%. The recall distribution over the 100 randomized iterations were tested against the respective models using the baseline features with the Wilcoxon test. The SVM models achieve p values of 0.492 (PCA emotion) and 0.77 (vitality based feature), both above the targeted p value for significance of 0.05. The MLR models where closer to 0.05 at 0.056 and 0.717. These p values do not support the rejection of the null hypothesis, so this study cannot reject the null hypothesis.

5.4 Future Work & Recommendations

While this study does not reject the null hypothesis, it was observed that the emotion recognition-based features involved fundamental frequency often. If this research was repeated with gender specific models, different results may have been achieved when considering the work of Jiang et al., (2017) and Bailey & Plumby, (2021).

Additionally, Jiang et al., (2017) reported model improvements when the undertone of questions asked of participants was considered. Rather than sampling an entire interview as this study did, cutting the audio to a specific question known to trigger more emotional response might yield audio better suited for emotion recognition-based features to capture the difference in response between a healthy participant and one with MDD. Additionally gathering data from additional speech task, such as reading a passage, has been reported to increase the acoustic feature differences between MDD and control participants.

Furthermore, additional work utilizing different models to predict extreme cases (PHQ9 scores > 17) versus cases closer to the cut-off point would improve recall. As mentioned in Kanter et al., 2008, depression severity does impact the extent symptoms can be identified, and for traditional machine learning approaches with

limited data for the MDD class can present issues in outlier data. One of the limitations of the study was the limit in of MDD patients across the entire range of MDD.

In summary future work in the research problems relating to emotional features assisting in MDD classifiers should focus from the data source up and consider adapting datasets to what is going to present the best activity to gather the differences MDD symptoms produced. Asking participants to describe a happy memory with a friend for example presents a positive undertone question with the aim of having participants answer in a genuinely happy way. For the DAIC-Woz database this would be achieved by a question such as when Ellie asks the participants to describe their hometown. While not guaranteed to be positive it takes a step in the right direction while using a dataset already available. Due to MDD symptoms including decrease in overall joy and please for things talking about a memory in which those feelings are associated with rather than a psychiatric interview creates a situation where the decrease in energy can be spotted between the two groups.

BIBLIOGRAPHY

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* 5th ed.). Washington, DC

Bailey, A., & Plumbley, M. D. (2021). Gender bias in depression detection using audio features. *2021 29th European Signal Processing Conference (EUSIPCO)*. doi:10.23919/eusipco54536.2021.9615933

Boersma, P., & Weenink, D. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.2.18, retrieved from http://www.praat.org/

Burdisso, S. G., Errecalde, M., & Montes-y-Gómez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications, 133*, 182-197. doi:10.1016/j.eswa.2019.05.023

Cameron, I. M., Crawford, J. R., Lawton, K., & Reid, I. C. (2008). Psychometric comparison of PHQ-9 and Hads for measuring depression severity in primary care. *British Journal of General Practice, 58*(546), 32-36. doi:10.3399/bjgp08x263794

Chen, X., & Pan, Z. (2021). A convenient and lowcost model of depression screening and early warning based on voice data using for public mental health. *International Journal of Environmental Research and Public Health*, 18(12). doi: 10.3390/ijerph18126441

Cohen, A. S., Cox, C. R., Le, T. P., Cowan, T., Masucci, M. D., Strauss, G. P., & Kirkpatrick, B. (2021). Using machine learning of computerized vocal expression to measure blunted vocal affect and alogia. PJ *schizophrenia, 6*. doi: https://doi.org/10.1038/s41537-020-00115-2

Cohn, J., Kruez, T., Matthews, I., Yang, Y., Nguyen, M., Padilla, M., . . . De la Torre, F. (2009, 10). Detecting depression from facial actions and vocal prosody. In (p.1-7). doi: https://doi.org/10.1109/ACII.2009.5349358

Corretge, R. (2022). Praat Vocal Toolkit. [Computer Program] retrieved from: https://www.praatvocaltoolkit.com

Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S. (2014) COVAREP – a collaborative voice analysis repository for speech technologies, *IEEE International Conference on Acoustics, Speech, and Signal Processings (ICASSP)* Florence, Italy.

Di Matteo, D., Fotinos, K., Lokuge, S., Yu, J., Sternat, T., Katzman, M. A., & Rose, J. (2020). The relationship between smartphone-recorded environmental audio and

symptomatology of anxiety and depression: Exploratory study. *JMIR Formative Research, 4*(8). doi:10.2196/18751

Epstein, R., Duberstein, P., Feldman, M., Rochlen, A., Bell, R., Kravitz, R., ... Paterniti, D. (2010, 09). ”i didn’t know what was wrong:” how people with undiagnosed depression recognize, name and explain their distress. *Journal of general internal medicine*, 25, 954-61. doi: 10.1007/s11606-010-1367-0

Feinberg, D., (2018) Measure Pitch, Jitter, Shimmer, and HNR [Computer Program] retrieved from: https://osf.io/huz7d/

France, D., Shiavi, R., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering, 47*(7), 829-837. doi:10.1109/10.846676

Gratch, J., Artstein, R., Lucas, G., Stratour, G., Scherer, S., Nazarian, A., … Morency, T. (2014) The distress analysis interview corpus of human and computer interviews. *Proceedings of Language Resources and Evaluation Conference (LREC)* Retrieved from: https://schererstefan.net/assets/files/papers/508_Paper.pdf

Huang, Z., Epps, J., & Joachim, D. (2020). Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments. In Icassp 2020 - 2020 ieee international conference on acoustics, speech and signal processing (icassp) (p. 6549-6553). doi: 10.1109/ICASSP40776.2020.9054323

Itani, S., & Rossignol, M. (2020). At the crossroads between psychiatry and machine learning: Insights into paradigms and challenges for clinical applicability. *Frontiers in Psychiatry, 11*. doi: 10.3389/fpsyt.2020.552262

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1-15. https://doi.org/10.1016/j.wocn.2018.07.001

Jiang, H., Hu, B., Liu, Z., Yan, L., Wang, T., Liu, F., ... Li, X. (2017). Investigation of different speech types and emotions for detecting depression using different classifiers. *Speech Communication, 90*, 39-46. doi: https://doi.org/10.1016/j.specom.2017.04.001

Kanter, J., Busch, A., Weeks, C., & Landes, S. (2008, 03). The nature of clinical depression: Symptoms, syndromes, and behavior analysis. *The Behavior analyst / MABA*, 31, 1-21. doi: 10.1007/BF03392158

Ke, X., Zhu, Y., Wen, L., & Zhang, W. (2018). Speech emotion recognition based on SVM and ANN. *International Journal of Machine Learning and Computing, 8*(3), 198-202. doi:10.18178/ijmlc.2018.8.3.687

Kelleher, J., Namee, B., & D'Arcy, A. (2020). *Fundamentals of Machine Learning for Predictive Data Analytics* (2nd ed.). Cambridge: MIT Press.

Liu, Z., Wang, D., Zhang, L., & Hu, B. (2020). *A novel decision tree for depression recognition in speech*. arXiv. Retrieved from https://arxiv.org/abs/2002.12759 doi: 10.48550/ARXIV.2002.12759

Lin, R. F., Leung, T., Liu, Y., & Hu, K. (2022). Disclosing critical voice features for discriminating between depression and insomnia—a preliminary study for developing a quantitative method. *Healthcare, 10*(5), 935. doi:10.3390/healthcare10050935

Lortie, C. L., Thibeault, M., Guitton, M. J., & Tremblay, P. (2015). Effects of age on the amplitude, frequency and perceived quality of voice. *AGE, 37*(6). doi:10.1007/s11357-015-9854-1

Louw, S. (2021). Automated transcription software in qualitative research. *Proceedings of the International Conference*.

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1), 96116. doi: https://doi.org/10.1002/lio2.354

McGinnis, E. W., Anderau, S. P., Hruschak, J., Gurchiek, R. D., Lopez-Duran, N. L., Fitzgerald, K., . . . McGinnis, R. S. (2019). Giving voice to vulnerable children: Machine Learning Analysis of speech detects anxiety and depression in early childhood. *IEEE Journal of Biomedical and Health Informatics, 23*(6), 2294-2301. doi:10.1109/jbhi.2019.2913590

Mitsuyoshi, S. (2008). *U.S. Patent No. 7340393*. Washington, DC: U.S. Patent and Trademark Office.

Muzammel, M., Salam, H., Hoffmann, Y., Chetouani, M., & Othmani, A. (2020). Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis. *Machine Learning with Applications*, 2, 100005. doi: https://doi.org/10.1016/j.mlwa.2020.100005

Narziev, N., Goh, H., Toshnazarov, K., Lee, S. A., Chung, K.-M., & Noh, Y. (2020). Stdd: Short-term depression detection with passive sensing. *Sensors, 20*(5). doi: https://doi.org/10.3390/S20051396

Ozkanca, Y., Göksu Öztürk, M., Ekmekci, M., Atkins, D., Demiroglu, C. and Hosseini Ghomi, R., 2019. Depression Screening from Voice Samples of Patients Affected by Parkinson's Disease. *Digital Biomarkers*, 3(2), pp.72-82.

Pan, W., Flint, J., Shenhav, L., Liu, T., Liu, M., Hu, B., & Zhu, T. (2019, 06). Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders. *PLOS ONE*, 14, e0218172. doi: 10.1371/journal.pone.0218172

Schuller, B., Rigoll, G., & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. doi:10.1109/icassp.2004.1326051

Schultebraucks, K., Yadav, V., Shalev, A. Y., Bonanno, G. A., & Galatzer-Levy, I. R. (2022). Deep learning-based classification of posttraumatic stress disorder fand depression following trauma utilizing visual and auditory markers of arousal and mood. *Psychological Medicine*, 52(5), 957–967. doi: 10.1017/S0033291720002718

Shen, Y., Yang, H., & Lin, L. (2022). Automatic depression detection: An emotional audio-textual corpus and a GRU/BILSTM-based model. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi:10.1109/icassp43922.2022.9746569

Shinohara, S., Nakamura, M., Omiya, Y., Higuchi, M., Hagiwara, N., Mitsuyoshi, S., ... Tokuno, S. (2021). Depressive mood assessment method based on emotion level derived from voice: Comparison of voice features of individuals with major depressive disorders and healthy controls. *International Journal of Environmental Research and Public Health*, 18(10). doi: https://doi.org/10.3390/IJERPH18105435

Srimadhur, N. and Lalitha, S., 2020. An End-to-End Model for Detection and Assessment of Depression Levels using Speech. *Procedia Computer Science*, 171, pp.12-21.

Thieme, A., Belgrave, D., & Doherty, G. (2020, aug). Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Trans. Comput.-Hum. Interact.*, 27 (5). doi: https://doi.org/10.1145/3398069

Tlachac, M., Toto, E., Lovering, J., Kayastha, R., Taurich, N., & Rundensteiner, E. (2021). Emu: Early mental health uncovering framework for depression screening. *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)* (p. 1311-1318). doi: 10.1109/ICMLA52953.2021.00213

Tong, L., Liu, Z., Jiang, Z., Zhou, F., Chen, L., Lyu, J., . . . Zhou, H. (2022). Cost-sensitive boosting pruning trees for depression detection on Twitter. *IEEE Transactions on Affective Computing,* 1-1. doi:10.1109/taffc.2022.3145634

Uddin, M. Z., Dysthe, K., Følstad, A., & Brandtzaeg, P. (2022, 01). Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applicatio*n, 34, 1-24. doi: 10.1007/s00521-021-06426-4

Va´zquez-Romero, A., & Gallardo-Antol´ın, A. (2020, 06). Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy*, *22*, 688. doi: https://doi.org/10.3390/E22060688

Xezonaki, D., Paraskevopoulos, G., Potamianos, A., & Narayanan, S. (2020). Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. *Interspeech 2020*. doi:10.21437/interspeech.2020-2819

Yamamoto, M., Takamiya, A., Sawada, K., Yoshimura, M., Kitazawa, M., Liang, K.-c. …, Kishimoto, T. (2020, 09). Using speech recognition technology to investigate the association between timing-related speech features and depression severity. *PLOS ONE*, 15(9), 1-10. doi: https://doi.org/10.1371/JOURNAL.PONE.0238726

APPENDIX A

Table Program Version Table

Software/ Program	Library	Version Number
Python		3.9.7
	Pandas	1.3.5
	Librosa	0.8.1
	Parselmouth	0.4.1
	Scipy	1.7.3
	IPython	7.28.0
	Opensmile	2.4.1
	Sklearn	1.1.2
	Soundfile	0.10.3
Praat		6.2
MATLab		R2009a
	COVAREP	1.0.1

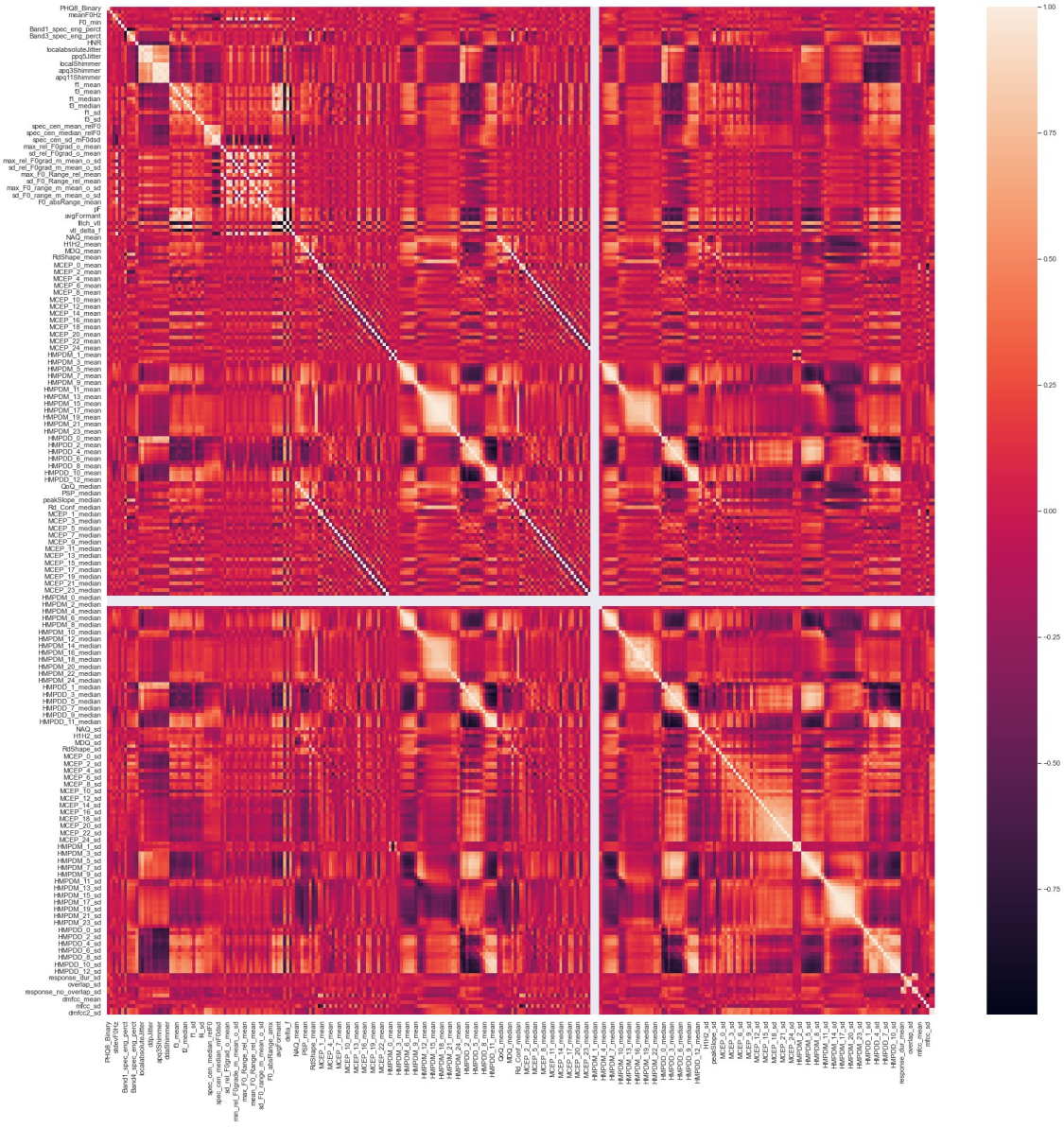


Figure A-1 Heatmap of Entire Acoustic Features Dataset