ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

Depression Recognition from Speech

by
Georgiadou Katerina

Advisor: Ilias Maglogiannis

Submitted

in partial fulfilment of the requirements for the degree of

Master of Big Data and Analytics

at the

UNIVERSITY OF PIRAEUS

June 2022

Thesis

Depression Recognition from Speech

Georgiadou Katerina

June 2022

Acknowledgments

I would like to thank my thesis advisor, Professor Ilias Maglogiannis of the Department of Digital Systems at the University of Piraeus, for his guidance and support.

I would also like to express my gratitude to my family for the patience, support, and encouragement they show at every step I take. Finally, I will be forever thankful to Ilias Mavropoulos, for the unconditional support he showed throughout the entire process of this thesis, and every day before and after.

Abstract

Depression, also known as major depressive disorder, is a major mental health disorder that is affecting ever more lives worldwide. It has a negative impact on the emotional, physical, and psychological state of a person. For a person to be diagnosed with depression, a series of tests must be performed while a series of symptoms must be present for at least 2 continuous weeks. Depression's most common symptoms include feeling down or feeling worthless, lack of interest in daily activities, anxiety, irritability, and reduced appetite. However, depression is possible to cure, and early detection increases exponentially the possibility of controlling the condition.

The complexity of the depression recognition process poses challenges for clinicians regarding both the accuracy of the diagnosis and the timely treatment, given that the disease can be undiagnosed for many months or years, and the fact that delays in the recognition and the treatment can be vital on the life of the patient. To that end, machine learning has been introduced to the medical field to provide tools capable of enhancing the time needed as well as the accuracy and precision of the recognition process, while minimizing human interference.

For this purpose, this thesis studies the use of machine learning models for Depression Recognition using audio data from the widely known database DAIC-WOZ which contains clinical interviews designed specifically to support the diagnosis of psychological distress conditions. Regarding the audio information, the collaborative voice analysis repository (COVAREP) features provided by the dataset were used. Classification is performed using the following models Decision Tree, Random Forest, AdaBoost, Support Vector Machine, and Multilayer Perceptron. AdaBoost achieved the best results and is considered a good model for depression prediction.

Περίληψη

Η κατάθλιψη, γνωστή επίσης ως μείζων καταθλιπτική διαταραχή, είναι μια σημαντική διαταραχή ψυχικής υγείας που επηρεάζει όλο και περισσότερες ζωές παγκοσμίως. Έχει αρνητικό αντίκτυπο στη συναισθηματική, σωματική και ψυχολογική κατάσταση ενός ατόμου. Για να διαγνωστεί ένα άτομο με κατάθλιψη, πρέπει να πραγματοποιηθεί μια σειρά εξετάσεων, ενώ μια σειρά συμπτωμάτων πρέπει να είναι παρούσα για τουλάχιστον 2 συνεχείς εβδομάδες. Τα πιο κοινά συμπτώματα της κατάθλιψης περιλαμβάνουν κακή διάθεση, αίσθηση απελπισίας, έλλειψη ενδιαφέροντος για τις καθημερινές δραστηριότητες, άγχος, ευερεθιστότητα, και μειωμένη όρεξη. Ωστόσο, η κατάθλιψη είναι δυνατόν να θεραπευτεί και η έγκαιρη ανίχνευση αυξάνει εκθετικά τη δυνατότητα ελέγχου της κατάστασης.

Η πολυπλοκότητα της διαδικασίας αναγνώρισης της κατάθλιψης θέτει προκλήσεις για τους κλινικούς ιατρούς όσον αφορά τόσο την ακρίβεια της διάγνωσης όσο και την έγκαιρη θεραπεία, δεδομένου ότι η ασθένεια μπορεί να είναι αδιάγνωστη για πολλούς μήνες ή ακόμη και χρόνια, καθώς και το γεγονός ότι οι καθυστερήσεις στην αναγνώριση και τη θεραπεία μπορεί να είναι ζωτικής σημασίας για τη ζωή του ασθενούς. Για το σκοπό αυτό, η μηχανική μάθηση έχει εισαχθεί στον ιατρικό τομέα για να παρέχει εργαλεία ικανά να βελτιώσουν τον απαιτούμενο χρόνο καθώς και την ακρίβεια της διαδικασίας αναγνώρισης, ελαχιστοποιώντας παράλληλα τις ανθρώπινες παρεμβολές.

Για το σκοπό αυτό, η παρούσα πτυχιακή εργασία μελετά τη χρήση μοντέλων μηχανικής μάθησης για την αναγνώριση της κατάθλιψης χρησιμοποιώντας ηχητικά δεδομένα από την γνωστή βάση δεδομένων DAIC-WOZ, η οποία περιέχει κλινικές συνεντεύξεις που έχουν σχεδιαστεί ειδικά για να υποστηρίξουν τη διάγνωση καταστάσεων ψυχολογικής δυσφορίας. Όσον αφορά τις ηχητικές πληροφορίες, χρησιμοποιήθηκαν οι δυνατότητες του συνεργατικού αποθετηρίου ανάλυσης φωνής (COVAREP) που παρέχονται από το dataset. Η ταξινόμηση πραγματοποιείται χρησιμοποιώντας τα παρακάτω μοντέλα Decision Tree, Random Forest, AdaBoost, Support Vector Machine and Multilayer Perceptron . Το μοντέλο AdaBoost πέτυχε τα καλύτερα αποτελέσματα και θεωρείται ένα καλό μοντέλο για την πρόβλεψη της κατάθλιψης.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Motivation

Depression is a serious mental health disorder that affects the way people think and behave. According to the latest information provided by WHO, it is estimated that 5% of the world's population is affected by depression. Some of the most common symptoms of depression can be the loss of interest in everyday activities, feelings of worthlessness, sadness, sleeping and eating disorders, or even thoughts of suicide, as stated by WHO. In worst cases, depression can lead to suicide. More than 700,000 suicide deaths are reported each year due to depression, while it is one of the leading causes of death among people between the ages of 15 to 29. [WHO-Depression]

Depression recognition is the problem of identifying signs of depression in individuals. Depression can be detected in people's speech, their use of language, or even their facial expressions. Naturally, the process of diagnosing depression is a difficult and complicated task, given that depression can affect each person in different ways. Both the symptoms as well as their severity and duration can vary from person to person. For instance, some people might experience mood changes and get angry and anxious, while others might be unable to make decisions. This process is thus subjective to some extent and the need to develop intelligent systems to help decision-making when it comes to depression is considered of great importance.

Given the growing amount of online available data, the prospects to perform data-driven analyses and developing complex algorithms to assist specialists in psychology, study depression and improve clinical methods and protocols are very promising. Depression diagnosis is a time-consuming process as specialists must consider patients' symptoms and medical history, similar disorders, and possible treatment all at once in order to provide an established diagnosis. AI systems, on the other hand, can process an enormous amount of data in a short time, meaning that provided the therapy sessions, an algorithm will be able to provide faster a recommendation on the diagnosis, that doctors will then evaluate and eventually use. It comes with no saying that the recognition of depression will be more accurate when the models can leverage both the linguistic information as well as the vocal cues and facial expressions, which have a major contribution to the recognition procedure.

Developing systems that are able to understand the behavioral functioning of individuals and predict their depression status, is a challenging task with a lot to offer to the medical society. This thesis focuses on the use of well established machine learning methods and provides them with audio data so as to train the models to recognize depression. The machine learning models used are Support Vector Machine, Decision Tree, Random Forest, AdaptiveBoosting and Multilayer Perceptron.

1.2. Structure

The structure of the thesis consists of five chapters, where each one contributes to a better understanding of the final results. The remainder of this thesis is organized as follows:

Chapter 2 starts by introducing machine learning methodologies, followed by an extensive overview of depression and all the different types of depression that a person can suffer. Additionally, some statistics are listed to emphasize the importance of this type of mental health condition. Following that, information about the Patient Health Questionnaire is provided along with a sample of the questions and the score calculation. After introducing depression, the next major characteristic this thesis is based on, sound is presented. While basic knowledge of sound and speech is described, most of the sound attributes and audio features that contribute to the necessary steps for the recognition process are extensively analyzed and depicted so as to give a better understanding of the fundamentals. Concluding this chapter, some of the existing studies that focus on the recognition of depression from data that derives by any means are presented.

In **Chapter 3** the general machine learning methodology is presented along with some of the most common techniques followed. An overview of machine learning models that fall under the supervised machine learning method is provided along with some of the most common metrics used to evaluate the ML models.

In **Chapter 4** the dataset is introduced along with some statistics that help to understand it better. Subsequently, an extensive analysis of the depression recognition methodology followed in this thesis is presented. More specifically, given that the dataset that will be used in this thesis is created under a specific protocol the most important step in the pre-processing phase includes the removal of the speech segments where Ellie, the virtual interviewer is talking. Finally, the optimal parameters for the following models are presented: Decision Tree, Random Forest, AdaBoost, LinearSVC and MLP.

Chapter 5 includes the results of the depression recognition along with detailed comments on the evaluation metrics of each algorithm that was performed.

Finally, in **Chapter 6** a summary of the findings is presented, as well as suggestions for further investigation on the recognition of depression through the medium of sound.

2. Background and related work

2.1. Machine learning

The subfield that accounts for the vast majority of achievements in the field in recent years is **Machine Learning** (ML). That is also the reason why most people confuse AI with Machine Learning. While AI is described as giving machines human like intelligence, ML is focused on giving them the ability to learn without any human intervention.

The roots of ML can be found during WWII, when Alan Turing worked on cracking the German Enigma machine. The Turing Test, a test proposed by Alan Turing setting the question "Can Machines Think?", states that in order for a machine to be considered intelligent it has to be indistinguishable from human intelligence, as in, the human would not be able to understand if he interacts with another human or a machine. [43] Later, Arthur Samuel, a pioneer of the field, who in 1959 developed one of the world's first self-learning systems, coined the term "machine learning" [62] where he states that

> *"A computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program."*

as well as

> *"Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort."*

Given the availability of large datasets as well as the improvement of computational techniques, Machine learning has now come to adapt in almost every sector imaginable.

Machine learning uses programmed algorithm which analyses the input data in order to learn and give accurate predictions. Depending on the way the algorithm learns, they can be categorized in four types: supervised, semi-supervised, unsupervised and reinforcement

2.1.1. Supervised learning

The most common type of machine learning is supervised training. In supervised learning the algorithm is trained by example, meaning it is given a known dataset with labeled data which includes the inputs and outputs. Even though the data are already known, the algorithm has to identify on its own the patterns, learn from the observations and finally reach to the most accurate prediction. To do so, it is going through a training process which includes a small part of the original dataset, the training dataset. When the training data are fed to the algorithm it adjusts the weights and tries to predict the output data. The weights are then updated as part of the cross-validation process to ensure that the model will avoid overfitting or underfitting. This process is repeated until the algorithm achieves the highest level of accuracy. At the end of the training process the algorithm is tested on the rest of the dataset and as so it continues to learn discovering new patterns from new data.

There are two main algorithms used in supervised learning: Classification and Regression algorithms. Both are used for prediction but the difference between them is in the outputs, classification problems have classes or categories as outputs while regression problems have real numbers as outputs, whether positive or negative.

In a classification model the task is to approximate a mapping function from input variables to *discrete* output variables. The model trains itself using the training dataset and tries to find the best pattern to classify the input data to the specific class labels. To have the best possible accuracy, the training data must include a sufficient amount of data for all given labels, as well as cover all possible scenarios. Once the model has been trained, it is of major importance to evaluate the classification by analyzing its accuracy and efficiency. The evaluation of the classification prediction model can be estimated in many ways, but one of the most common ones is to evaluate the model based on the predicted classes with classification accuracy. Accuracy is the ratio of correctly predicted observations to the total observations of the dataset.

Some of the most common algorithms used in supervised learning are Logistic Regression, Support Vector Machines, Random forests, Naïve Bays, Decision trees and Neural Networks.

On the other hand, regression attempts to determine the correlations between dependent and independent variables. The goal of a regression model is to approximate a mapping function from input variables to a *continuous* output variable. Following the training, some of the most common metrics that are used to evaluate the prediction are:

Mean Squared Error (MSE), which calculates the deviation between the predicted and the actual value.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE), an extension of MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)}$$

Mean Absolute Error (MAE) that calculates the absolute difference between the predicted and actual values. As in RMSE, in MAE the units of the score are also in the same unit as the output variable.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Some of the most common algorithms used in regression are Linear Regression, Polynomial Regression and Support Vector Regression.

2.1.2. Unsupervised learning

Unsupervised learning is a machine learning technique which implements a totally different approach where the machine is made to learn by itself without any human intervention. The model uses unlabeled data to learn by detecting patterns and analyzing the given data in order to obtain the correlations and relationships among them so they can be properly clustered or arranged in a more organized way. Unsupervised learning is a perfect fit when dealing with customer segmentations, exploratory data analysis, and image recognition. One of the most important unsupervised learning methods is clustering.

Clustering is referring to the technique in which unlabeled data are grouped together (clustered) based on their similarities or differences. The most common as well as the most used clustering algorithm is k-means clustering as it is known to return more accurate results in a short period of time compared to other clustering algorithms. In k-means, there can be several variations, but the principal remains the same for all algorithms: *n* observations are grouped into *k* clusters based on their distance. Each data point is assigned in one of the *k* groups. The clustering is based on feature similarity so the distance between each data point to the k cendroids is calculated using Euclidian distance and the data points are later re-assigned to the nearest cluster. The process repeats until there is no re-assignment required.

Unsupervised learning can be utilized also for association and dimensionality reduction.

2.1.3. Semi-supervised learning

As machine learning becomes a necessity and is nowadays of great help in many sectors, the disadvantages of supervised learning are causing a significant problem. The lack of large amounts of labeled data, as well as the valuable time needed to label data, whether that's from human hands or machines, are leading to an increased use of semi-supervised learning.

Semi-supervised learning is a technique which takes advance of the benefits of the two machine learning models mentioned above. During the training process, the algorithm is trained using a combination of a small amount of labeled data (supervised learning) and a larger amount of unlabeled data (unsupervised learning). When given the labeled data the algorithm trains itself and thus creates a partially trained model. The new model will later label all the unlabeled data that were given resulting in a new set of "pseudo-labeled" data. The two datasets are linked and trained resulting in reduced errors and improved accuracy of the model.

Semi-supervised learning can be used both in classification and clustering algorithms.

2.1.4. Reinforcement learning

Inspired by the field of behaviorist psychology, reinforcement learning is a learning model where the algorithm gets trained based on a system of rewards and punishment. While in supervised learning the model is trained on labeled data, in reinforcement learning the model is bound to learn by itself using trial and error to reach optimal results. The algorithm enables an agent that gets trained by exploring different paths to solve a problem using constant feedback for its actions. The agent is given either a reward or a punishment depending on its actions forcing it to continuous learning. This way, even though the model was originally given limited information ends up with superhuman skills achieving the optimal solution. The goal of the agent in this process, is to increase its total cumulative rewards, so as be able to compile sophisticated tactics in any environment, no matter how uncertain or complex.

2.2. Depression

Depression, also known as Major depressive disorder (MDD), is a mental health disorder that negatively affects the everyday life of the individual suffering from it. It is a common but serious disorder that impacts the individual emotionally, physically as well as psychologically. People suffering from depression usually experience feelings of sadness, anxiety, loss of pleasure or interest in normally enjoyable activities and reduced self-esteem. They are prone to sleep disturbances as they might find it difficult to sleep or wake up and have pessimistic views of the future. They may suffer from suicidal thoughts and develop slowed thinking, speaking or body movements [29] [55]. Even thought, those symptoms can also be associated with sadness or grief, depression differs in many ways from the grieving process one goes through when dealing with the death of a loved one or the loss of a job. [55]

While Major Depressive Disorder is one of the most common disorders, there are various kinds of depression. The Diagnostic and Statistical Manual for Mental Disorders (DSM-5-TR) divides depression into two basic categories: Bipolar Disorder and Unipolar Depression.

Although **bipolar disorder** differs from depression, it is included in this list as those who suffer from it can have depressive episodes. Bipolar disorder causes mood and energy shifts from overly elated and elevated mood, also known as manic episode, to depressive lows. Sometimes manic and depressive symptoms may be experienced at the same time, which is called a mixed episode. Given the mood changes, the disorder can be separated in three forms: [20] [76]

1. **Bipolar I disorder**, where at least one manic episode that lasted at least seven days has occurred, or caused the need for hospitalization
2. **Bipolar II disorder**, where a combination of a depression episode and a hypomanic period occurs.
3. **Cyclothymic Disorder** (also called Cyclothymia), where there are periods of hypomanic along with periods of depression that last for at least 2 years.

Unipolar Depression also known as **Major Depressive Disorder** is mainly focused on the “lows”, the negative emotions a person is experiencing. According to the DSM-5 the diagnostic criteria for major depressive disorder require five or more symptoms to be present during the same 2-week period, most of the day, nearly every day while at least one of the symptoms must be depressed mood or loss of interest/pleasure. The symptoms of MDD that are in the diagnostic criteria include: [75] [40] [31]

1. depressed mood (or irritability in children and adolescents)
2. decreased interest or pleasure in daily activities
3. weight loss or gain or a noticeable change in appetite
4. difficulty sleeping or sleeping more than usual
5. noticeable changes in physical movements
6. fatigue or loss of energy
7. feeling worthless or inappropriately guilty
8. difficulty thinking and concentrating
9. thoughts of death or suicide

MDD is also divided into subtypes – called specifiers, depending on the duration of the diagnosis and the defining characteristics of each one. Some of those are mentioned bellow:

1. **Persistent depressive disorder**, also called dysthymia.
   This type refers to a depressed mood that lasts for at least two years however it does not reach the intensity of MDD and thus might also be referred to as "high-functioning" depression. [70] The patient is able to function day to day but finds it difficult to feel happy even on joyful occasions. Some other symptoms are low self-esteem, hopelessness, low energy, poor concentration or difficulty making decisions or sleep changes. [75] [52] [76]
2. **Seasonal affective disorder (SAD)**
   Depression symptoms occur seasonally, and usually during autumn or winter months, when there is less sunlight. For a patient to be diagnosed with SAD, symptoms must be present for 2 consecutive years during a specific season and a full remission (or a shift to mania/hypomania) must happen at a characteristic time of the year. [70] [75] [52] [76]
3. **Postpartum depression**
   Although women are at higher risk for general depression, they are also at risk for postpartum depression, also called perinatal depression, which is influenced by reproductive hormones. Women during pregnancy or in the first 12 months after delivery suffer from major or minor depression episodes. Symptoms include feeling depressed most of the days, feeling distant from family and friends, loss of interests, feeling tired or irritated and having feelings of anxiety or panic attacks. [70] [52].
4. **Psychotic depression**
   This type of depression can be developed when a person has severe depression episodes in addition to some form of psychosis, like delusions or hallucinations. [52].

Depression can affect anyone, at any age, but it often begins at adulthood [52]. In the US, roughly 21.0 million people experienced at least one episode of depression in 2020 Figure 1. [46] It can be triggered by a single factor or a combination of genetics, biological, biochemistry, environmental, and psychological factor. For some people, major life changes or traumatic events, like a bereavement or giving birth, can be the cause, while for others low self-esteem or the exposure to violence, neglect or abuse can make them more vulnerable to depression. Furthermore, depression can run in the family, but it can also be caused by differences in certain chemicals in ones' brain due to illness or even medications. [21] [55]

Even though depression is a disorder that is hard to endure, it is among the most treatable disorders. With an estimation of an 80% to 90% of the people will depression eventually responding to the treatment, even the most severe cases can be treated. Depending on how severe the disorder is, there will be a different approach to treatment. Certainly, the earlier the treatment begins, the more effective it is. Treatment usually involves *psychotherapy*, *medication*, or a combination of the two. Psychotherapy also known as "talk therapy" includes methods such as cognitive behavioral therapy (CBT) and interpersonal therapy (IPT) while antidepressants might be prescribed as to improve the way the brain used certain chemicals to control mood or stress. For the most severe cases, that do not respond to other treatment, *brain stimulation therapies* such as electroconvulsive therapy (ECT) can be reserved. [52] In addition to the above suggested methods, small changes in the lifestyle and the mindset such as exercising and setting realistic goals, can contribute to the treatment process.

Depression affects about 1 in 15 adults and 1 in 6 people will experience depression at some time in their life. [32]

However, it has to be mentioned that there are some studies indicating that the DSM is leading to misdiagnosis. [48]

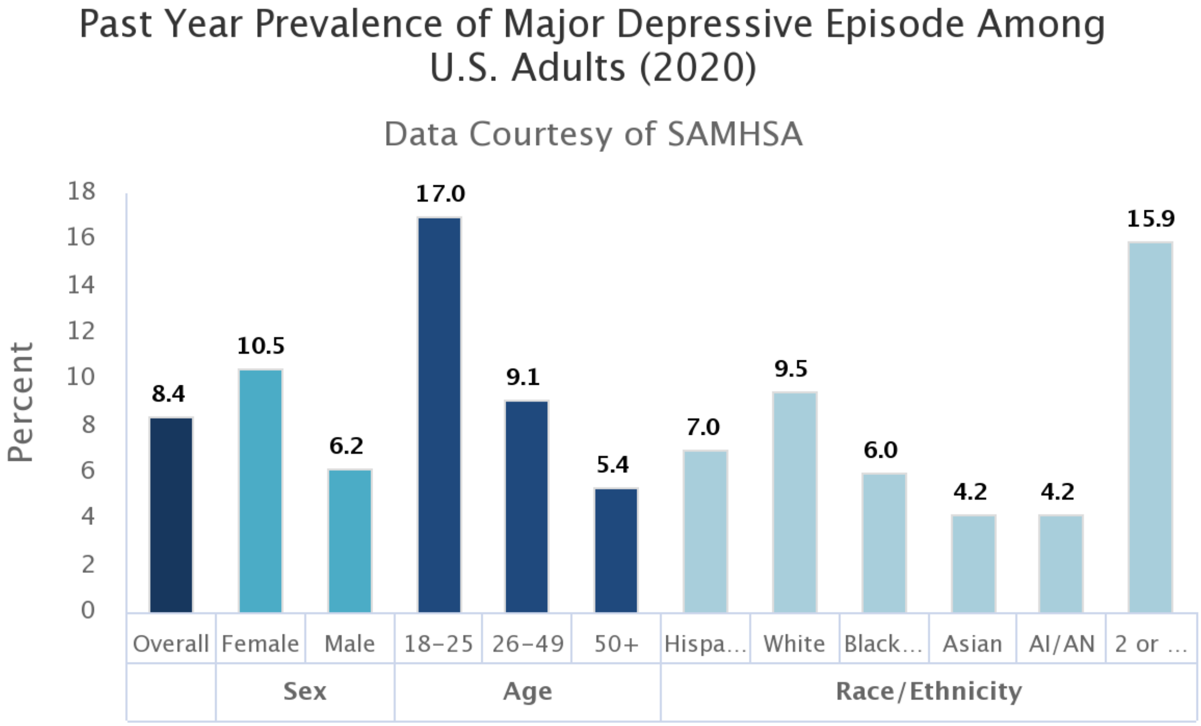


Figure 1 Depression Statistics

**Persons of Hispanic origin may be of any race; all other racial/ethnic groups are non-Hispanic* |

Covid-19 has definitely affected the mental health of people across the globe. A recent study [54] has concluded that during the COVID outbreak, depression prevalence was approximately 7 times higher than in 2017, which is the last global estimated prevalence of depression. However, they highlight the need interpretation with the results. WHO has also published a brief regarding the impact of the COVID-19 pandemic in mental health which comes to the conclusion that “Evidence suggests the pandemic and associated PHSMs have led to a worldwide increase in mental health problems, including widespread depression and anxiety”. [25]

2.2.1. PHQ-8

Given the need to measure the severity of depression, many scales/methods have been developed. The eight-item Patient Health Questionnaire depression scale (PHQ- 8) is established as a valid diagnostic and severity measure for depressive disorders in large clinical studies. Patient Health Questionnaire is a test that determines if somebody is depressed based on a series of questions and it corresponds to the DSM-IV major depressive disorder criteria. In the dataset used in this thesis, instead of the PHQ-9, the PHQ-8 is used. The PHQ-8 is identical to the PHQ-9, but with the suicidal ideation question removed for ethical reasons.

The respondent of PHQ-8 is asked specific question in which he has to reply giving the number of days in the past two weeks he/she has experienced particular depressive symptoms. The responses are separated into three categories, 0 to 1 day=“not at all”, 2 to 6 days=“several days”, 7 to 11 days=“more than half the days”, and 12 to 14 days=“ nearly every day,”, each one assigned with a point (0-3). The questions are the following:

1. Little interest or pleasure in doing things
2. Feeling down, depressed, or hopeless
3. Trouble falling or staying asleep, or sleeping too much
4. Feeling tired or having little energy
5. Poor appetite or overeating
6. Feeling bad about yourself – or that you are a failure or have let yourself or your family down
7. Trouble concentrating on things, such as reading the newspaper or watching television
8. Moving or speaking so slowly that other people could not have noticed. Or the opposite – being fidgety or restless that you have been moving around a lot more than usual

The total score, which can be between 0 and 24 points, is determined by adding together the scores of each item as presented in Figure 2. A total score of 0 to 4 represents no significant depressive symptoms, a total score of 5 to 9 represents mild depressive symptoms, a score of 10 to 14, moderate, 15 to 19 represent moderately severe and 20 to 24 represent severe depression symptoms. [68] However, as mentioned before, there are two ways for a person to be diagnosed with depression, 1. the first or second item must be present for "more than half the days" and at least 5 of the total symptoms must be present "more than half the days", 2. Get a total PHQ-8 score of ≥10.

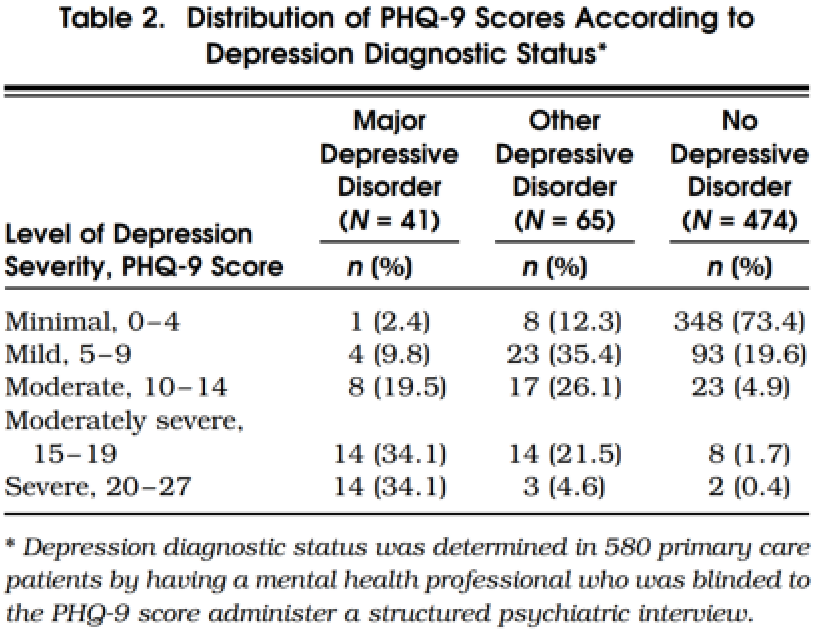


Table 2. Distribution of PHQ-9 Scores According to Depression Diagnostic Status*

Level of Depression Severity, PHQ-9 Score	Major Depressive Disorder (*N* = 41) *n* (%)	Other Depressive Disorder (*N* = 65) *n* (%)	No Depressive Disorder (*N* = 474) *n* (%)
Minimal, 0–4	1 (2.4)	8 (12.3)	348 (73.4)
Mild, 5–9	4 (9.8)	23 (35.4)	93 (19.6)
Moderate, 10–14	8 (19.5)	17 (26.1)	23 (4.9)
Moderately severe, 15–19	14 (34.1)	14 (21.5)	8 (1.7)
Severe, 20–27	14 (34.1)	3 (4.6)	2 (0.4)

** Depression diagnostic status was determined in 580 primary care patients by having a mental health professional who was blinded to the PHQ-9 score administer a structured psychiatric interview.*

Figure 2 Distribution of PHQ-9 Scores

2.3. Sound

Sound is produced by vibrating objects. Sound sources, like tuning forks, the strings on a guitar, the larynx etc. vibrate and those vibrations cause particles in the surrounding medium to oscillate and push on the particles near them. This disturbance changes the state of the mediums pressure in the local region creating a sound wave.

Sound waves can be categorized into three categories: longitudinal waves, mechanical waves, and pressure waves. In a longitudinal wave, the particles are moving in a parallel direction to the wave movement A mechanical wave is a wave that depends on the oscillation of matter, meaning that it transfers energy through a medium to propagate. A pressure wave, or compression wave, has a regular pattern of high- and low-pressure regions. A region of increased pressure on a sound wave is called a compression or condensation while a region of decreased pressure on a sound wave is called a rarefaction or dilation. [79] [65]

So, sound can be defined as a vibration that typically propagates as an audible longitudinal wave of pressure through a transmission medium such as a gas, liquid or solid. Energetically, sound is a mechanical wave with the required medium to travel being air.[53] [19] It can also be considered as a pressure wave since sound waves consist of compressions and rarefactions meaning their regions fluctuate between low and high-pressure patterns.

2.3.1. Speech

A key element on how people produce speech is the vocal tract. The vocal tract is a complex system that includes among others the tongue, the teeth, the nasal cavity and the throat (Figure 3). Depending on how a person is shaping his vocal tract, produces different sounds/ linguistic phonemes. It basically acts as a filter.

Speech generation requires a series of events. Initially, it all starts with a glottal pulse, a noisy high-pitched signal that gets generated by the vocal cords. That signal passes through the vocal tracts which act as a filter on the glottal pulse and thus creating the speech signal. Depending on how the vocal tract is shaped, a different speech signal is generated by the glottal pulse. The intuition is that the glottal pulse procures information about pitch while the vocal tract caries information about the timbre.

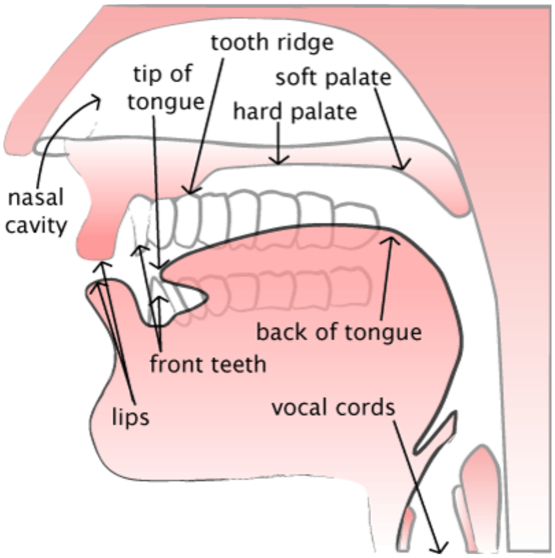


Figure 3 Representation of the vocal tract

2.3.2. Sound attributes

1. Frequency

   Frequency is connected with the idea of period. While a period is the time required to produce a complete cycle, frequency is the number of cycles completed in a particular amount of time. There is an inverse relation between a wave's frequency and its period:

$$f = \frac{1}{T}$$

   The relationship between frequency and sound is that the higher the frequency, the higher the sound that is perceived. Human hearing mechanisms is only capable of perceiving a frequency range of 20 cycles per second up to 20,000 cycles per second. [19]

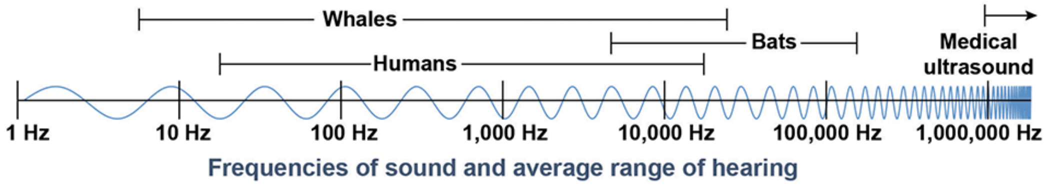


Figure 4 Frequencies of sound and average range of hearing

2. Amplitude

   The amplitude of a sound wave determines its relative loudness and can be observed as the high or low perturbation in air pressure. So, in a sense, amplitude is the distance from rest to crest. Sound is perceived as louder if the amplitude increases, and softer if the amplitude decrease.

3. Pitch

   Perceiving frequency is extremely subjective. However, the way frequency is perceived through pitch differs from the way frequency is perceived itself as pitch has a logarithmic perception of frequency.

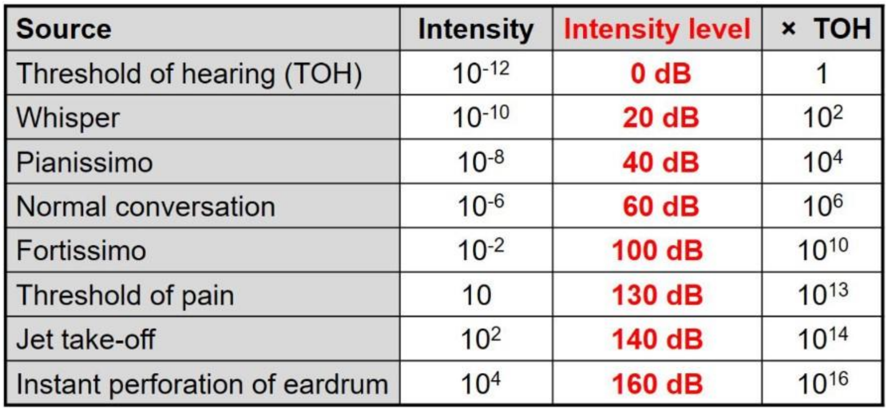
4. Sound Power

   Sound power is the rate at which sound energy is emitted from a source per unit time.

5. Sound Intensity

   Intensity is defined as the average rate of energy transmission per unit area perpendicular to the direction of propagation of the wave. The greater the amplitude of vibrations of the particles of the medium, the greater the rate at which energy is transported through it, and the more intense that the sound wave is.

6. Threshold of Hearing (TOH) – Threshold of Pain (TOP)
   The minimum amplitude of pressure variation that can be sensed by the human ear, also known as the threshold of hearing (TOH) is 0 dB, which corresponds to an intensity of 1*10-12W/m2, while the pressure amplitude at the threshold of pain is 130dB – intensity of 1*101 W/m2.



Source	Intensity	Intensity level	× TOH
Threshold of hearing (TOH)	10^{-12}	**0 dB**	1
Whisper	10^{-10}	**20 dB**	10^{2}
Pianissimo	10^{-8}	**40 dB**	10^{4}
Normal conversation	10^{-6}	**60 dB**	10^{6}
Fortissimo	10^{-2}	**100 dB**	10^{10}
Threshold of pain	10	**130 dB**	10^{13}
Jet take-off	10^{2}	**140 dB**	10^{14}
Instant perforation of eardrum	10^{4}	**160 dB**	10^{16}

Figure 5 Intensity levels

7. Loudness
   Loudness is a subjective perception of sound intensity, and it depends on the duration and frequency of a sound. Each individual perceives differently the loudness of a sound and one of the reasons behind this lies in the age of the person.

8. Timbre

Even tough timbre has been studies by researchers for a long time, they still have to come up with a comprehensive definition. Timbre can be defined as the “color of sound”. It can be considering as the diff between two sounds which have same intensity, frequency and duration. It’s what allows humans to quickly identify sounds (e.g. a piano note, running water, the sound of a friend’s voice).

2.3.3. Audio features

In order to train any Machine Learning model, some of the most useful features must be first extracted from the audio signals. Audio feature extraction is as important as it is necessary given that any ML model needs robust and discriminatory features to learn faster and more accurate.

An audio signal is a representation of sound which encodes all the information needed to reproduce the sound once again - to reconstruct it.

Audio features are descriptors of sound. Different audio features will provide different aspects of sound.

There are a few strategies that can be used for audio feature categorization:

- Level of abstraction
  Mainly covers music signals. Can be divided in 3 levels:
  - High-level: Abstract features that tent to map to musical constructs that are perceived by humans. Features include key, chords, melody, lyrics, genre etc.
  - Mid-level: Features that make sense from a perceptual perspective. Those features include pitch and beat related descriptors, MFCCs etc.
  - Low level: Features that make sense to machines but not so much to humans. Statistical features that get extracted directly from audio such as amplitude envelope, energy, spectral centroid, zero-crossing rate etc. [50]
- Temporal scope
  Applies to any type of sound and can be divided to 3 categories:
  - Instantaneous: As the name suggests there are audio features that give instantaneous information about the audio signal. They are usually very short chunks of signal ~50-100ms. It should be noted that the minimal temporal resolution that people are capable of appreciating is approximately 10ms.
  - Segment-level: Audio features that can be calculated in segments of the audio signal in the range of seconds, from 2-15seconds.
  - Global: Provides features about the whole sound aggregating the results from instantaneous and segment level.
- Music aspect
  This strategy is clearly focused only on music. Related to beat, timbre, pitch, harmony etc.
- Signal domain

Signal domain is one of the most important strategies for categorizing audio features as it consists of the most descriptive features for audio.

- Time domain: Features extracted from waveforms of the raw audio. Some of those features are Amplitude envelope, Root-mean square energy, Zero crossing rate.

  The *amplitude envelope (AE)* of a signal consists of the maximum amplitude value of all samples in each frame. With this feature a rough idea of loudness is given however, it is really sensitive to outliers.

  $$AE_t = max_{k=t\cdot k}^{(t-1)\cdot k-1} s(k)$$

  Root-mean square energy (RMS) is based on all samples in a frame. RMS is also an indicator of loudness but in regard to amplitude envelope, it is less sensitive to outliers, as it gets information from all the samples and not a single sample value from a frame.

  $$RMS_t = \sqrt{\frac{1}{k}\cdot \sum_{k=t\cdot k}^{(t-1)\cdot k-1} s(k)^2}$$

  Zero-crossing rate (ZCR) provides information about the number of times a signal crosses the horizontal time axis. ZCR is extensively used in speech recognition. It can be used in recognition of percussive vs pitched sounds (percussive sounds tend to have random ZCR while pitch tends to be more stable. It can also be used as a monophonic pitch estimator and can distinguish signals which contain voice.

  $$ZCR_t = \frac{1}{2} \sum_{k=t\cdot k}^{(t+1)\cdot k-1} \left|sgn\big(s(k)\big) - sgn(s(k+1))\right|$$

- Frequency domain: Those features focus on the frequency components of the audio signal. The signal of the time domain representation of the raw audio is translated from the time domain to the frequency domain using the Fourier Transform. When Fourier Transform is applied to the time domain representation, it returns a spectrum. Some of the frequency domain features are Band energy ratio, Spectral Centroid, Spectral flux etc.
- Time-Frequency representation: These features combine information about both time and frequency. In order to obtain the time-frequency representation, Short Time Fourier Transform is applied to the time-domain representation of the signal, resulting to a spectrogram. Some of the Time-

Frequency domain features are Spectrogram, Mel-Spectrogram and Constant-Q transform

- ML approach
    - Tradition machine learning considers all possible audio features, both in the time domain and the frequency domain by handpicking the ones that fit better to the model performance. Some of the most used features are Amplitude envelope, Root-mean square energy, Zero crossing rate, Band energy ratio, Spectral centroid, Spectral flux etc. The chosen features are extracted from the audio files and then fed to the traditional ML algorithm

      Band Energy Ratio (BER) provides information about the relation between the energy in the lower and the higher frequency bands. It can be seen as a measure of how dominant low frequencies are.

      $$BER_t = \frac{\sum_{n=1}^{F-1} m_t(n)^2}{\sum_{n=F}^{N} m_t(n)^2}$$

      The *Spectral Centroid (SC)* provides the center of gravity of the magnitude spectrum, meaning that it provides the frequency below band where most of the energy is concentrated. As depicted bellow, spectral centroid is the weighted mean of the frequencies.

      $$SC_t = \frac{\sum_{n=1}^{N} m_t(n)^2}{\sum_{n=1}^{N} m_t(n)}$$

      Bandwidth or Spectral Spread is related to the spectral centroid. It can be perceived as the spectral range that is around the centroid. It is the variance from the spectral centroid and has a direct relation with the perceived timbre. Mathematically, it is the weighted mean of the distances of frequency bands from SC

      $$BW_t = \frac{\sum_{n-1}^{N} |n - SC_t| \cdot m_t\,(n)}{\sum_{n=1}^{N} m_t(n)}$$

    - In Deep learning unstructured audio representations like the spectrogram, MFCCs and Mel-spectrogram are used.

      A *Spectrogram* is a visualization of an audio signal. It includes both the time and frequency aspects of the signal and is obtained by applying the Short-Time Fourier Transform (STFT) on the signal.

The frequency representation of a normal spectrogram is linear and uses Hz, which is problematic as humans perceive frequency logarithmically. Mel scale, a logarithmic scale based on the principle that equal distances on the scale have the same perceptual distance, solves that problem. So, a *Mel-spectrogram* is nothing more than a spectrogram where the frequencies are converted to the mel scale.

$$m = 2595 \cdot \log 1 + \frac{f}{500}$$

The *Mel-Frequency Cepstral Coefficients (MFCC)* are the coefficients that make up the mel-frequency cepstrum. More specifically, cepstrum is a wordplay of spectrum. It is a spectrum of the log of the spectrum of the time signal and since it is neither in the time domain nor in the frequency domain, it was named the quefrency domain. The cepstrum conveys the different values that construct the formants (a characteristic component of the quality of a speech sound) and timbre of a sound. MFCCs are able to describe the large structures of the spectrum while ignoring fine spectral structures and thus are useful for speech and music processing. However, they are not robust to noise.

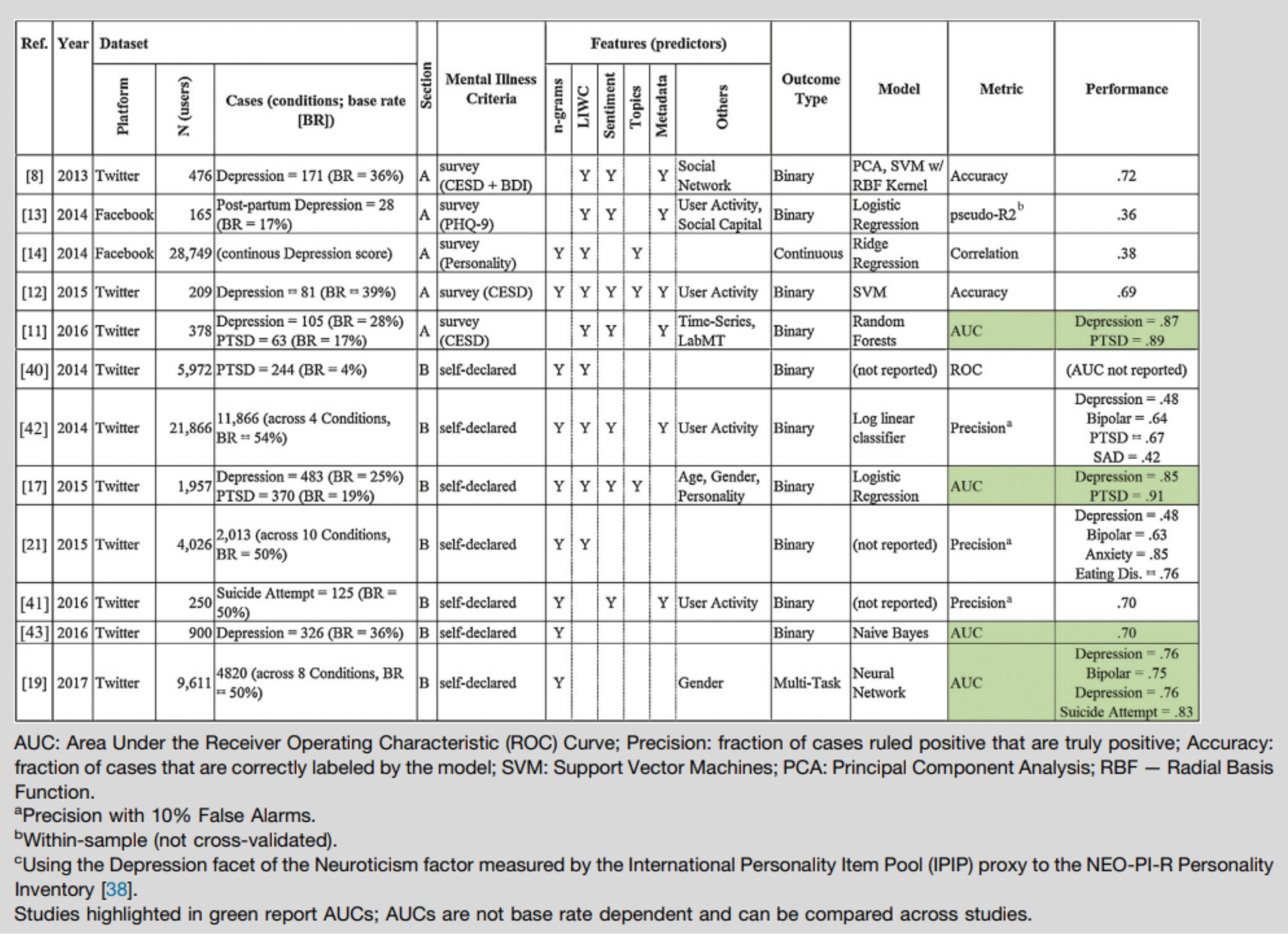
$$C\big(x(t)\big) = F^{-1}[\log F[x(t)]])]$$

2.4. Previous studies

Over the last few years, many research studies in Computer Science have been proposed to deal with mental health disorders. Even more, recently, numerous related papers use spectral and prosodic features extracted from raw audio signals to proceed with the recognition of emotions and mental health conditions.

In a study proposed by [35] various methods were used to identify depression using data from different social media platforms (Twitter and Facebook). This study focused on using text data from two of the largest social networks. The findings of the review are presented in Figure 6, including the type of the models used, the performance of each of the models, and of course the source of the data.

One of the findings of this study is the distinction between diagnosing depression through a survey and being self-declared with depression. It was remarkable that the machine learning models were able to identify depression more accurately when the user was self-identifying the condition, as opposed to being diagnosed via a survey.



Ref.	Year	Dataset			Section	Mental Illness Criteria	Features (predictors)						Outcome Type	Model	Metric	Performance
		Platform	N (users)	Cases (conditions; base rate [BR])			n-grams	LIWC	Sentiment	Topics	Metadata	Others				
[8]	2013	Twitter	476	Depression = 171 (BR = 36%)	A	survey (CESD + BDI)		Y	Y		Y	Social Network	Binary	PCA, SVM w/ RBF Kernel	Accuracy	.72
[13]	2014	Facebook	165	Post-partum Depression = 28 (BR = 17%)	A	survey (PHQ-9)		Y	Y		Y	User Activity, Social Capital	Binary	Logistic Regression	pseudo-R2[b]	.36
[14]	2014	Facebook	28,749	(continous Depression score)	A	survey (Personality)	Y	Y		Y			Continuous	Ridge Regression	Correlation	.38
[12]	2015	Twitter	209	Depression = 81 (BR = 39%)	A	survey (CESD)	Y	Y	Y	Y	Y	User Activity	Binary	SVM	Accuracy	.69
[11]	2016	Twitter	378	Depression = 105 (BR = 28%) PTSD = 63 (BR = 17%)	A	survey (CESD)		Y	Y		Y	Time-Series, LabMT	Binary	Random Forests	AUC	Depression = .87 PTSD = .89
[40]	2014	Twitter	5,972	PTSD = 244 (BR = 4%)	B	self-declared	Y	Y					Binary	(not reported)	ROC	(AUC not reported)
[42]	2014	Twitter	21,866	11,866 (across 4 Conditions, BR = 54%)	B	self-declared	Y	Y	Y		Y	User Activity	Binary	Log linear classifier	Precision[a]	Depression = .48 Bipolar = .64 PTSD = .67 SAD = .42
[17]	2015	Twitter	1,957	Depression = 483 (BR = 25%) PTSD = 370 (BR = 19%)	B	self-declared	Y	Y	Y	Y		Age, Gender, Personality	Binary	Logistic Regression	AUC	Depression = .85 PTSD = .91
[21]	2015	Twitter	4,026	2,013 (across 10 Conditions, BR = 50%)	B	self-declared	Y	Y					Binary	(not reported)	Precision[a]	Depression = .48 Bipolar = .63 Anxiety = .85 Eating Dis. = .76
[41]	2016	Twitter	250	Suicide Attempt = 125 (BR = 50%)	B	self-declared	Y		Y		Y	User Activity	Binary	(not reported)	Precision[a]	.70
[43]	2016	Twitter	900	Depression = 326 (BR = 36%)	B	self-declared	Y						Binary	Naive Bayes	AUC	.70
[19]	2017	Twitter	9,611	4820 (across 8 Conditions, BR = 50%)	B	self-declared	Y					Gender	Multi-Task	Neural Network	AUC	Depression = .76 Bipolar = .75 Depression = .76 Suicide Attempt = .83

AUC: Area Under the Receiver Operating Characteristic (ROC) Curve; Precision: fraction of cases ruled positive that are truly positive; Accuracy: fraction of cases that are correctly labeled by the model; SVM: Support Vector Machines; PCA: Principal Component Analysis; RBF — Radial Basis Function.

[a]Precision with 10% False Alarms.

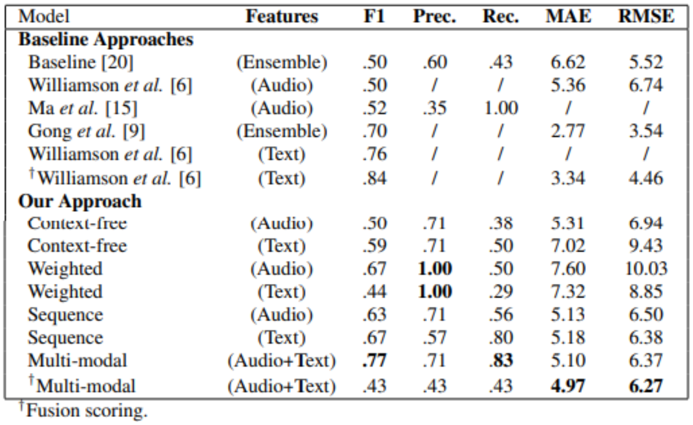
[b]Within-sample (not cross-validated).

[c]Using the Depression facet of the Neuroticism factor measured by the International Personality Item Pool (IPIP) proxy to the NEO-PI-R Personality Inventory [38].

Studies highlighted in green report AUCs; AUCs are not base rate dependent and can be compared across studies.

Figure 6 Representation of first study results

Another study on depression recognition using voice and text data is presented by [38] using data from the DAIC database. It includes three different experiments conducted using text and voice data both independently as well as together. The model that performed better than the rest approaches was the LSTM model using both text and audio data, indicating that not only did a combination of modalities provide additional discriminative power, but that they contained complementary information.



Model	Features	F1	Prec.	Rec.	MAE	RMSE
Baseline Approaches						
Baseline [20]	(Ensemble)	.50	.60	.43	6.62	5.52
Williamson *et al.* [6]	(Audio)	.50	/	/	5.36	6.74
Ma *et al.* [15]	(Audio)	.52	.35	1.00	/	/
Gong *et al.* [9]	(Ensemble)	.70	/	/	2.77	3.54
Williamson *et al.* [6]	(Text)	.76	/	/	/	/
†Williamson *et al.* [6]	(Text)	.84	/	/	3.34	4.46
Our Approach						
Context-free	(Audio)	.50	.71	.38	5.31	6.94
Context-free	(Text)	.59	.71	.50	7.02	9.43
Weighted	(Audio)	.67	**1.00**	.50	7.60	10.03
Weighted	(Text)	.44	**1.00**	.29	7.32	8.85
Sequence	(Audio)	.63	.71	.56	5.13	6.50
Sequence	(Text)	.67	.57	.80	5.18	6.38
Multi-modal	(Audio+Text)	**.77**	.71	**.83**	5.10	6.37
†Multi-modal	(Audio+Text)	.43	.43	.43	**4.97**	**6.27**

†Fusion scoring.

Figure 7 Representation of second study results

In the AVEC depression sub-challenges throughout the years, many papers have presented satisfactory results. [27] conducted content analysis of transcripts to manually select depression related questions. They constructed a decision tree based on the selected questions to predict the patients' depressive conditions achieving F1 s-center of 0.857 for class depressed and 0.964 for not depressed. Despite the overfitting in training, the classification plot still got fulfilling execution on the test dataset. [41] presented a deep learning approach, using CNN on audio and resNet of 50 layers on visual data, in assessing emotion as well as the depression state of a person.

3. Overall Methodology

3.1. Introduction

This thesis aims to recognize whether a person is depressed or not, and for that, a classification task is performed to predict who the depressed person is. The Daic-Woz dataset provides the binary classification of an individual's depression (depressed / not depressed) based on the severity of depression, measured using the Patients Health Questionnaire (PHQ-8) which was introduced in Section 2.2.1.

The predictions are performed under the assumption that there are some similarities in the speech signals extracted from different people suffering from depression. For example, diminished, prosodic and monotonous speech is often strongly correlated with depression [11]. Given that the problem in this thesis is a supervised binary classification, the following algorithms are to be implemented.

1. SVC
2. Logistic Regression
3. Decision tree
4. Random Forest
5. AdaBoost

The Daic-Woz dataset is dealing with imbalanced data, which as mentioned before will probably cause problems given that most algorithms will tend to predict the majority class for every instance. To get the best evaluation of the classification result, it was decided to calculate the following values: Precision, Recall, Accuracy as well as the f1 score to have a better understanding of the precision and recall scores.

A brief overview of the study workflow is presented in Figure 8. First, the audio is preprocessed to obtain a more accurate representation of the participants' speech. Next, audio features are properly extracted using the Covarep toolbox. Finally, implemented in a list of machine learning models is implemented while the results of each models' predictions are compared and analyzed. Overall, the goal within this thesis is to compare Machine Learning method performances across different models' experimentation.

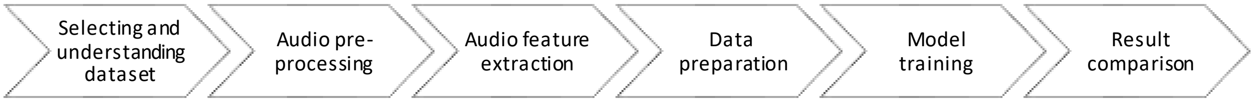


Figure 8 Study workflow

This study is implemented in python using known and effective python open-source libraries including numpy, scipy, matplotlib, sklearn.

3.2. Data Balance

While it is optimal for any classification problem to have balanced data, real word datasets are rarely perfectly balanced. An imbalanced dataset can compromise the model's effectiveness as the algorithm receives considerably more instanced from one class, causing it to be biased towards that one particular class. It is also a very important issue as the minority class is often the one that it is of the most interest in the classification. Some of the most common approaches that handle the imbalanced data, include random resampling, synthetic minority oversampling, and class weight.

- *Random Sampling*: The random resampling method refers to oversampling or undersampling the minority or majority class respectively, Figure 9. Both oversampling and undersampling involve the introduction of a bias that will result in more samples; however, this is considered to be a naïve technique because it assumes nothing of the given data when it is been performed. More specifically, *oversampling* is an approach that expands the dataset through random duplication. It duplicates random samples from the minority class until the instances of both class match. With this method, loosing information is avoided but there is a risk of overfitting the model. The second resampling technique reduces the instances of the major class through random selection by eliminating as many instances needed to match the number of samples of the majority class. Compared with oversampling, one advantage is that undersampling generates a smaller balanced training sample thereby reducing the training time [60]. Yet, the main disadvantage with the undersampling method is that there is a possibility that relevant information might be lost.

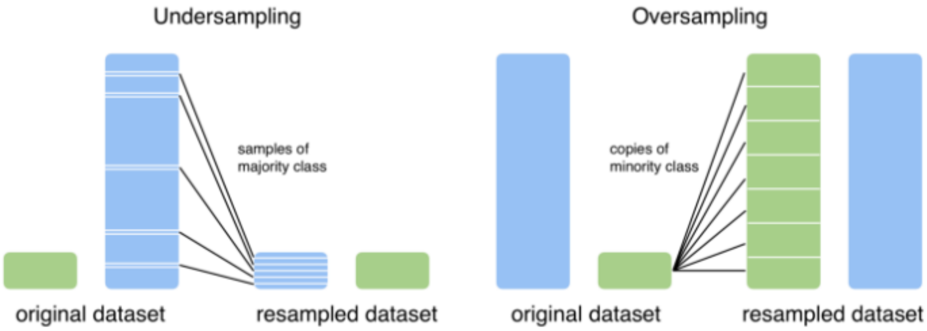


Figure 9 Undersampling - Oversampling

- *SMOTE*: Synthetic Minority Oversampling Technique, commonly known as SMOTE, is an advanced version of oversampling where instead of simply adding duplicates, new instances are synthesized from the existing data. The minority class is oversampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. [23]
- Another common method to address the class imbalance is class weight. Class weighting directly adjusts the cost function of the model by penalize the misclassification of an instance from the minority class greater than the misclassification of an instance from the majority class. By doing so, the class distribution is rebalanced, and the accuracy improved.

3.3. Feature Extraction

Audio features were already extracted and included in the DAIC-Woz dataset. The COVAREP toolbox for speech analysis was utilized to extract specific features through well validated and tested feature extraction methods that aim to capture prosodic characteristics of the speaker as well as voice quality and spectral features. More specifically the extracted prosodic features include the Fundamental frequency (F0) and the voicing boundaries (VUV). The Normalized amplitude quotient (NAQ), quasi open quotient (QOQ), the difference in amplitude of the first two harmonics of the differentiated glottal source spectrum (H1H2), the parabolic spectral parameter (PSP), the maxima dispersion quotient (MDQ), the spectral tilt/slope of wavelet responses (peakslope), as well as the shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics (Rd) are features that correspond to the voice quality. Lastly, Spectral features include 25

Mel cepstral coefficients (MCEP0- 24), Harmonic model and Phase distortion mean (HMPDM0-24) and Harmonic model and Phase distortion deviations (HMPDD0-12)

3.4. Classifiers

In machine learning, classification refers to the process of the categorization of a given set of data into classes. The basic process involves the prediction of the class of each instance in a given dataset. The algorithms implementing classifications are known as classifiers. Regarding modeling classification tasks, various classification algorithms can be deployed. Yet not all classifiers can be successfully applied to all tasks and vice versa. It is generally recommended to conduct experiments and discover which method results in the best performance for each task.

3.4.1. Support Vector Machine (SVM)

Support Vector Machines (SVM) are machine learning models that perform supervised learning for both classification and regression tasks, however, they are mostly used in classification problems.

The main goal of an SVM model is to find a hyperplane in an N-Dimensional hyperplane that distinctively categorizes the data points of each class. Of course, there can be many different hyperplanes that do separate the two classes of a given dataset. The objective is to find a plane that has the maximum margin between the instances of each class, meaning that it must have the maximum distance from the nearest data point of each side. Maximizing the margin offers the probability that future data points can be classified with more certainty in the correct class.

Given a dataset with M number of samples $x_1, x_2 \dots, x_M$ and y_i corresponding labels. The goal is to find a hyperplane $f(x) = 0$ that classifies the given dataset

$$f(x) = w^T x + b$$

where w is a M-dimensional vector and b is a scalar, and they are used to define the hyperplane.

This equation can be either $f(x_i) > 0$ or $f(x_i) < 0$ for x_i belonging to the first and second class respectively. Given this is a 2-dimentional example, the aim is to find the hyperplane for which the minimum distance between the two classes is the widest possible. In order to maximize the margin, the norm $\|w\|$ must be minimized. Ideally, the value $y_i(w^T x_i + b)$ would be ≥ 1 for all samples, indicating a perfect prediction, but, as it is expected, problems are usually not perfectly separable with a hyperplane.

A hyperplane is the optimal decision boundary that helps classify the data points. Given the number of features, the hyperplane can be either a single line or a 2-dimentional plane for 2 and 3 features respectively. It becomes difficult to imagine when the number of features exceeds 3.

Support vectors are the coordinates of individual instances. These are the data point closest to the hyperplane that are the most difficult to classify and thus influence the position and orientation of the hyperplane. Those support vectors are used to maximize the margin of the classifier, any change in these support vectors will lead to a change of the hyperplane.

Marginal distance significance: The marginal distance is the distance created by two parallel lines with respect to the nearest positive and nearest negative point. When separating with respect to the positive and negative, any point above and below the hyperplanes is easily classified to the corresponding class. So, been that in order to get better accuracy using any kind of data, the model needs to be generalized, the margin distance between the instances of each data point needs to be maximum.

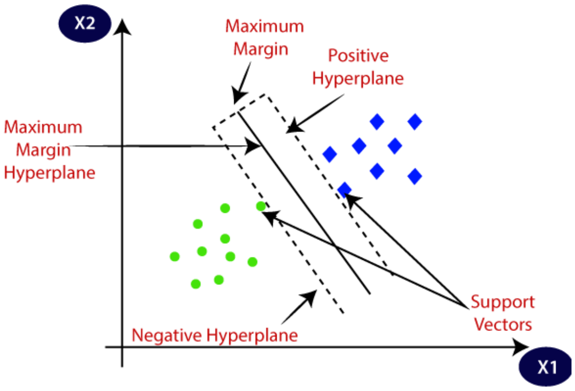


Figure 10 Representation of SVM

SVM can be used for linearly separable data, meaning the dataset can be classified into two classes just by using a single line. Furthermore, it can be also used for non-linearly separated data, meaning that the data can not be classifies with the use of a straight line *Figure 11.*

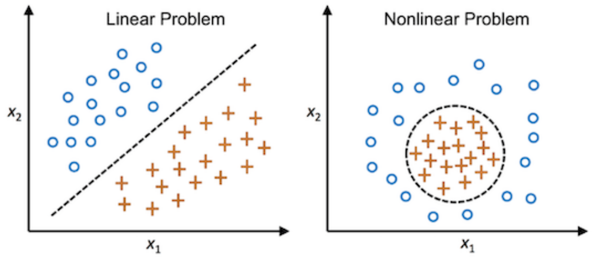


Figure 11 Representation of Linear and Nonlinear Problem

The example presented previously represents a linear problem. In order to solve non-linear classification tasks, kernel functions are applied to the model. By applying kernel functions, the input vectors are mapped to a higher-dimensional feature space, in which linear classification is possible. This mapping is implemented using a kernel function $k(x, y)$, which must satisfy the Mercer's conditions, $k(x,y)$ must equal to $k(y, x)$. Some of the most common kernel functions are:

Linear: $k(x, y) = x^T \cdot y$

Polynomial: $k(x, y) = (x \cdot y + 1)^d$, where d is the polynomial degree

Gaussian RBF: $k(x, y) = \exp(-\gamma \|x - y\|^2)$ for $\gamma > 0$

3.4.2. Decision Tree

Decision Tree is a supervised machine learning technique that can be used for both classification and regression problems, however, it is usually preferred in cases of classification. A decision tree is a hierarchical model composed of discriminant functions, or decision rules, that are applied recursively to partition the feature space of a dataset into pure, single class subspaces [10]. More specifically, their representation resembles somehow a flow chart with a tree structure, wherein instances are classified according to their feature values. In this technique, data are continuously split into smaller parts until all data points are isolated and assigned to each class. The main components of a decision tree model, as illustrated in Figure 12, are one root node, several decision and leaf nodes, and branches.

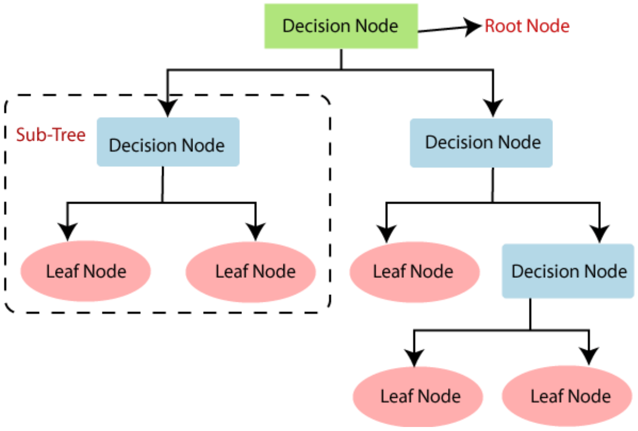


Figure 12 Decision Tree

All in all, there are three types of nodes. The first node, called *root node,* of a decision tree represents the entire dataset that will later be divided into two or mode sets. This is the only type of node that doesn't have any incoming branches. All the other nodes have exactly one incoming edge. The nodes that do not have any outgoing branches are the *leaf nodes,* while all the other nodes are referred to as *decision nodes.*

A decision tree classifies the instances by sorting them down the tree starting from the root and ending at specific leaf nodes, depending on the outcome of the tests occurring in the meantime. Each node in the tree behaves as a test case for an attribute while each branch descending from that node corresponds to one of the possible outputs to that test case. Given that, the process of such a model starts from the root node while each branch that corresponds to the value of the attribute creates a new node. Each node afterward splits the instance space into two or mode sub-trees according to the value of the result of the test. Moving to the sub-tree, created by the last branch, the whole process is then repeated. All instances are classified when there are no more outgoing edges, meaning all instances are associated with a leaf node that represents the class labels.

When building a model, one of the most important tasks is to identify the important input variables that will determine the best way to split the records. This is achieved with the help of characteristics that are related to the degree of purity, such as information gain, Gini index, Gain ration, entropy and Chi-square. [28] [26]

3.4.3. Random Forest

As the name implies, Random Forest is a machine learning algorithm that is based on the concept of ensemble learning, meaning that it combines multiple classifiers, in this case decision trees, in order to solve a complex problem and to improve the model's performance.

There are two types of ensemble method:

Bagging, also known as bootstrap aggregation, which chooses a random set by selecting with replacement from the training dataset and after independently training each model, leads to the final output which is the majority of those predictions.

Boosting which combines weak learners into strong learners by generating sequential models so that the final model has the highest accuracy.

Random forest is a bagging-type ensemble of decision trees that creates an uncorrelated forest of decision trees, trains them in parallel and establishes the outcome of the model based on the majority decision of the trees. This process is also the reason why random forest eradicates one of the most usual issues of decision tree models, which is the problem of overfitting that often occurs.

3.4.4. Adaboost

AdaBoost, short for **Ada**ptive **Boost**ing, is an ensemble method technique in machine learning suited for imbalanced datasets. As mentioned before, there are two basic types of ensemble learning. The one that is mainly used to decrease bias (*Bagging*) and a second one that is used for variance decrease (B*oosting*). Random Forest is a bagging technique that generates models in parallel, exploiting the independence between models by averaging out the mistakes. AdaBoost on the other hand is a boosting algorithm that works on the principle of sequentially growing learners. It basically combines multiple weak classifiers into a single strong classifier

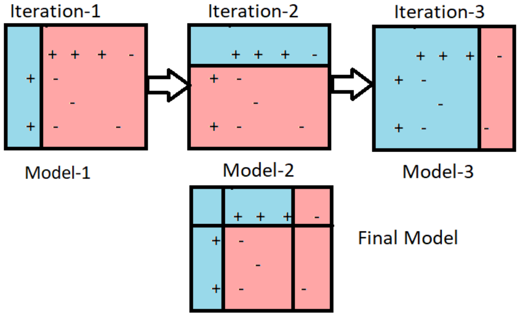


Figure 13 Adaboost iterations

Boosting is an ensemble modeling technique that attempts to build a strong classifier from the mistakes of several weak ones. Initially, the algorithm builds a model on the training data and assigns equal weights to each data point. After the first model is created, the wrongly classified data points are identified, and higher weights are assigned to them so that the subsequent model will focus on correctly classifying the previously misclassified data. Based on the accuracy, weights are also assigned to the classifier after every training. More accurate classifiers will be assigned higher weights in order to have more impact on the final outcome. The previous steps are then repeated until all data points are correctly classified.

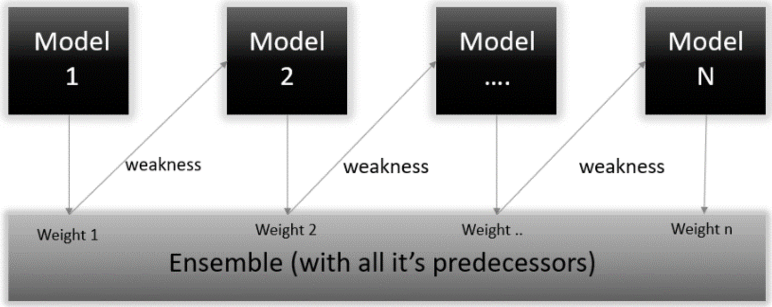


Figure 14 Ensemble modeling technique

The most common algorithms that is used with AdaBoost is one node Decision Trees, meaning they only have one split, and are called Decision Stumps.

3.4.5. MLP

A specific type of Neural Network that is very common and was one of the first neural networks that were engineered is called the Multilayer Perceptron (MLP). MLP is a feed-forward artificial neural network, meaning that the data flows in the forward direction from input to output and does not revisit nodes that have been encountered before.

MLPs are composed of neurons called *perceptions*. A perceptron, as shown in Figure 15, receives n features as input and each of these features is associated with a weight. Those input features are fed to an input function h that computes its weighted sum.

$$h = \sum_{i=1}^{n} w_i \cdot x_i$$

The result of this sum is then passed on to an activation function f, that will generate the output of the perceptron.

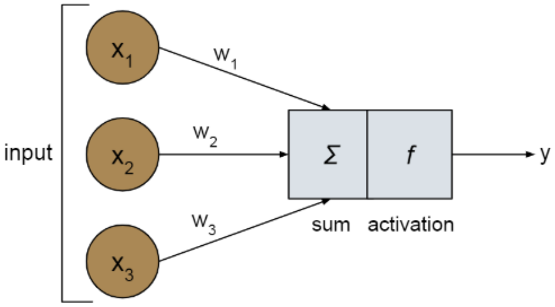


Figure 15 Representation of a perceptron

MLP consists of one input layer, one or more hidden layers and one output layer. Figure 66 represents an MLP with 4 inputs and thus 4 input nodes and one hidden layer with 3 nodes. The output layer gives one output, so naturally, there is one node.

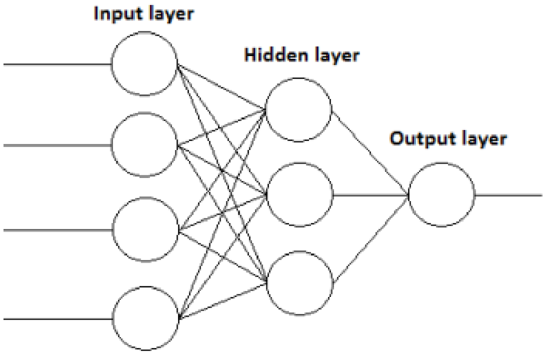


Figure 16 Representation of a Multilayer Perceptron

Each of these units forms a weighted sum of its inputs which passes through a non-linear function, called activation function. Activation functions are responsible for calculating the sum of the various weights to determine the final output value for the current hidden layer, which will be the input for the next layer. Some of the most common activation functions are the logistic function, also known as the **sigmoid** function, the hyperbolic tangent activation function (**tanh**), and **ReLU** or rectified linear unit activation function. The representation of those activation functions is presented in Figure 67

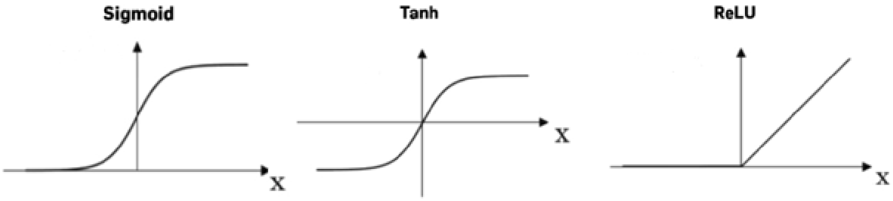


Figure 17 Activation functions

- The *Sigmoid* activation function, one of the most widely used activation functions, transforms any real value to the range of 0 to 1. The mathematical definition of the sigmoid function alongside with its derivative are shown in Equation 1 and Equation 2 accordingly, while their representation is presented in Figure 18.

$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad \sigma'(x) = \frac{\sigma(x)}{(1-\sigma(x))}$$

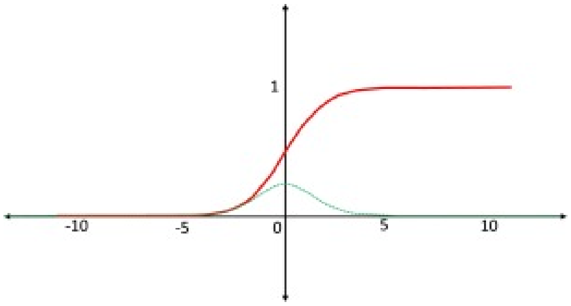


Figure 18 Sigmoid activation function

Sigmoid is continuously differentiable monotonic function, most used in prediction and probability models. The gradient is steep around zero and flattens as it moves farther away on either side. However, converting large number in the range 0 to 1 can rises the problem of vanishing gradients which adds to the fact that the output of the function is not symmetrical around zero leading to a more difficult and unstable training of the neural network. This issue can be improved by scaling the sigmoid function. [5] [8]

- The *Tanh* activation function is very similar to the sigmoid function with the difference that Tanh is symmetrical around the origin. It is continuous and differentiable with values ranging from -1 to 1. Even though the derivative of Tanh is steeper in comparison to the sigmoid function, the output values are still bounded so the gradient still tends to vanish. [5] [8]. The definition of the Tanh function and its derivative are expressed as

$$\tanh(x) = \frac{(e^{x}-e^{-x})}{(e^{x}+e^{-x})} \qquad \tanh'(x) = 1 - tanh^{2}(x)$$

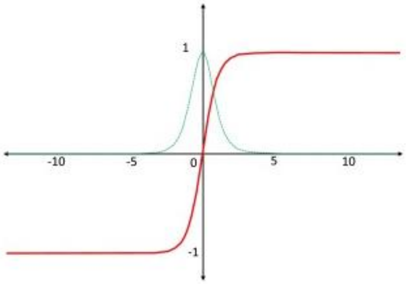


Figure 19 Tanh Activation Function

- To address the issue of vanishing gradients *ReLU* activation function was proposed and is nowadays perhaps the most used one.
  The ReLU activation function is differentiable at all points except point 0. For values greater than 0, the max of the function is considered. This can be interpreted as follows.

$$ReLU(x) = max\{0, x\}$$

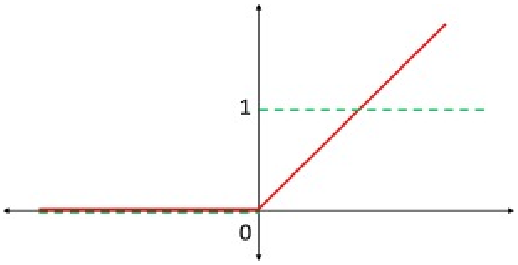


Figure 20 Relu activation function

The ReLU function retains only positive values and discards all negative values by setting the corresponding activations to 0. Shifting all negative values to zero folds the feature space and in result it enforces nonlinearity. Moreover, when working with a large network, deactivating a number of neurons brings sparsity in the system resulting in a more efficient and less computationally expensive system.

The downside of ReLU can be found in the negative regions of the activations, where the gradient is 0 leading to the weights not being updated during training. This creates a problem called the dying ReLU problem, where the neurons affected stop responding. [51] [2]

The complexity of such a model means that the weight-update procedure gets more complicated as well. MLP used Backpropagation as the learning mechanism to iteratively update all the weights in the network, with the condition of minimizing the output error using the method of gradient descent. The gradient of the loss function is computed for each pair with respect to each weight individually, using the chain rule. However, for this to hold, the activation functions must be continuous and differentiable.

3.5. Evaluation Measures

The effectiveness of a model's predictions is evaluated using evaluation metrics. Using different metrics to evaluate a model's performance can improve its overall predictive power.

Confusion Matrix

One of the most common evaluation metrics is confusion matrix. As the name suggests, the values are presented in the form of a matrix, where the Y-axis depicts the actual classes, and the X-axis depicts the predicted classes.

Table 1 Confusion matrix

		Predicted Values	
		Negative	**Positive**
Actual values	**Negative**	*TN*	*FP*
	Positive	*FN*	*TP*

The predicted values are labelled as "positive" and "negative" depending on the actual value that was predicted by the model. The output "TP" stands for true positive and indicates the number of correctly identified positive examples. "TN" stands for true negative; it displays the number of correctly identified negative cases. "FP" is for false positive, depicting the number of actual negative examples classified as positive, and "FN" stands for false negative, which is the number of actual positive examples classified as negative. In this thesis, the negative would present the label for the non-depressed participants while the label for the actual depressed participants would be positive.

After obtaining the confusion matrix of the model, there are also some other rates that can be computed.

Accuracy

Accuracy is the ratio of correct predictions to the total predictions made. It is a metric that can possibly hide details of the model's performance and thus is usually not a good indicator of its efficiency. Mathematically, it is the sum of True Positives (TP) and True Negatives (TN) divides by the sum of the total – True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Accuracy, as well, is insufficient when dealing with imbalanced classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

Precision measures the level of error caused by False Positives (FP). It indicates how many of the prediction of the positive class instances where correct. This is another metric that is unreliable to use when measuring the performance of imbalanced data

$$Precision = \frac{TP}{TP + FP}$$

Recall

Recall, also referred as Sensitivity, is the opposite of precision as it presents the extent of error caused by False Negatives (FNs). It measures the instances out of all the positive instances that were indeed classified as positive.

$$Recall = \frac{TP}{TP + FN}$$

F1 Score

F1 score can be interpreted as the harmonic mean of precision and recall. It reaches its best value at 1 – perfect precision and recall, and worst score at 0. The purpose of F1 score is to obtain an equal balance between precision and recall, which is extremely useful in cases when performing with imbalanced datasets.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

The reason F1 score uses hormonic mean is because it punishes extreme values more. Meaning that harmonic mean discourages hugely unequal values as well as extremely low values.

Macro F1-score

The macro-averaged F1 score (macro-F1), is an arithmetic mean of the per-class F1-scores. The macro-average will first compute the metric independently for each class and then take the average, hence it treats all classes equally regardless of their support values.

4. Depression Recognition Methodology

4.1. Dataset

The Daic-Woz depression dataset, which stands for Distress Analysis Corpus – Wizard of Oz, is part of a larger corpus, the Distress Analysis Interview Corpus (DAIC) [64]. It contains clinical interviews designed specifically to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. These interviews were collected as part of a larger effort to create a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illness [61]. In this subset of the dataset, a virtual interviewer named Ellie, controlled by a human interviewer in another room, is conducting the interviews with the participants, hence the name Wizard of Oz.

The participants of those interviews were recruited through online ads posted on Craigslist.org while the interviews took place at the USC Institute for Creative Technologies (ICT) in Los Angeles, California. All participants were fluent English speakers, and the interviews were conducted in English ranging from approximately 5 to 20 minutes. Each participant has completed the Patient Health Questionnaire alone on a computer and then went on with the interview. The questions asked started off as simple questions to make the participant as comfortable as possible and then progressed to more specific ones, focusing on symptoms and events related to depression and PTSD, while they concluded with more simpler questions once again to ensure that the participant would not leave in a distressed state of mind. [64]

The dataset contains the audio of 189 interviews, their associated transcripts and facial data, as well as the PHQ-8 score based on answers to the PHQ-8 questionnaire answered by each participant. More specifically, the audio file of the interviews consists of raw .wav files and can be represented as shown in figure 21.

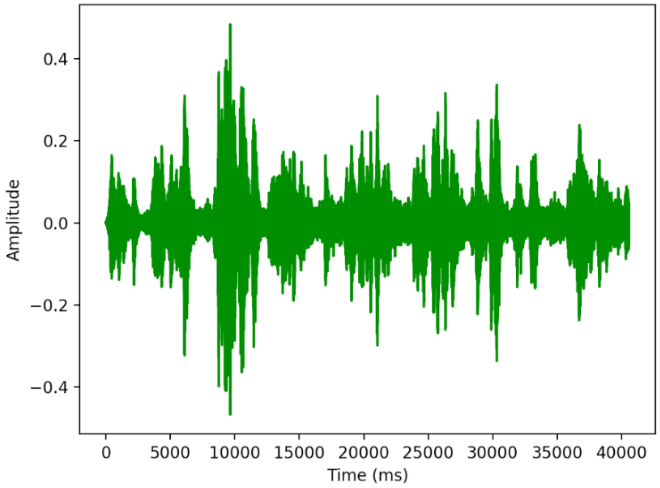


Figure 21 Soundwave of one data sample

* Participant were recorded by a camera, high quality close-talking microphone, and Kinect. Audio has been recorded with the use of head-mounted microphone (Sennheiser HSP 4---EW---3) at 16kHz.

The transcript file contains the sentences spoken by the participant and the virtual assistant Ellie. Each row consists of the start time -the time at which the speaker starts speaking, the stop time - the time at which the speaker stops speaking, the speaker, denoting whether the speaker is Ellie or the participant, and the value - the exact sentence spoken by the speaker. An example is represented in Figure 22.

```
165.854  166.324      Ellie  yeah3 (yeah)
```

Figure 22 Sample of time representation in transcript

The COVAREP and Formant feature files of each interview. Those features were extracted using the COVAREP framework. The COVAREP and formant feature files contains various features from both the participant and the virtual interviewers' voice.

The dataset also has provided the training, development, and test split files. On the training and development split files the following are presented: participant ID, PHQ-8 Binary label, PHQ-8 score, Participant gender, and the responses to every question of the PHQ-8 questionnaire. The test split file contains the participants ID and the participants' gender.

Table 2 Train Development Test set Attributes

	Train	Development	Test
Participant_ID	✓	✓	✓
Gender	✓	✓	✓
PHQ8_Binary lebels	✓	✓	
PHQ8_Score	✓	✓	
PHQ8_NoInterest	✓	✓	
PHQ8_Depressed	✓	✓	
PHQ8_Sleep	✓	✓	
PHQ8_Tired	✓	✓	
PHQ8_Appetite	✓	✓	
PHQ8_Failure	✓	✓	
PHQ8_Concentrating	✓	✓	
PHQ8_Moving	✓	✓	

As mentioned before, the dataset contains the audio of 189 interviews. The train, test and dev sets are given, and each set contains 107, 35 and 47 audios respectively, the percentage of each set can be seen in Figure 23.

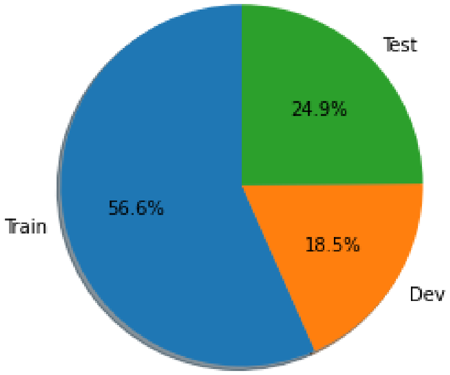


Figure 23 Dataset split representation

The distribution of depressed and non-depressed samples in the train and the development subset are presented bellow in Figure 24:

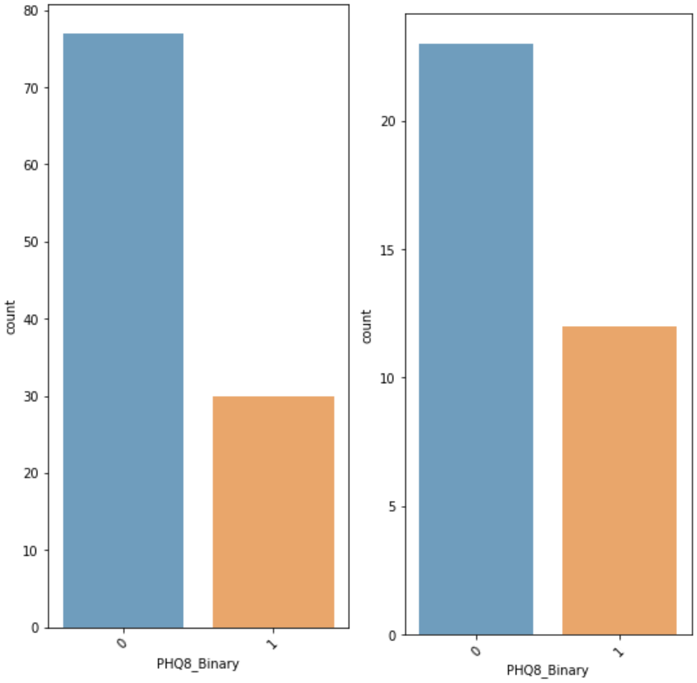
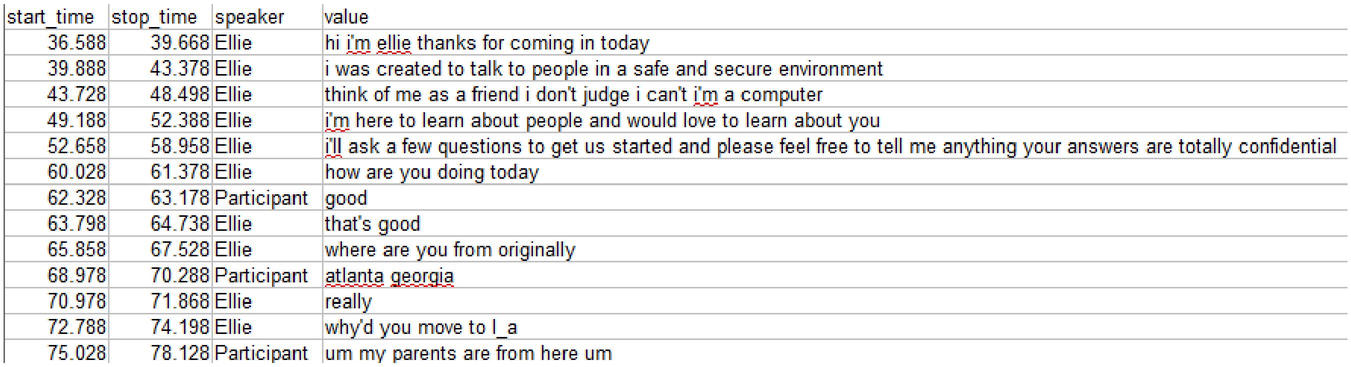


Figure 24 Dataset distribution

4.2. Preprocess

The quality of data is always one major factor that impacts the outcome of almost any ml model. In the Daic-Woz dataset in order to provide the best possible data a very specific protocol was followed, resulting to a clean dataset in regard to noise and disruptions. Of course, no data can be perfect and having in mind that the interviews took place with a virtual interviewer it was expected that there might be some technical issues. The documentation of the dataset provides specific information about interviews with interruptions. Those interviews, 7 in total, were removed.

The original audio recordings consist of the speech segments of both the interviewer, Ellie, and the participants. Since the interviewers' voice does not relate to the severity of the participants' depression, her speech segments would only add noise to the audios. In order to resolve that issue, the transcript provided by the dataset is used. The dataset includes the transcripts of all audio recording in comma-separated (csv) files, that distinguish the speech segments of each speaker with a time frame as presented in Figure 25. Therefore, based on the time frames given in the transcript the speech segments of the participant are easily obtained.



start_time	stop_time	speaker	value
36.588	39.668	Ellie	hi i'm ellie thanks for coming in today
39.888	43.378	Ellie	i was created to talk to people in a safe and secure environment
43.728	48.498	Ellie	think of me as a friend i don't judge i can't i'm a computer
49.188	52.388	Ellie	i'm here to learn about people and would love to learn about you
52.658	58.958	Ellie	i'll ask a few questions to get us started and please feel free to tell me anything your answers are totally confidential
60.028	61.378	Ellie	how are you doing today
62.328	63.178	Participant	good
63.798	64.738	Ellie	that's good
65.858	67.528	Ellie	where are you from originally
68.978	70.288	Participant	atlanta georgia
70.978	71.868	Ellie	really
72.788	74.198	Ellie	why'd you move to l_a
75.028	78.128	Participant	um my parents are from here um

Figure 25 Transcript sample

4.3. Model Training

Following the documentation of the dataset, the train-test-evaluation method was followed with respect to the model training and evaluation. This is a method that can evaluate the performance of a machine learning algorithm and can be used in both classification and regression tasks. For this thesis, the dataset is split into 57%-18%-25% for training-evaluation-test datasets respectively. The split was performed and given from the dataset.

This technique separates the dataset into three subgroups, train, validation, and test. The training set is the first subset and is the largest section of the dataset reserved for the training of the model. A smaller section is separated from the whole dataset and is used during the training phase to evaluate how well the model is training. The remaining section of the dataset is the test set. After the training experiments have concluded the model predicts values using the train set and compares them with the knowledge provided by the test set.

However the recommendation, the final decision was to concatenate the train and validation sets. So, the final train-test set was 57% - 43%.

4.4. Parameter tuning

In the section below, the optimal parameters, for the models mentioned before, are represented. For some of the model, the parameters were determined with the use of grid search, however, due to the fact that it was computational expensive, most of the parameters were set out of random (small experimentation or general knowledge).

Decision tree

As mentioned in section 3.3 most machine learning algorithms are very sensitive to biased class data. One solution to this kind of problem is the modification of the training algorithm in order to consider the skew distribution of classes. This can be achieved by assigning different weights to the minority and majority class. In training, the weight difference will affect the classification, since the algorithm will provide higher or lower punishment to the respected class, giving the algorithm the attention needed to reduce the error of the minority class. Decision tree model, as most sklearn classifier modeling libraries, has a built-in parameter called **class_weight** that can be set to balance or given a custom dictionary to declare how to rank the importance of imbalanced data. It was decided to set class_weight as “balance” to automatically adjust the weights.

Random Forest

With respect to parameter tuning and in order to improve the classification of Random Forest, a few of the model’s parameters have been modified. More specifically, as in the decision tree model, the parameter **class_weight** is again set to “balance” in order to resolve the imbalanced class problem that the dataset is facing.

AdaBoost

One of the most important parameters of the Adaboost is **n_estimators**, as in many cases, by changing the number of weak learners the accuracy of the model is also adjusted. A higher number of trees means more weak learners which leads to better performance. However, the right value must be chosen as the higher the number the slower the algorithm will get. After experimentation with number of estimators set to 10, 20, 30, 50, 70 and 100, the results presented in table 3 were achieved. Based on those results, the **n_estimators** parameter, is set to 50. As for the base estimator of the Adaboost the default DecisionTreeClassifier was used.

Table 3 n-estimator experimentation results

N_estimators	Test set accuracy	F1 score	Macro F1
10	0.6	0.6	0.49
20	0.6	0.6	0.46
30	0.6	0.1	0.42
50	0.71	0.71	0.62
70	0.73	0.45	0.63
100	0.62	0.19	0.47

Linear SVC

Regarding the Support Vector Machine model, it was decided that instead of using SVC with kernel = ‘linear’, the best approach would be to use LinearSVC (Linear Support Vector Classification). Even though the two are analogous, their distinction lies in the fact that the underlying estimators in LinearSVC are liblinear, while SVC uses libsvm. That is also the reason that LinearSVC is faster and scales a lot better than SVC with kernel set as linear.

LinearSVC is one of the models that has the **class_weight** parameter built-in, so in this model the parameter is set to “balance” as well. Another parameter that is modified is **tol** which is set to 1e-5.

MLP

For the Multi-layer Perceptron model, the parameters that were modified were the following. The **hidden_layer_size** was set to (8,4,2) meaning that the model consists of 3 hidden layers with 8, 4 and 2 units respectively. The strength of the L2 regularization represented by the parameter **alpha** is set to 0.00001 and the **solver** parameter is set to “lbfgs”.

5. Depression Recognition Binary Classification

In the present thesis, as mentioned in the previous chapters, five machine learning models were developed. These models were chosen with the aim to produce enough evaluation measures to compare and observe how the DAIC-WOZ dataset performs in each of them. The results of those models are presented in the sections bellow.

5.1. Parameter tuning summary

After the parameter tuning that was described in the previous chapter, a summary of the parameters chosen for each algorithm are presented in Table 34

Table 4 Parameters summary

Models	**Parameter tuning**
Decision Tree	random_state=0, class_weight='balanced'
Random Forest	max_depth=4, random_state=0,class_weight='balanced'
AdaBoost	n_estimators=50, random_state=0
Linear SVC	random_state=0, tol=1e-5, class_weight='balanced'
MLP	solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(8,4,2), random_state=1

5.2. Results

The results of the modeling phase of each algorithm are presented in Table 5. Furthermore, the precision, recall, accuracy and f1 score are presented for each one of the 5 classification algorithms that were performed. Additionally, the macro f1-score was calculated and presented for a better evaluation of the results obtained by each model.

Table 5 Model performance summary

Model	Precision	Recall	Accuracy	F1	Macro F1
Decision Tree	0.31	0.29	0.58	0.30	0.50
Random Forest	0.00	0.00	0.69	0.00	0.41
Adaboost	0.30	0.21	0.71	0.71	0.62
LinearSVC	0.36	0.36	0.60	0.36	0.53
MLP	0.33	0.57	0.51	0.42	0.50

As can be observed from the results presented above, Random Forest seems to obtain better results with respect to accuracy. However, given that the classes of the dataset were imbalanced, the accuracy is no longer a valid metric as can be observed from the precision and recall of the same model, where their values are both 0, meaning that the algorithm failed to distinguish the two categories and thus giving a macro f1 score of 0.41 that derives from the data of the one class.

Decision Tree has the lowest performance with an F1 score of 0.3 and 0.7 for the not depressed and depressed class respectively, and macro f1 score of 0.5. The LinearSVC model performed slightly better than the previous one, as it obtained F1 score of 0.36 for the not depressed class, 0.71 for the depressed class and macro F1 score of 0.53. all the above models reaching f1 score of 0.36 and macro f1 score 0.53. The impact of the imbalance of the two classes can be seen once again in the results of the Multilayer Perceptron as it has an accuracy of 0.51, which is the lowest of all the models, while the f1 score of the minority class (0.42) is the highest of the 3 models that have been described so far.

Given that the dataset is highly imbalanced, the AdaBoost algorithm was expected to perform better than the rest models, having in mind that it is one of the most suitable algorithms for this type of problems. In fact, Adaboost achieved an accuracy of 0.71 and an F1 score of 0.71 as well, for the not depressed class which is the minority. Provided that the chosen metric for the final evaluation of the ML models is F1 score, it is safe to say that Adaboost is the model that had the best performance.

6. Conclusions and future proposals

The thesis investigated methods that tackle the task of depression detection, from provided audio of clinical interviews. The valuable insights of the patients' audio features that can be provided through the speech during the interview time were explored and from that the task of audio-based classification was performed.

Audio data have been properly pre-processed and audio features provided by the dataset were extracted by the toolbox COVAREP. The final dataset was used to train five machine learning algorithms. Given the imbalance of the classes in the dataset, it comes with no surprise that the AdaBoost model performed overall better than the rest of the models, obtaining F1 score of 0.71 and accuracy of 0.71 as well.

Overall, the implemented machine learning models have proven to be able to detect depression using audio. This, relatively new, methodology opens new opportunities for patients suffering from depression to be accurately and in time diagnosed and treated for depression.

It is clear that in further research, the dataset can be additionally processed to achieve better results regarding the accuracy of the recognition of depression from the audio files provided by the DAIC-WOZ dataset. Furthermore, the imbalance of the classes in the dataset could be resolved using the oversampling method and the results of each method could be compared in order to have a clear perspective. In addition, with respect to parameter tuning, all models could be tested with modified parameters to achieve better evaluation measures, or given that it is computationally possible, it would be even more helpful to utilize useful tools to fine tune the parameters.

References

[1] “2020_Book_UnderstandingAcoustics.pdf.” https://library.oapen.org/viewer/web/viewer.html?file=/bitstream/handle/20.500.12657/42912/2020_Book_UnderstandingAcoustics.pdf?sequence=1&isAllowed=y.

[2] M. Morales, S. Scherer, and R. Levitan, “A Linguistically-Informed Fusion Approach for Multimodal Depression Detection,” in Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, New Orleans, LA, Jun. 2018, pp. 13–24. doi: 10.18653/v1/W18-0602.

[3] H. Negi, T. Bhola, M. S. Pillai, and D. Kumar, “A Novel Approach for Depression Detection using Audio Sentiment Analysis,” p. 4.

[4] T. Giannakopoulos, A Python library for audio feature extraction, classification, segmentation and applications. 2021. [Online]. Available: https://github.com/tyiannak/pyAudioAnalysis

[5] A. D. Rasamoelina, F. Adjailia, and P. Sincak, “A Review of Activation Function for Artificial Neural Network,” in 2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI), Herlany, Slovakia, Jan. 2020, pp. 281–286. doi: 10.1109/SAMI48414.2020.9108717.

[6] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” Speech Communication, vol. 71, pp. 10–49, Jul. 2015, doi: 10.1016/j.specom.2015.03.004.

[7] A. Belouali et al., “Acoustic and language analysis of speech for suicidal ideation among US veterans,” BioData Mining, vol. 14, no. 1, p. 11, Feb. 2021, doi: 10.1186/s13040-021-00245-y.

[8] S. Sharma, S. Sharma, and A. Athaiya, “ACTIVATION FUNCTIONS IN NEURAL NETWORKS,” IJEAST, vol. 04, no. 12, pp. 310–316, May 2020, doi: 10.33564/IJEAST.2020.v04i12.054.

[9] K. Fagan, “Amplitude and Intensity,” Discovery of Sound in the Sea, Sep. 09, 2016. https://dosits.org/science/sound/characterize-sounds/intensity/.

[10] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, “An introduction to decision tree modeling,” J. Chemometrics, vol. 18, no. 6, pp. 275–285, Jun. 2004, doi: 10.1002/cem.873.

[11] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, An Investigation of Depressed Speech Detection: Features and Normalization. 2011, p. 3000.

[12] C. Toh and J. P. Brody, Applications of Machine Learning in Healthcare. IntechOpen, 2021. doi: 10.5772/intechopen.92297.

[13] “Artificial Intelligence in Healthcare Market by Offering, Technology, Application, End User and Geography - Global Forecast to 2027.” https://www.reportlinker.com/p04897122/Artificial-Intelligence-in-Healthcare-Market-by-Offering-Technology-Application-End-User-Industry-and-Geography-Global-Forecast-to.html?utm_source=GNW.

[14] “Audio Data Analysis Using Deep Learning with Python (Part 1),” KDnuggets. https://www.kdnuggets.com/audio-data-analysis-using-deep-learning-with-python-part-1.html/.

[15] S. Scherer et al., “Automatic audiovisual behavior descriptors for psychological disorder analysis,” Image and Vision Computing, vol. 32, no. 10, pp. 648–658, Oct. 2014, doi: 10.1016/j.imavis.2014.06.001.

[16] A. Vázquez-Romero and A. Gallardo-Antolín, “Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks,” Entropy, vol. 22, no. 6, Art. no. 6, Jun. 2020, doi: 10.3390/e22060688.

[17] M. Valstar et al., “AVEC 2016 - Depression, Mood, and Emotion Recognition Workshop and Challenge.” arXiv, Nov. 22, 2016. [Online]. Available: http://arxiv.org/abs/1605.01600

[18] M. Valstar et al., “AVEC 2016 - Depression, Mood, and Emotion Recognition Workshop and Challenge.” arXiv, Nov. 22, 2016. [Online]. Available: http://arxiv.org/abs/1605.01600

[19] “Basics Of Audio.” https://www.music-production-guide.com/basics-of-audio.html.

[20] “Bipolar Disorder,” National Institute of Mental Health (NIMH). https://www.nimh.nih.gov/health/topics/bipolar-disorder.

[21] “Causes - Clinical depression,” nhs.uk, Feb. 15, 2021. https://www.nhs.uk/mental-health/conditions/clinical-depression/causes/.

[22] K. Santosh, N. Das, and S. Ghosh, “Chapter 2 - Deep learning: a review,” in Deep Learning Models for Medical Imaging, K. Santosh, N. Das, and S. Ghosh, Eds. Academic Press, 2022, pp. 29–63. doi: 10.1016/B978-0-12-823504-1.00012-X.

[23] Y. Fan, X. Cui, H. Han, and H. Lu, “Chiller fault diagnosis with field sensors using the technology of imbalanced data,” Applied Thermal Engineering, vol. 159, p. 113933, Aug. 2019, doi: 10.1016/j.applthermaleng.2019.113933.

[24] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, “COVAREP — A collaborative voice analysis repository for speech technologies,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2014, pp. 960–964. doi: 10.1109/ICASSP.2014.6853739.

[25] “COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide,” 2022. https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide.

[26] “Decision Tree Algorithm, Explained,” KDnuggets. https://www.kdnuggets.com/decision-tree-algorithm-explained.html/.

[27] L. Yang, D. Jiang, L. he, E. Pei, M. Oveneke, and H. Sahli, “Decision Tree Based Depression Classification from Audio Video and Language Information,” Nov. 2016, doi: 10.1145/2988257.2988269.

[28] Y. SONG and Y. LU, “Decision tree methods: applications for classification and prediction,” Shanghai Arch Psychiatry, vol. 27, no. 2, pp. 130–135, Apr. 2015, doi: 10.11919/j.issn.1002-0829.215044.

[29] “Depression.” https://www.who.int/news-room/fact-sheets/detail/depression.

[30] “Depression (major depressive disorder) - Symptoms and causes,” Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/depression/symptoms-causes/syc-20356007.

[31] “Depression Definition and DSM-5 Diagnostic Criteria,” Psycom.net - Mental Health Treatment Resource Since 1996. https://www.psycom.net/depression-definition-dsm-5-diagnostic-criteria/.

[32] “Depression Rates By Country 2021.” https://worldpopulationreview.com/country-rankings/depression-rates-by-country.

[33] S. Dham, A. Sharma, and A. Dhall, “Depression Scale Recognition from Audio, Visual and Text Analysis,” Sep. 2017.

[34] Y. Ozkanca, M. Göksu Öztürk, M. N. Ekmekci, D. C. Atkins, C. Demiroglu, and R. Hosseini Ghomi, “Depression Screening from Voice Samples of Patients Affected by Parkinson’s Disease,” DIB, vol. 3, no. 2, pp. 72–82, 2019, doi: 10.1159/000500354.

[35] S. C. Guntuku, D. Yaden, M. Kern, L. Ungar, and J. Eichstaedt, “Detecting depression and mental illness on social media: an integrative review,” Current Opinion in Behavioral Sciences, vol. 18, pp. 43–49, Dec. 2017, doi: 10.1016/j.cobeha.2017.07.005.

[36] M. Tasnim and E. Stroulia, “Detecting Depression from Voice,” in Advances in Artificial Intelligence, vol. 11489, M.-J. Meurs and F. Rudzicz, Eds. Cham: Springer International Publishing, 2019, pp. 472–478. doi: 10.1007/978-3-030-18305-9_47.

[37] E. Victor, Z. M. Aghajan, A. Sewart, and R. Christian, “Detecting Depression Using a Framework Combining Deep Multimodal Neural Networks With a Purpose-Built Automated Evaluation,” Psychological Assessment, vol. 31, May 2019, doi: 10.1037/pas0000724.

[38] T. Al Hanai, M. Ghassemi, and J. Glass, “Detecting Depression with Audio/Text Sequence Modeling of Interviews,” in Interspeech 2018, Sep. 2018, pp. 1716–1720. doi: 10.21437/Interspeech.2018-2522.

[39] D. Shin et al., “Detection of Minor and Major Depression through Voice as a Biomarker Using Machine Learning,” JCM, vol. 10, no. 14, p. 3046, Jul. 2021, doi: 10.3390/jcm10143046.

[40] Diagnostic and statistical manual of mental disorders: DSM-5™, 5th ed. Arlington, VA, US: American Psychiatric Publishing, Inc., 2013, pp. xliv, 947. doi: 10.1176/appi.books.9780890425596.

[41] P. Tzirakis, G. Trigeorgis, M. Nicolaou, B. Schuller, and S. Zafeiriou, “End-to-End Multimodal Emotion Recognition Using Deep Neural Networks,” IEEE Journal of Selected Topics in Signal Processing, vol. PP, Apr. 2017, doi: 10.1109/JSTSP.2017.2764438.

[42] M. Müller, Fundamentals of Music Processing, 1st ed. Springer, 2015. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-21945-5

[43] A. M. TURING, “I.—COMPUTING MACHINERY AND INTELLIGENCE,” Mind, vol. LIX, no. 236, pp. 433–460, Oct. 1950, doi: 10.1093/mind/LIX.236.433.

[44] W. Ertel, Introduction to Artificial Intelligence, 2nd ed. Springer International Publishing, 2017. doi: 10.1007/978-3-319-58487-4.

[45] S. Theodoridis, Machine learning: a Bayesian and optimization perspective, 2nd edition. London: Elsevier, Academic Press, 2020.

[46] “Major Depression,” National Institute of Mental Health (NIMH). https://www.nimh.nih.gov/health/statistics/major-depression

[47] “Major depressive disorder,” Wikipedia. Oct. 16, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Major_depressive_disorder&oldid=1050177387

[48] A. Koukopoulos, G. Sani, and S. N. Ghaemi, “Mixed features of depression: why DSM-5 is wrong (and so was DSM-IV),” The British Journal of Psychiatry, vol. 203, no. 1, pp. 3–5, Jul. 2013, doi: 10.1192/bjp.bp.112.124404.

[49] M.-C. Popescu, V. Balas, L. Perescu-Popescu, and N. Mastorakis, “Multilayer perceptron and neural networks,” WSEAS Transactions on Circuits and Systems, vol. 8, Jul. 2009.

[50] Music Similarity and Retrieval. [Online]. Available: https://link.springer.com/book/10.1007/978-3-662-49722-7

[51] M. A. Nielsen, “Neural Networks and Deep Learning,” 2015, [Online]. Available: http://neuralnetworksanddeeplearning.com

[52] “NIMH » Depression.” https://www.nimh.nih.gov/health/topics/depression.

[53] abbottds, “Perception of sound: Overview”, Available: https://sound.pressbooks.com/chapter/vibrations-in-sound-overview/

[54] J. Bueno-Notivol, P. Gracia-García, B. Olaya, I. Lasheras, R. López-Antón, and J. Santabárbara, “Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies,” International Journal of Clinical and Health Psychology, vol. 21, no. 1, p. 100196, 2021, doi: 10.1016/j.ijchp.2020.07.007.

[55] “Psychiatry.org - What Is Depression?” https://psychiatry.org:443/patients-families/depression/what-is-depression.

[56] T. Giannakopoulos, “pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis,” PLoS One, vol. 10, no. 12, p. e0144610, Dec. 2015, doi: 10.1371/journal.pone.0144610.

[57] “Random Forest | Introduction to Random Forest Algorithm,” Analytics Vidhya, Jun. 17, 2021. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest.

[58] J. Zhang, X. Wen, and M. Whang, “Recognition of Emotion According to the Physical Elements of the Video,” Sensors (Basel), vol. 20, no. 3, p. 649, Jan. 2020, doi: 10.3390/s20030649.

[59] “Self-Reported Symptoms of Depression and PTSD Are Associated with Reduced Vowel Space in Screening Interviews,” IEEE Trans. Affective Comput., vol. 7, no. 1, pp. 59–73, Jan. 2016, doi: 10.1109/TAFFC.2015.2440264.

[60] Z. Lin, Z. Hao, X. Yang, and X. Liu, Several SVM Ensemble Methods Integrated with Under-Sampling for Imbalanced Data Learning. 2009, p. 544. doi: 10.1007/978-3-642-03348-3_54.

[61] D. DeVault et al., “SimSensei kiosk: a virtual human interviewer for healthcare decision support,” in Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, Richland, SC, May 2014, pp. 1061–1068.

[62] A. L. Samuel, “Some Studies in Machine Learning Using the Game of Checkers,” p. 20.

[63] D. Jouvet, “Speech Processing and Prosody,” 2019, pp. 3–15. doi: 10.1007/978-3-030-27947-9_1.

[64] J. Gratch et al., “The Distress Analysis Interview Corpus of human and computer interviews,” in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), Reykjavik, Iceland, May 2014, pp. 3123–3128. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf

[65] G. Elert, “The Nature of Sound,” The Physics Hypertextbook, 2021, [Online]. Available: https://physics.info/sound/

[66] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," Psychol Rev, vol. 65, no. 6, pp. 386–408, Nov. 1958, doi: 10.1037/h0042519.

[67] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," Psychological review, 1958, doi: 10.1037/H0042519.

[68] K. Kroenke, T. Strine, R. Spitzer, J. B. W. Williams, J. T. Berry, and A. Mokdad, "The PHQ-8 as a measure of current depression in the general population.," Journal of affective disorders, 2009, doi: 10.1016/j.jad.2008.06.026.

[69] L. Lin, X. Chen, Y. Shen, and L. Zhang, "Towards Automatic Depression Detection: A BiLSTM/1D CNN-Based Model," Applied Sciences, vol. 10, no. 23, Art. no. 23, Jan. 2020, doi: 10.3390/app10238701.

[70] "Types of Depression: The 10 Most Common Depressive Disorders," Psycom.net - Mental Health Treatment Resource Since 1996. https://www.psycom.net/depression/types-of-depression.

[71] G. L. Team, "Types of Neural Networks and Definition of Neural Network," GreatLearning Blog: Free Resources what Matters to shape your Career!, Apr. 29, 2020. https://www.mygreatlearning.com/blog/types-of-neural-networks/.

[72] Understanding Acoustics. [Online]. Available: https://link.springer.com/book/10.1007/978-3-030-44787-8

[73] Abbott, Understanding Sound. [Online]. Available: https://sound.pressbooks.com/

[74] "Valerio Velardo," Valerio Velardo. https://valeriovelardo.com/.

[75] F. Benazzi, "Various forms of depression," Dialogues Clin Neurosci, vol. 8, no. 2, pp. 151–161, Jun. 2006.

[76] "What Are the Different Types of Depression?," Verywell Health. https://www.verywellhealth.com/different-types-of-depression-overview-5209204.

[77] "What is Artificial Intelligence (AI)?," Sep. 16, 2021. https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

[78] "What is Artificial Intelligence (AI)? - AI Definition and How it Works," SearchEnterpriseAI. https://searchenterpriseai.techtarget.com/definition/AI-Artificial-Intelligence.

[79] "Sound Waves," PASCO scientific. https://www.pasco.com/products/guides/sound-waves.