



WPI

Machine Learning for Mental Health Screening

A Major Qualifying Project submitted to the faculty of Worcester Polytechnic Institute in partial fulfillment of the requirements for the Degrees of Bachelor of Science in Computer Science

Submitted By:

Praise Eteng

Advised By:

Emmanuel Agu

March 25, 2022

This report represents the work of one or more WPI undergraduates submitted to the faculty as evidence of completion of a degree requirement. WPI routinely publishes these reports on its website without editorial or peer review.

Acknowledgements

There are a few people who I would like to thank for their contributions to this project. I would like to thank Professor Emmanuel Agu of the Computer Science Department at Worcester Polytechnic Institute for his guidance and advice. I would also like to thank Ph.D. student Wen Ge for explaining many of the machine learning concepts used in this project, answering my questions every step of the way, as well as providing some example code for machine learning classification. I would also like to thank WPI for providing me with the necessary resources needed to complete this project. Without the use of their Turing clusters, I would not have been able to generate any results from my experiments. Last but certainly not least, I would like to thank my family, their continuous support has not only helped me get through this project, but also helped me complete my tenure at WPI.

Abstract

Depression is one of the most prevalent mental disorders in the world, which can worsen existing medical conditions and could lead to suicide if left untreated. . The goal of this project is to use features from facial, audio and GPS modalities that can be gathered from a smartphone to assess and track trajectories of depression. Features were extracted from the DAIC-WOZ and StudentLife datasets and the Naïve Bayes, random forest classifier, Support Vector Machine with stochastic gradient descent, and XGBoost classifiers were used to detect depression levels based on the PHQ score. The best performing classifier used was the XGBoost algorithm, with a mean accuracy of 0.82 for 2 bin classification and 0.639 for 3 bin classification. Features from the GPS modality had the highest metrics overall with a mean accuracy of 0.8875 for two 2-bin classification and 0.6625 for 3-bin classification.

Table of Contents

<i>Acknowledgements</i>	1
<i>Abstract</i>	2
<i>Table of Contents</i>	3
<i>List of Figures</i>	5
<i>List of Tables</i>	6
<i>1 Introduction</i>	7
1.1 The Goal of this Major Qualifying Project (MQP)	8
1.2 Prior Work	9
<i>2 Related Work</i>	10
<i>3 Methodology</i>	18
3.1 Datasets	19
3.2 Pre-processing	20
3.3 Patient Health Questionnaire (PHQ)	21
3.4 Features	23
3.4.1 Facial Features	23
3.4.2 Audio Features	30
3.4.3 GPS Features	31
3.5 Evaluation Metrics	32
3.6 Classification Algorithms	34
3.7 Grid Search	37
3.7.1 Hyperparameters	38
<i>4 Implementation</i>	40
<i>5 Results</i>	41
<i>6 Discussion</i>	46
6.2 Limitations	46

7 Conclusion 49

7.1 Future Work 49

Bibliography 51

List of Figures

1	Figure 1: PHQ-9: depression scale and pre and post class outcomes	12
2	Figure 2: Correlation between depression severity and sensor data that was collected.....	12
3	Figure 3: 10 eye points.....	14
4	Figure 4: Facial landmarks.....	15
5	Figure 5: Baseline results for emotion on the Development (D) and Test (T) partitions from audio, video, and text feature sets, and their early fusion (multimodal)	16
6	Figure 6: Baseline results for depression severity estimation on the Development (D) and Test (T) partitions from audio video and audio-video modalities	17
7	Figure 7: Machine learning pipeline.....	19
8	Figure 8: PHQ 9 Questionnaire.....	21
9	Figure 9: Facial landmark coordinates on a subject's face.....	26
10	Figure 10: Facial pose movements.....	28
11	Figure 11: Gaze direction.....	29
12	Figure 12: Confusion Matrix.....	33

List of Tables

1	<i>Table 1: Action Units and their representations</i>	23
2	<i>Table 2: Hyperparameters used for random forest classifier</i>	38
3	<i>Table 3: Hyperparameters used for Support Vector Machine</i>	38
4	<i>Table 4: Hyperparameters used for XGBoost</i>	39
5	<i>Table 5: Hyperparameters used for Naïve Bayes</i>	39
6	<i>Table 6: ROC-AUC for 2 bin classification</i>	41
7	<i>Table 7: Accuracy for 2 bin classification</i>	41
8	<i>Table 8: Table 4: F1-score for 2 bin classification</i>	42
9	<i>Table 9: Confusion matrix for 2 bin classification</i>	42
10	<i>Table 10: ROC-AUC for 3 bin classification</i>	43
11	<i>Table 11: Accuracy for 3 bin classification</i>	43
12	<i>Table 12: F1 score for 3 bin classification</i>	44
13	<i>Table 13: Confusion matrix for 3 bin classification</i>	44

1 Introduction

Depression, or Major Depressive Disorder is a mood disorder that causes loss of interest and a persistent feeling of sadness [28]. About 17.3 million adults (7.1% of the adult population) in the United States have had a Major Depressive episode, defined as a period of a least two weeks where one suffers from symptoms of depression such as problems with concentration, eating, energy, self-worth, and sleep [27]. Other symptoms of depression include, but are not limited to: Persistent sad, anxiety, or “empty” mood, feelings of hopelessness, or pessimism, irritability, loss of interest or pleasure in hobbies and activities, moving or talking more slowly, and feeling restless or having trouble sitting still [40]. Risks of depression include the accentuation and worsening of already present illnesses in the body such as diabetes, cancer, and heart disease., as well as thoughts of death or suicide, or suicide attempts. According to the AFSP, suicide has claimed the lives of 47,511 Americans, making it the 10th leading cause of death in the country [6]. One possible solution this issue is to detect symptoms early, which can be done through using data from smart devices that users carry with them almost all the time.

Smartphones are the most popular electronic platform in the United States. As of 2021, 84% of Americans own a smartphone compared to 74% of Americans owning a computer, 45% owning a tablet, and about 21% of adults own a smartwatch or a fitness tracker [7]. Smartphones are heavily used, with 65.6% Americans claiming to check their phone at least 160 times or more in a day and 47% of adults stating that they cannot live without their phones [1]. Smartphones can track a wide variety of data through their sensors, which can then be analyzed to detect various ailments and for tracking trajectories of depression based on distinct changes in behavior. For example, a person with depression is less likely to move [40], which can be detected using data from the

accelerometer and gyroscope to track a user's step count and activities [34]. Other symptoms such as slow speech patterns and sad, anxious, or "empty" moods can be detected using a smartphone's microphone and infrared sensor [36]. These methods have been utilized in the past with promising results, for example Canzian *et al* had a mean absolute correlation of 0.432 and an average p-value of 0.068 when attempting to find a relationship between mobility metrics gathered from GPS data on a phone and depressive moods [8].

1.1 The Goal of this Major Qualifying Project (MQP)

The goal of this MQP is to:

- Use data from facial, audio and GPS modalities to track trajectories of depression using machine learning models.
- Evaluate the models using various machine learning classification metrics such as accuracy, AUC-ROC, F1 score, and a confusion matrix to assess their performance.

These modalities were chosen because they can easily be recorded from a mobile phone in a passive manner, which is important because it eliminates any possible bias that can arise when a user has to manually enter data. This MQP also aims to evaluate what modality would be the best at identifying depression. The features used from the facial modality will be 2D and 3D points on the face, Action Units, gaze of the eyes, and facial Poses. Features extracted from the audio modality will be the collection of features obtained from the COVAREP algorithm. The features extracted from GPS modality are Location Variance, Speed Mean, Total Distance and Transition time. The machine learning models that will be used are Support Vector Machine with Stochastic Gradient Descent, random forest classifier, Naïve Bayes, and XGBoost. These models will be trained using the DAIC-WOZ [35], and the StudentLife [42] dataset. Based on the results, experts

can have a foundation as to what algorithms to use to best identify a patient with depression as well as what symptoms to look out for.

1.2 Prior Work

There have been many MQPs in the past that have attempted to track trajectories of depression. A project in 2018 created a mobile application that instantly collected as much data as possible from a subject's social media usage, GPS, calls and texts two weeks prior to whenever they initiated an assessment [12]. The information is then used to provide instant feedback on the severity of their depression to the doctor and the patient. MQP teams in 2019 and 2021 used that foundation to further build upon the app, surveying users on what data they were willing to share , as well as using different datasets to test their application, such as the Moodable dataset and the Amazon Mechanical Turk platform [34]. This MQP differs from their work because they looked to gather all features instantaneously, while this MQP will be detecting depression using datasets that were gathered over a long period of time. In addition to finding the best modality, this study also aims to find the best machine learning algorithm to identify depression severity as well. Another MQP conducted in 2020 aimed to identify depression using machine learning models with sub clip boosting, convolutional neural networks, and long-term short-term memory models on audio and text modalities [37]. The difference between this MQP and the 2020 study is that only machine learning models will be used on audio, facial, and GPS modalities.

2 Related Work

Matteo *et al* looked to find the relationship between environmental audio and symptoms of depression and anxiety [27]. To do this an Android application was developed that was used to gather environmental audio and track the presence of English-speaking voices from 84 participants for two weeks. This data was then analyzed using as ground truth self-reported Liebowitz Social Anxiety Scale (LSAS), the Generalized Anxiety Disorder seven-item scale, the Patient Health Questionnaire eight-item scale (PHQ-8), and the Sheehan Disability Scale (SDS). Based on the audio extracted and the reports on the four scales mentioned above, no statistically significant relationships were found between the collected environmental audio and severity of anxiety. However when it came the severity of depression, the inferred patterns of daily activity and inactivity from the environmental audio volume had a correlation ($r=-0.37$; $P<.001$). Sleep disturbance inferred from the environmental audio volume was also correlated with the severity of depression ($r=0.23$; $P=.03$). A measure of social interaction based on the detection of speaking voices in the environmental audio was also correlated with depression ($r=-0.37$; $P<.001$) and functional impairment ($r=-0.29$; $P=.01$) [27]. The approach taken by this MQP is similar to this study in the sense that audio and the PHQ will be used to track symptoms of depression. However, in contrast, this MQP will use multiple modalities such as GPS, facial images, and audio. In addition, features extracted from these modalities will only be used to track symptoms of depression, not anxiety which is a different ailment.

A 2019 study by Ray *et al* aimed to create a multilevel neural network to predict depression from audio, visual, and text to predict depression [33]. The dataset used to train their network was the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ). This dataset contains the

transcripts, as well as audio and video recordings of conversations conducted between participants and Ellie, a virtual interviewer. The Bi-LSTM and LSTM neural networks models were used for text and visual modalities, while a Deep Spectrum and VGG network was used to extract features from the audio modality. The audio-based models outperformed the base models by 9.3%, 6.6%, and 8.7% [30]. The models that classified visual features outperformed the four baseline models by 29.80%, 29.04%, 10.04%, and 14.46% [33]. This study uses audio and visual modalities which are some of the modalities that will be used in this MQP, however it will not use the text modality. This is due to the findings of 2020 study of what features users would be willing to share. They found that only 49%, 48%, and 45% of users were willing to give access to Twitter tweets, Facebook posts and data from Text Chat apps respectively [13], which are features that would be extracted from the text modality. Another area where this study differs from the goals of this MQP is that the models will classify various data types separately and not fused them into a hybrid model or trained on only the DAIC-WOZ dataset. Instead, models will be trained using the StudentLife dataset in addition to the DAIC-WOZ dataset.

Wang *et al* conducted a study in 2014 aimed to look at the changes in behavior and mental health of students at Dartmouth College over the course of a 10-week period. 48 students took a pre and post psychological survey (41 students completed the post psychological survey) that included depression questions and used a mobile application called Student Life to track sleep patterns, the number of conversations and the duration of each conversation per day. Physical activity, location, the number of people around a student through the day, outdoor and indoor mobility stress level through the day, eating habits, app usage, and in-situ comments on current events were also measured with the app [42].

depression severity	minimal	minor	moderate	moderately severe	severe
score	1-4	5-9	10-14	15-19	20-27
number of students (pre-survey)	17	15	6	1	1
number of students (post-survey)	19	12	3	2	2

Figure 1: PHQ-9 depression scale and pre and post class outcomes [42]

Figure 1 above shows the results of psychological surveys taken before and after a 10-week period. While more students fell into the “minimal” category of the PHQ-9 depression scale, one student fell into the moderately severe and severe category each. In figure 2 below the features collected are displayed with their associated r-scores, showing that there is a statistically significant correlation between sleep duration, conversation frequency, and number of co-locations and

automatic sensing data	r	p-value
sleep duration (pre)	-0.360	0.025
sleep duration (post)	-0.382	0.020
conversation frequency during day (pre)	-0.403	0.010
conversation frequency during day (post)	-0.387	0.016
conversation frequency during evening (post)	-0.345	0.034
conversation duration during day (post)	-0.328	0.044
number of co-locations (post)	-0.362	0.025

Figure 2 Correlation between depression severity and sensor data that was collected [37]

depression severity. This is because people who sleep less, have less conversational interactions, converse later in the day, and have fewer co-locations with other students are more likely to be depressed [37]. This MQP differs from the StudentLife study because it will go beyond seeing

how mental health changes over a certain length of time. It will use the GPS data gathered from this study and compare it with the audio and facial modality from another dataset to see which one is the best indicator for detecting depression.

Hatton *et al* investigated how useful machine learning can be in predicting the persistence of depressive symptoms in older adults. To do this, they used data from a previous trial consisting of 284 patients who were at least 65 years old and met the criteria for depression. To meet the criteria, they had to screen positive for the Whooley questions, a two-question instrument used to determine depression in patients. They also had to meet sub threshold depressive symptoms according to the DSM-IV [18]. The study used a logistic regression model that used a backwards stepwise approach to select predictor variables. For their machine learning model, they used a form of Extreme Gradient Boosting (XGBoost) that was implemented in R [18]. When comparing the ability of the models to predict depressive symptoms (a PHQ-9 score over 10) they found that the AUC value was higher on the machine learning approach (0.72) compared to the logistic regression approach (0.67). This is significant because the AUC value is an overall metric of the potential utility as a screening method [18]. While the XGBoost model will be used in this MQP, it will be implemented in Python instead of R. Another difference between this MQP and the study conducted by Hatton *et al* is that different machine learning models will be compared to each other, instead of being compared to a logistic regression model.

Ringeval *et al* [35] compared the relative merits of the various approaches to depression and emotion recognition from real-life data. In the study they went over two sub challenges: the affect sub challenge, and the depression sub challenge [35]. The affect sub challenge uses the Sentiment

Analysis in the Wild (SEWA) database, which contains 64 subjects aged between 18 and 60 having a discussing an advertisement they watched. Annotators later give a label of either arousal, valence, or liking to the recordings. The audio features extracted from the SEWA dataset were functionals, bag of audio words, Low level descriptors extracted every 10ms consisting of energy, spectral and cepstral features, pitch, voice quality, and micro prosodic features. The video features extracted were face orientation (Pitch, yaw, roll), pixel coordinates for 10 eye points (for both the x and y coordinate) and for 49 facial landmarks (for both the x and y coordinates). The Features are shown in figure 3 and 4 below

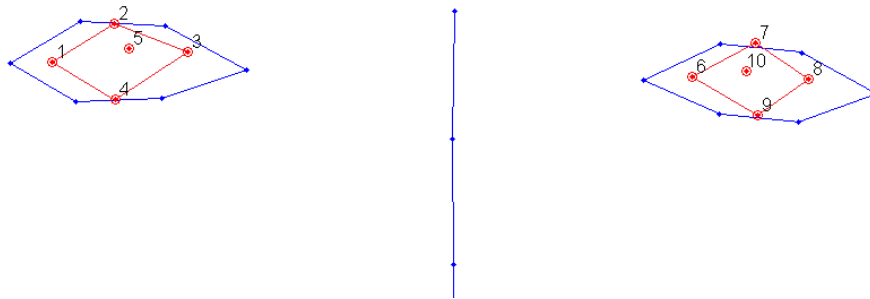


Figure 3: 10 eye points



Figure 4: 49 facial landmarks

The only text feature extracted was the bag of text words. The dataset was trained using a Support Vector Regressor and the metrics used for evaluation was the concordance correlation coefficient [35]. The results are shown in figure 5 below

Modality	Arousal	Valence	Liking
D-Audio	.344	.351	.081
D-Video	.466	.400	.155
D-Text	.373	.390	.314
D-Multimodal	.525	.507	.235
T-Audio	.225	.244	-.020
T-Video	.308	.455	.002
T-Text	.375	.425	.246
T-Multimodal	.306	.466	.048

Figure 5: Baseline results for emotion on the Development (D) and Test (T) partitions from audio, video, and text feature sets, and their early fusion (multimodal) [35]

The Concordance Correlation Coefficient for the arousal and liking label was highest for the text partition at 0.375 and 0.246 respectively, while the valence label was highest when using the multimodal partition at 0.466.

The Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) dataset was used for the depression sub challenge, which consists of clinical interviews designed to support the diagnosis of psychological distress conditions [35]. The participants were labeled as depressed if they had a PHQ-8 score greater than 10. The audio features extracted from the DAIC-WOZ dataset were fundamental frequency, and voicing. Voice quality, normalized amplitude quotient (NAQ), quasi open quotient (QOQ), the difference in amplitude of the harmonics of the differentiated glottal source spectrum (H1H2) parabolic spectral parameter (PSP), maxima dispersion quotient (MDQ), spectral tilt/slope of wavelet responses (peak-slope), and shape parameter of the Liljencrants-Fant model of the global pulse dynamics (rd) Mel cepstral coefficients (MCEP0-24),

harmonic model and phase distortion mean (HMPDM0-24) and deviations (HMPDD0-12). The video features extracted were 2D and 3D points on the face, Histogram of Oriented Gradients (HOG), gaze direction estimate for both eyes, 3D position and orientation of the head, and action units (AUs) [26]. DAIC-WOZ was trained using random forest regression and the metrics used to evaluate the regressors are the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). The baseline results are shown in figure 6 below

Partition	Modality	RMSE	MAE
Development	Audio	6.74	5.36
Development	Video	7.13	5.88
Development	Audio-Video	6.62	5.52
Test	Audio	7.78	5.72
Test	Video	6.97	6.12
Test	Audio-Video	7.05	5.66

Figure 6: Baseline results for depression severity estimation on the Development (D) and Test (T) partitions from audio video and audio-video modalities [35]

The best performance was obtained using the video modality [35] with an RMSE of 6.97. Ringeval *et al's* study differs from this MQP because it will go beyond just using audio and video modalities, it will also analyze features from the GPS modality. This MQP will also use classifiers instead of regressors to detect depression. One final difference is that this MQP focuses on detecting depression and will not look at emotion like this study has.

3 Methodology

For this project, we will be training features from the audio and facial modalities from the DAIC-WOZ dataset and examining which features can best detect depression. In addition to DAIC-WOZ, the Student Life dataset will also be used to extract features from the GPS modality. The datasets will be split randomly with an 80-20 ratio, where 80% of the dataset will be used for training the machine learning models, while 20% of the dataset will be used to test their performance.

PHQ scores from participants in both datasets will be used as ground truth labels. Participants in the DAIC-WOZ dataset took the PHQ-8 test while participants in the StudentLife dataset took a PHQ-9 test. To maintain consistency, the PHQ score for participants in the StudentLife dataset was recalculated to not include the final question about suicide ideation. Participants in the Studentlife dataset also took the PHQ both before and after the experiment. To discover any bias that may come by picking one score over the other, both PHQ-8 scores were used as ground truth for depression classification separately.

The machine learning models used on the dataset were XGBoost, random forest classifier, Support Vector Machines, and Naïve Bayes. Using the training dataset, the machine learning models had to classify the testing dataset into both two and three bins. In order to get the best performance out of the classifiers a grid search was performed on them. The categories used for two bins were 0-10 (mild depression) and 11-21 (severe depression) while the categories for three bins were 0-7 (mild depression) 8-14 (moderate depression) and 15-21 (severe depression).

The performance of the machine learning models were measured by the following evaluation metrics: accuracy, confusion matrix, F1 score, and AUC-ROC. A diagram outlining this pipeline is shown in figure 7 below

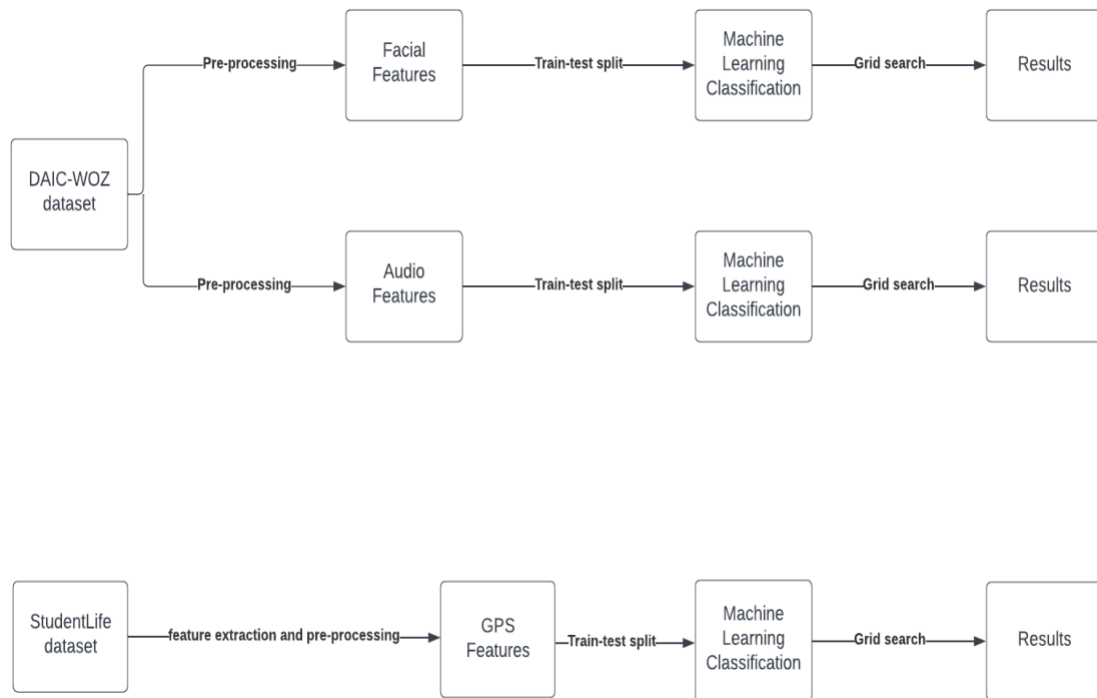


Figure 7: Our Machine learning pipeline for classifying depression.

3.1 Datasets

DAIC-WOZ The Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) is a dataset that was part of a larger effort aimed to create a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illness in 2014. The dataset contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and PTSD. Data collected include audio and video recordings and extensive questionnaire responses; and includes interviews of 189 patients conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room [35].

StudentLife The StudentLife dataset is from 48 undergrads and grad students at Dartmouth over a 10-week spring term [42]. This data was gathered to gauge how a student's mental health could be impacted. The dataset is comprised of:

- objective sensing data: sleep (bedtime, duration, wake up); conservation duration, conversation frequency; physical activity (stationary, walk, run).
- location-based data: location, co-location, indoor and outdoor mobility.
- other phone data: light, Bluetooth, audio, Wi-Fi, screen lock/unlock, phone charge, app usage.
- self-reports: affect (PAM), stress, behavior, Boston bombing reaction, cancelled classes, class opinion, comment, Dartmouth now, Dimension incident, Dimension protest, dining halls, events, exercise, Green Key, lab, mood, loneliness, social and study spaces.
- pre-post surveys: PHQ-9 depression scale, UCLA loneliness scale, Positive and Negative Affect Schedule (PANAS), perceived stress scale (PSS), big five personality, flourishing scale, Pittsburgh Sleep Quality Index (PSQI), veterans RAND 12 item health (VR12)
- academic performance data: class information, deadlines, grades (grades, term GPA, cumulative GPA), piazza data
- Dining data: meals data, location and time
- Seating data: seating position of students in Android programming
- Entry and exit surveys

3.2 Pre-processing

Before any classification or testing was done all the data was organized by modality, and then by feature type. Once complete, the data was normalized using a min-max scaler, which translates every feature to be within a range of zero and one.

3.3 Patient Health Questionnaire (PHQ)

The Patient Health Questionnaire (PHQ) is a validated instrument used by clinicians to diagnose depression disorders and measure depression severity [23]. It contains 9 questions shown in figure 8 below.

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

Figure 8: PHQ 9 Questionnaire







The PHQ is self-administered by the patient and assesses 8 diagnoses, divided into threshold disorders that correspond to specific DSM-IV diagnoses such as major depressive disorder, panic disorder, other anxiety disorder, and bulimia nervosa, as well as subthreshold disorders whose criteria encompass fewer symptoms than are required for any specific DSM-IV diagnoses such as









other depressive disorders, probable alcohol abuse/dependence, somatoform, and binge eating disorder. The PHQ-9 score can range from 0 to 27, since each of the 9 items can be scored from 0 (not at all) to 3 (nearly every day). In this project however, the PHQ-8 questionnaire will be used, which does not include the final question of suicide ideation. In this MQP, patients PHQ-9 scores generated from their answers to the 9 questions, were utilized as ground truth labels for machine learning prediction of their depression levels.

3.4 Features

3.4.1 Facial Features

Action Units Action Units refer to the facial action coding system, which are a set of facial movements used to determine the emotion of a participant [39]. The DAIC-WOZ dataset utilizes Action Units 1, 2, 4, 5, 6, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, and 45. Each Action Unit value was either a regression output marked as “_r” or a binary label “_c” indicating whether the unit is present or not. Their representations are listed in table 1 below:

Action Unit	Image	Description	Regression or Binary?
1		Inner Brow Raiser	Regression
2		Outer Brow Raiser (unilateral, right side)	Regression
4		Brow Lowerer	Both
5		Upper Lid Raiser	Regression
6		Cheek Raiser	Regression
9		Nose Wrinkler	Regression

10		Upper Lip Raiser	Regression
12		Lip Corner Puller	Both
14		Dimpler	Regression
15		Lip Corner Depressor	Both
17		Chin Raiser	Regression
20		Lip Stretcher	Regression
23		Lip Tightener	Binary
25		Lips Part	Regression



26		Jaw Drop	Regression
28		Lip Suck	Binary
45		Blink	Binary

Table 1: Action Units and their representations

A study performed in 2009 aimed to incorporate facial expressions and voice observations to identify mental disorders [9]. They used the manual facial action coding system (FACS), Active Appearance Modeling (AAM) and pitch extraction to measure facial and vocal expression. Using SVM classifiers for AAM and FACS and logistic regression for voice, they were able to accurately detect depression 88% of the time using FACS, and 79% of the time using AAM and voice [9].

Facial landmarks This feature consists of 68 2D and 3D points on the face, determined by video. These coordinates aim to identify facial features such as the chin, eyebrows, eyes, jaw, nose, and mouth. An image of the points on the face is shown in figure 9 below

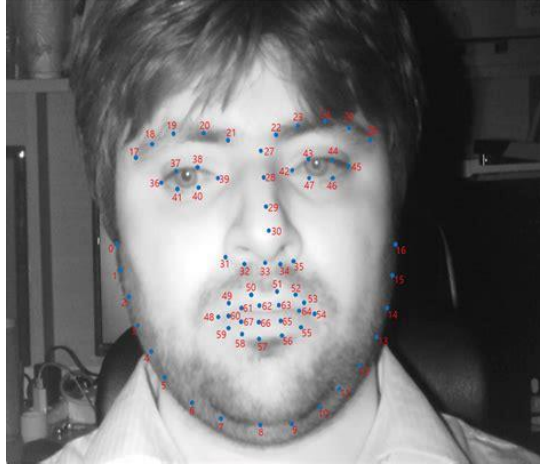


Figure 9: Facial landmark coordinates on a subject's face

A plethora of algorithms have been used to identify key points on the face. A study conducted in 2016 categorizes these algorithms into three major groups: holistic methods, Constrained Local Model (CLM) methods, and regression-based methods [43]. Holistic methods aim to build models to represent the entire facial appearance and shape information. A classic example of a holistic Method would be the Active Appearance Model (AAM) which fits facial images in accordance with a select number of coefficients, controlling both the facial appearance and shape variations [43]. Facial landmarks are found using the formula:

$$x = cR_{2d}(\Theta) \left(s_0 + \sum_{n=1}^{K_s} p_n * s_n \right) + t$$

Where $R_{2d}(\Theta)$ is the rotation matrix, c and t are the scale and translation parameter respectively, s_n is the facial shape bases p_n is the shape coefficient and K_s is the number of bases [43].

CLMs estimate the landmark locations based on the global facial shape patterns [43]. This is done by minimizing the misalignments error subject to the shape patterns:

$$\tilde{x} = arg_x min Q(x) + \sum_{d=1}^D D_d(x_d, \mathcal{J})$$

x_d represents the positions of different landmarks in x . $D_d(x_d, \mathcal{J})$ represents the local confidence score around x_d . $Q(x)$ represents a regularization term to penalize the infeasible or anti-anthropology face shapes in a global sense [43].

Regression-Based Methods skip the process of building a global face model and directly learn the mapping from image appearance to the landmark locations [43]. One type of method in particular, the cascaded regression method, performs an initial estimate of landmark locations before gradually upgrading. The formulas below are for the initial landmark locations and shape updates, respectively:

$$\delta \tilde{\mathbf{x}} = \underset{\delta \mathbf{x}}{\operatorname{arg\,min}} \|\Phi(\mathcal{J}(x^*)) - \Phi(\mathcal{J}(x_0 + \delta \mathbf{x}))\|_2^2$$

$$\delta \mathbf{x} = -2\mathbf{H}_f(\mathbf{x}_0)^{-1} \mathbf{J}_\Phi^T(\boldsymbol{\phi}(\mathcal{J}(\mathbf{x}_0)) - \boldsymbol{\phi}(\mathcal{J}(\mathbf{x}^*)))$$

2D and 3D features have been utilized to detect depression in patients many times before. For example, a study conducted in April 2021 used 2D and 3D points on a subject's face while they were doing emotional stimulus tasks to detect depression with an accuracy of 0.774 [15].

Poses This feature refers to a subject's 3D position and orientation of the head relative to the view of a camera. Examples of possible facial poses are shown in figure 10 below

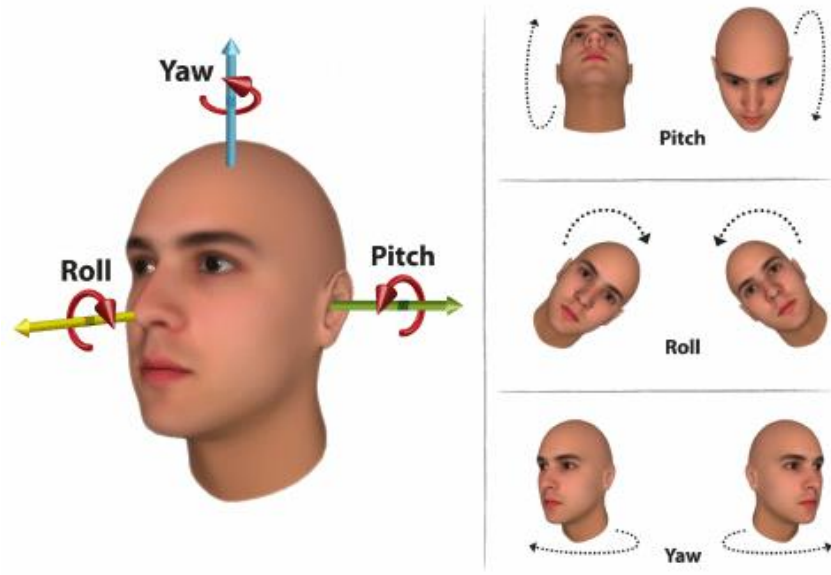


Figure 10: Facial pose movements

A study conducted in 2008 found that there are eight categories that methods for head pose estimations fall into [30]. The first category is appearance template methods, where the image of a head is compared to various other archetypes and uses the pose of the most similar one [30]. The next category is the detector array methods which train groups of head detectors in a specific pose and assign a discrete pose to the detector with the greatest support [30]. There are nonlinear regression methods that develop a functional mapping from the image or feature data using nonlinear regression tools. Manifold embedding methods seek low-dimensional manifolds that model the continuous variation in head pose. New images can be embedded into these manifolds and then used for embedded template matching or regression. Flexible models fit a non-rigid model to the human facial structure in the image plane. Head pose is estimated from feature-level comparisons or from the instantiation of the model parameters. Geometric methods use the location of features such as the eyes, mouth, and nose tip to determine pose from their relative configuration. Tracking methods recover the global pose change of the head from the observed

movement between video frames. Hybrid methods combine one or more of these methods to overcome the limitations inherent in any single approach.

Alghowinem *et al* conducted a study with the goal of detecting depression using head poses and analyzing movement [3]. On average, they were able to recognize a subject with depression 71.3% of the time [3], showing that utilizing poses is effective in depression detection.

Gaze This feature refers to the direction of a person's eyes and has a close relation with facial poses as illustrated in figure 11 below

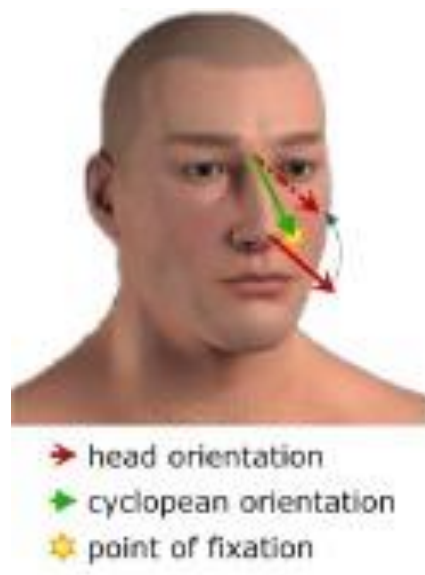


Figure 11: Gaze direction

Given a head pose a rough estimation of gaze can be determined if the eyes are obscured, but when the eyes are in view, head pose is required to accurately predict depression. This is because the orientation of the head dictates gaze direction [40].

Just like the pose feature, gaze has been successfully used as an indicator of depression, with Alghowinem *et al* used a Gaussian Mixture Models and Support Vector Machine hybrid

classifier and achieved a 70% accuracy, and 75% accuracy when using only a Support Vector Machine classifier [2].

3.4.2 Audio Features

Cooperative Voice Analysis Repository for Speech Technologies (COVAREP) Audio features were obtained using COVAREP, an open-source toolbox that uses feature extraction methods to capture voice quality and prosodic characteristics of the speaker [35]. The features used fall into three categories: prosodic, voice quality, and spectral.

Prosodic features consist of fundamental frequency (F0) and voicing (VUV). Fundamental frequency refers to the lowest rate at which a waveform repeats itself. The formula used to find the F0 is $f_0 = v/4L$ where v is the speed of the wave and L is the tube length. Voicing refers to whether a signal was produced via vocal fold action, as this would make a periodic vibration.

The voice quality features are made up of the Normalized Amplitude Quotient (NAQ), a method that parametrizes the glottal closing phase using two amplitude-domain measurements from waveforms estimated by inverse filtering [4]. The Quasi-Open Quotient (QQQ), which measures the amount of time that the pulse amplitude is above a certain limit [4]. The difference in amplitude of the first two harmonics of the differentiated glottal source spectrum (H1H2). The Parabolic Spectral Parameter (PSP), a method that generates a number that describes how the spectral decay of a given glottal flow performs in accordance with the theoretical limit corresponding with the maximal spectral decay [4]. The Maxima Dispersion Quotient (MDQ) is a method that analyzes the Linear Prediction residual to quantify how impulse-like glottal excitation is. Finally, the spectral tilt/slope of wavelet responses (peak-slope), and shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics (Rd) make up the voice quality features as well.

The spectral features used were the Mels cepstral coefficients (MCEP0-24), harmonic model and phase distortion mean (HMPDM0-24), and deviations (HMPDD0-12).

The COVAREP toolbox has previously been used to extract features to detect depression with great success. Yalamanchili *et al* used the DAIC-WOZ database to extract prosodic, spectral, and voice quality features. These features were then trained on a classification model which was then used within an app to detect depression on 50 subjects with 90% accuracy [44]

3.4.3 GPS Features

The GPS features were collected from a study conducted by Gerych et al to detect depression using low level features that can be detected using a smartphone [16] some of the features used were:

Location Variance The Location Variance is calculated by:

$$Location\ Variance = \log(\sigma_{lat}^2 + \sigma_{long}^2)$$

Where σ_{lat}^2 and σ_{long}^2 are the variance of a user's latitude and longitude, respectfully [16]. This feature could be useful for detecting depression because lack of movement is a symptom of depression [40].

Speed Mean The Speed Mean is calculated by:

$$Speed\ mean = \frac{1}{n} \sum_i^n \sqrt{\left(\frac{lat_i - lat_{i-1}}{t_i - t_{i-1}}\right)^2 + \left(\frac{long_i - long_{i-1}}{t_i - t_{i-1}}\right)^2}$$

Where n is the number of timestamps, and lat_i and $long_i$ are the user's latitude and longitude at time i [16]. The feature could be useful at detecting it. A lower speed mean indicates a lack of moment, could be an indicator of depression.

Total Distance A user's total distance is calculated by:

$$Total\ Distance = \sum_i^n \sqrt{(lat_i - lat_{i-1})^2 + (long_i - long_{i-1})^2}$$

Where n is the number of timestamps, and lat_i and $long_i$ are the user's latitude and longitude at time i [16]. This feature could be helpful at detecting depression since a lack of movement is a symptom of depression [40]. If a user with depression is less likely to move then they will cover less distance, which can be captured by the feature.

Transition Time A user's transition time is calculated by:

$$Transition\ Time = \frac{Time\ Moving}{Total\ Time}$$

If a user has a low transition time it could indicate that they don't spend much time moving, which can be useful at detecting depression since lack of movement is a symptom of the disorder [40].

A study that used the following GPS features were able to detect users with depression at high rate, earning an AUC-ROC score of 0.92 [16].

3.5 Evaluation Metrics

F1-score The F1 score measures a model's accuracy on a dataset, used to evaluate binary classification systems [9]. The F1 score is derived from combining the precision and recall score:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives

Accuracy The accuracy metric measures how what percentage of predictions the model guesses correctly [9].

$$Accuracy = \frac{True\ Negative + True\ Positive}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

Confusion Matrix The Confusion Matrix is a 2x2 table that illustrated the performance of an algorithm [38]. The rows of the matrix represent the actual values while the columns represent the predicted values. Figure 12 below gives an example of what a matrix looks like for binary classes.

		ACTUAL	
		Negative	Positive
PREDICTION	Negative	TRUE NEGATIVE	FALSE NEGATIVE
	Positive	FALSE POSITIVE	TRUE POSITIVE

Figure 12: Confusion Matrix

A True Negative (TN) happens when the model prediction is negative, and the actual value is also negative. A False Negative (FN) occurs when the model prediction is negative, but the actual value is positive. A true positive (TP) happens when the model prediction is positive, and the actual value is also positive. A false positive (FP) occurs when the model prediction is positive, but the actual value is negative. Using these values, the True Positive Rate (TPR) True Negative Rate (TNR) False Positive Rate (FPR) and False Negative Rate (FNR) can be calculated [38]:

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

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

$$TNR = \frac{TN}{TN + FP}$$

$$FPR = \frac{FP}{TN + FP}$$

AUC-ROC The ROC, or the receiver operating characteristic curve is a graph that visualizes the performance of a binary classifier. It is plotted with the sensitivity against the 1-specificity, where the sensitivity is on the y-axis and 1-specificity is on the x-axis [9]. The AUC is the area under the ROC. The AUC measures how well the model can discriminate between classes [9]. The higher the AUC the better the performance. The formula for the AUC is [31]

$$AUC(f) = \frac{\sum_{t_0 \in D^0} \sum_{t_1 \in D^1} 1[f(t_0) < f(t_1)]}{|D^0| \cdot |D^1|}$$

Where $1[f(t_0) < f(t_1)]$ is an indicator function that returns 1 if and only if $f(t_0) < f(t_1)$ otherwise, it returns 0. D^0 is the set of negative examples and D^1 is the set of positive examples.

3.6 Classification Algorithms

Support Vector Machines The Support Vector Machine is a method that is widely used for classification tasks. This is done by maximizing the distance between the data points of two separate classes in an N-dimensional space, where N is the number of features. The data points are classified using hyperplanes, which are boundaries that separate one class from another. The dimension of the hyperplane is dependent on the number of input features. For example, when

there are two input features the hyperplane is a line, and when there are three input features the hyperplane becomes a 2-D plane. The placement of the hyperplane is dependent on support vectors. A support vector is a data point that dictates the placement of the hyperplane [14]. SVMs maximize the distance between those data points and the hyperplane by calculating the hinge loss:

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

If the predicted value is the same sign as the actual value, then the cost is 0. If not the loss value is calculated. A regularization parameter is also added to the cost function. The regularization parameter is used to balance the margin maximization and loss. The cost function with the added regularization is [14]:

$$\min_w \lambda \|w\|^2 + \sum_{i=1}^n (1 - y_i \langle x_i, w \rangle)_+$$

The next step is to take partial derivatives with respect to the weights to find the gradients [14].

This is done to update the weights.

$$\frac{\delta}{\delta w_k} \lambda \|w\|^2 = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k} (1 - y_i \langle x_i, w \rangle)_+ = \begin{cases} 0, & \text{if } y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

When the model correctly predicts the class of a data point, only the regularization parameter is used to update the gradients [14].

$$w = w - \alpha \cdot (2\lambda w)$$

If the model makes a mistake on the prediction of the class of our data point, both the loss and the regularization parameter are used to update the gradients [13].

$$w = w - \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

In this project, the SVM will be implemented with stochastic gradient descent (SGD) due to its versatility with datasets both big and small.

XGBOOST Extreme gradient boosting, or XGBoost for short, is a Machine Learning classification algorithm based on decision trees that uses an enhanced and more optimized version of the gradient boosting framework [26]. This optimization is done first by building sequential trees using parallelization, changing the way it iterates through the trees by doing a depth first search, and by optimizing the hardware by allocating internal buffers of each thread to store gradient statistics. The Algorithmic Enhancements of XGBoost are regularization to prevent overfitting, 'learning' the best missing value based on the training loss, effectively finding the optimal split point using the distributed weighted quantile sketch algorithm and using a cross validation method at each iteration [29].

NAIVE-BAYES The Naïve Baye's classifier determines the class a data point belongs to using the value of features given. The classifier is based on Baye's theorem which states that the probability of an event A given an event B is based on the probability of event A and B divided by the probability of event B [45]. Adjusting this formula for the classifier gives us the formula is

$$p(y_i|x_1, x_2, \dots, x_n) = \frac{p(x_1, x_2, \dots, x_n|y_i) \cdot p(y_i)}{p(x_1, x_2, \dots, x_n)}$$

Where y_i is a class and x_1, x_2, \dots, x_n is the number of observations. However, to overcome the potential issue of requiring large datasets to have an estimate on the probability distribution for every feature combination the Naïve Bayes algorithm assumes that every feature is independent

of one another [45]. This assumption removes the need of dividing the number of observations with class y_i and reduced the formula to:

$$p(x_1, x_2, \dots, x_n | y_i) = p(x_1 | y_i) \cdot p(x_2 | y_i) \cdot \dots \cdot p(x_n | y_i)$$

Since the features used are continuous, this MQP will use the Gaussian Naïve Bayes.

RANDOM FOREST CLASSIFIER The random forest classifier (RFC) is made up of a multitude of individual decision trees generated from different subsets of the dataset, which work together to reach a consensus [46]. The RFC does this by having each decision tree predict what class a data point falls into, and the RFC picks whatever class is chosen the most as its final prediction [46]. To perform well the RFC needs an indicator in the features so that models built using said features have some foundation for their prediction and are not randomly guessing [46]. An RFC also needs the predictions made by its individual trees to have as little correlation as possible with each other [46]. This low correlation happens by letting every single individual tree take a random sample from the dataset with replacement, resulting in different trees [46]. This process is known as bagging. An RFC also ensures low correlation by having each tree pick from a random subset of features, doing so allows for more diversification [46].

3.7 Grid Search

To optimize the performance of deep learning models, grid search, a tuning technique was used to select optimal values for various hyperparameters. A hyperparameter is a parameter that is modified by the user, rather than learned by the model itself. A grid search gets the best performance out of a model by using every combination of hyperparameters possible and selects the best combination based on a given evaluation metric, such as accuracy or F1 score for example.

3.7.1 Hyperparameters

The hyperparameters used are shown in tables 2 through 5 below:

Random Forest	N_estimators	Criterion	max_depth	<i>min_samples_leaf</i>	min_samples_split	class_weight	random_state
Audio	50, 100, 200	gini, entropy	5	--	2	balanced	1
Facial	50, 100, 200	gini, entropy	5	--	2	balanced	1
GPS	10, 50, 100, 200, 500	gini, entropy	3	3	2	balanced	1

Table 2: Hyperparameters used for random forest classifier

Support Vector Machine	loss	penalty	Max_iter	tol	Class_weight	Random-state	Early_stopping
Audio	log	l2	10000	1e-4	balanced	1	True
Facial	log	l2	10000	1e-4	balanced	1	True
GPS	log	l1, l2	10000	1e-4	balanced	1	True

Table 3: Hyperparameters used for Support Vector Machine

XGBoost	n_estimators	Learning_rate	Use_label_encoder	Max_depth	Min_child_weight	Random_state	Gpu_id	Tree_method	gamma	subsample	Colsample_bytree
Audio	50, 100, 200	0.0001, 0.001, 0.01	False	5	1	1	0	gpu_hist	--	--	--
Facial	50, 100, 200	0.0001, 0.001, 0.01	False	5	1	1	0	gpu_hist	--	--	--
GPS	10,	0.0001,	--	2	5	--	--	--	5	0.8	0.8

	50, 100, 200	0.001 ,0.01									
--	--------------------	----------------	--	--	--	--	--	--	--	--	--

Table 4: Hyperparameters used for XGBoost

Naïve Bayes	Var_smoothing
Audio	1e-9
Facial	1e-9
GPS	1e-9

Table 5: Hyperparameters used for Naïve Bayes

4 Implementation

This study utilized the Scikit learn package version 1.0.1 in Python version 3.10.0 to classify features and calculate the evaluation metrics.

5 Results

Resultant values of various evaluation metrics for the machine learning models for 2-bin (0-10, 11-21) classification for all three modalities are shown in tables 5 through 8 below.

AUC-ROC	Naïve Bayes	RFC	XGBoost	SVM
Facial	0.517	0.557	0.517	0.433
Audio	0.438	0.436	0.419	0.460
GPS (before experiment)	0.3333	0.0	0.0	0.0
GPS (after experiment)	0.8571	0.7143	0.7143	0.4286

Table 5: AUC-ROC for 2 bin classification

Accuracy	Naïve Bayes	RFC	XGBoost	SVM
Facial	0.319	0.553	0.766	0.617
Audio	0.3617	0.426	0.745	0.468
GPS (before experiment)	0.1	0.4	0.9	0.0
GPS (after experiment)	0.25	0.75	0.875	0.875

Table 6: Accuracy for 2 bin classification

F1-score	Naïve Bayes	RFC	XGBoost	SVM
Facial	0.385	0.40	0.154	0.182
Audio	0.25	0.308	0.0	0.286
GPS (before experiment)	0.1818	0.0	0.0	0.0

GPS (after experiment)	0.25	0.0	0.0	0.0
------------------------	-------------	-----	-----	-----

Table 7: F1-score for 2 bin classification

Confusion Matrix	Naïve Bayes	RFC	XGBoost	SVM
Facial	[5 30] [2 10]	[19 16] [5 7]	[35 0] [11 1]	[27 8] [10 2]
Audio	[12 23] [7 5]	[14 21] [6 6]	[35 0] [12 0]	[12 18] [7 5]
GPS (before experiment)	[0 9] [0 1]	[4 5] [1 0]	[9 0] [1 0]	[0 9] [1 0]
GPS (after experiment)	[1 6] [0 1]	[6 1] [1 0]	[7 0] [1 0]	[7 0] [1 0]

Table 8: Confusion matrix for 2 bin classification

The highest AUC-ROC for facial features was 0.557, achieved using the random forest classifier, while Support Vector Machines (SVM) was the best classifier type for audio features achieving 0.460. Naïve Bayes had the best AUC-ROC for GPS features before and after the experiment with 0.3333 and 0.8571 respectively. This drastic difference in scores could be attributed to the smaller dataset used in GPS features after the experiment. With a smaller dataset, it could be easier to get a higher score since the model does not have to discriminate between as often as it would in a larger dataset. For accuracy metric, XGBoost was the best machine learning classifier for all three modalities, achieving 0.766, 0.745, 0.9, 0.875 respectively. It's worth mentioning that the Support Vector Machine also had an accuracy of 0.875 for GPS features after the experiment as well. Random forest classifier had the highest F1-score for facial and audio features with scores of 0.40 and 0.308 respectively. Naïve Bayes on the other hand had the best F1-score for GPS features taken before and after the experiment, with scores of 0.1818 and 0.25 respectively. As for the confusion matrix, the XGBoost had the best predictions for the facial features, audio features and

GPS features taken before the experiment while the Support Vector Machine and XGBoost performed the best for GPS features taken after the experiment.

AUC-ROC	Naïve Bayes	RFC	XGBoost	SVM
Facial	0.466	0.543	0.506	0.575
Audio	0.518	0.521	0.486	0.552
GPS (before experiment)	0.1905	0.0952	0.1429	0.1429
GPS (after experiment)	0.5333	0.9333	0.9333	0.8

Table 10: AUC-ROC for 3 bin classification

Accuracy	Naïve Bayes	RFC	XGBoost	SVM
Facial	0.170	0.340	0.638	0.553
Audio	0.340	0.277	0.596	0.596
GPS (before experiment)	0.0	0.3	0.7	0.7
GPS (after experiment)	0.625	0.625	0.625	0.625

Table 11: Accuracy for 3 bin classification

F1-score	Naïve Bayes	RFC	XGBoost	SVM
Facial	0.138	0.251	0.390	0.306
Audio	0.293	0.268	0.249	0.294
GPS (before experiment)	0.0	0.1667	0.4118	0.4118

GPS (after experiment)	0.3846	0.4722	0.3846	0.3846
------------------------	--------	---------------	--------	--------

Table 12: F1 score for 3 bin classification

Confusion Matrix	Naïve Bayes	RFC	XGBoost	SVM
Facial	[3 5 20] [0 0 12] [1 1 5]	[13 5 10] [2 0 10] [4 0 3]	[28 0 0] [11 1 0] [0 6 0]	[24 4 0] [10 2 0] [7 0 0]
Audio	[10 1 17] [3 1 8] [2 0 5]	[7 7 14] [2 3 7] [3 1 3]	[28 0 0] [12 0 0] [7 0 0]	[27 1 0] [11 1 0] [6 1 0]
GPS (before experiment)	[0 0 7] [0 0 3] [0 0 0]	[3 2 2] [2 0 1] [0 0 0]	[7 0 0] [3 0 0] [0 0 0]	[7 0 0] [3 0 0] [0 0 0]
GPS (after experiment)	[5 0 0] [3 0 0] [0 0 0]	[3 1 1] [0 2 1] [0 0 0]	[5 0 0] [3 0 0] [0 0 0]	[5 0 0] [3 0 0] [0 0 0]

Table 13: Confusion matrix for 3 bin classification

The evaluation metrics for the machine learning models for 3-bin (0-7, 8-14, 15-21) classification for all three modalities are shown above in tables 10 through 13. The Support Vector Machine classifier had the highest AUC-ROC for facial and audio features with a score of 0.575 and 0.552 respectively. Naïve Bayes had the best AUC-ROC score for GPS features taken before the experiment at 0.1952, while random forest and XGBoost were tied at 0.9333 for the best score for GPS features taken after the experiment. The best machine learning classifier in terms of accuracy for the facial features was the XGBoost at 0.638 while both the Support Vector Machine classifier and XGBoost had the best score for audio features as well as GPS features taken before the experiment at 0.596 and 0.7 respectively. As for the GPS features extracted from data gathered

after the experiment, all four machine learning models were equally accurate with a score of 0.625. When it came to the F1-score, XGBoost was the best machine learning classifier type for facial features, achieving a score of 0.390. On the other hand, SVM was the best classifier type for audio features with an F1-Score of 0.294. Both of the aforementioned classifiers were tied at 0.4118 for the best F1-score for GPS features taken before the experiment while the random forest classifier had the best F1-score for GPS features taken before the experiment at 0.4722. As for the confusion matrix the XGBoost had the best output for facial, audio, GPS both before and after the experiment. It is worth mentioning that SVM performed as well as XGBoost for GPS features before and after the experiment.

6 Discussion

Best classifier type Overall, it appears that XGBoost was the best at detecting depression based on the results of the evaluation metrics. The classifier had the highest score 17 times, compared to the Support Vector Machine at 10 times, random forest classifier at 6, and Naïve Bayes at 5. Most of the success with XGBoost might be attributed to using more options per parameter when performing a grid search as well. Conversely the lack of parameters to grid search with might've been the reason why the Naïve Bayes and random Forest classifiers produced such lackluster results.

Best data modality for depression classification The GPS features performed the best out of the modalities, the GPS metrics for predicting PHQ scores after the experiment were the highest 4 times, while the metrics before the experiment and the facial features were the highest twice. While the GPS features were the most promising, they were the smallest dataset by far compared to the audio and facial modalities. It is also worth noting that the number of instances in the GPS features for PHQ scores recorded after the experiment were less than before the experiment. It is possible that if more instances were available for the GPS modality the evaluation metrics would not be as high.

6.2 Limitations

Limitations to the study include the size of the datasets. While DAIC-WOZ and StudentLife dataset contained thousands of instances of data in their respective modalities, they did not have many participants in their studies. DAIC-WOZ had 192 participants while StudentLife had only 48 participants, so when it came to detect depression using their PHQ scores, the machine learning models did not have adequate data to work with. This can be seen in the confusion matrix, where even the best performing classifiers were not able to accurately predict the PHQ score of patients

who had higher scores compared to patients who had mild or even moderate scores. If the machine learning models had a bigger dataset they could have performed better when they received higher PHQ scores. To add to the relatively small datasets, the lack of features for the GPS modality poses as another limitation to this study. Due to time constraints, only 4 features could be extracted from the GPS modality, compared to the 5 features in the facial modality and 12 features that make up the audio modality. While it may seem that the number of facial features is only one more than the number of GPS features, each facial feature consists of more elements than the GPS features do. For example, the location variance at a given time is only one number, while the 3D points on the face consisted of 68 numbers.

In addition to the GPS features consisted of fewer elements and were recorded less frequently than the audio and facial features were. The GPS features from the StudentLife dataset records raw data every 20 minutes while the facial and audio features from the DAIC-WOZ dataset were recorded every 10 milliseconds. Recording data more often gives the machine learning algorithms a wider range of numbers to work with, allowing for more accurate predictions. More features could have given more comprehensive results, would make up for the lack of data that was recorded in the StudentLife dataset. Another limitation was the hyperparameter selection during the grid search. As mentioned earlier the more hyperparameters used the better the machine learning model performed, a thorough grid search looking at even more hyperparameters could have possibly given a more optimized machine learning model with an even better performance.

The most significant limitation for this study was that it was conducted by only one person, which ultimately limited the amount of work accomplished. While I did receive assistance, having an

additional three to four group members also working on the study full time would have helped fix or at least mitigate all the other limitations previously mentioned. Having more group members would have enable more features to be extracted, more classification models explored, more experiments conducted and more time finetuning the hyperparameters in the grid search, which could possibly lead to better evaluation metrics and a more conclusive result.

7 Conclusion

In summary, Depression is one of the most prevalent mental disorders in the world. The goal of this MQP was to use data from facial, audio and GPS modalities to track trajectories of depression using machine learning models. This was done by extracting GPS features from the StudentLife dataset and audio and facial features from the DAIC-WOZ dataset. Using patient health questionnaire scores as the ground truth, the features were trained using XGBoost, random forest classifier, Support Vector Machines, and the Naïve Bayes classifier. The XGBoost classifier performed the best out of all the machine learning algorithms used, with an accuracy of 0.82 for 2 bin classification and 0.639 for 3 bin classification. Much of this can be attributed to using a grid search to get the best performance possible. Overall, this study is a good baseline in comparing the performance of machine learning algorithms and what modalities and features were most useful for detecting depression.

7.1 Future Work

Future work in this study can expand on the features, data modalities, and classifiers used. More GPS features have been used to track trajectories of depression in the past with promising results. For example, Canzian *et al* had a mean absolute correlation of 0.432 and an average p-value of 0.068 when attempting to find a relationship between mobility metrics gathered from GPS data and depressive moods [9]. Using the current features in addition to other GPS features such as entropy, raw entropy, normalized entropy, percent of time at home, the routine index, the radius of gyration, the number of significant places visited, and more could have given the classifiers more data to use and predict depression with. Using other algorithms or perhaps building on existing classifiers could also be helpful in tracking trajectories of depression. Future studies could

further implement other advanced algorithms such as Neural Networks, which prior work has found to produce more accurate results than traditional machine learning when adequate data is available. This could also be done to existing classifiers such as XGBoost as well. Lastly, other modalities could also be used to detect depression such as the text modality. Jacob *et al* used CNN and LSTM to detect depression using features from the text modality gathered from twitter and had an accuracy of 99.46% [17]. These promising results show that these features can be an indicator of depression and including them could provide even more insight into trajectories of depression. Future studies can also go beyond focusing on what individual features are best at detecting depression and instead look at what combinations of features are optimal for detecting depression. For instance, audio/facial or GPS/facial features could be combined.

Bibliography

- [1] *67+ revealing statistics about smartphone usage in 2021*. TechJury. (2021, October 2). Retrieved October 14, 2021, from <https://techjury.net/blog/smartphone-usage-statistics/#gref>.
- [2] Alghowinem, Sharifa & Goecke, Roland & Wagner, Michael & Parker, Gordon & Breakspear, Michael. (2013). Eye movement analysis for depression detection. 2013 IEEE International Conference on Image Processing, ICIP 2013 - Proceedings. 10.1109/ICIP.2013.6738869.
- [3] Alghowinem, Sharifa & Goecke, Roland & Wagner, Michael & Parker, Gordon & Breakspear, Michael. (2013). Head Pose and Movement Analysis as an Indicator of Depression. Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013. 10.1109/ACII.2013.53.
- [4] Alku, P., Bäckström, T., & Vilkmán, E. (2002). Normalized amplitude quotient for parametrization of the glottal flow. *The Journal of the Acoustical Society of America*, 112(2), 701–710. <https://doi.org/10.1121/1.1490365>
- [5] Alsop, T. (2020, April 22). *Computer ownership in the US 2019*. Statista. <https://www.statista.com/statistics/756054/united-states-adults-desktop-laptop-ownership/>. Retrieved October 14, 2021.
- [6] American Foundation for Suicide Prevention. (2021, September 9). *Suicide statistics*. American Foundation for Suicide Prevention. Retrieved October 14, 2021, from <https://afsp.org/suicide-statistics/>.
- [7] Anderson, M. (2020, May 30). *American demographics of digital device ownership*. Pew Research Center: Internet, Science & Tech. Retrieved October 14, 2021, from <https://www.pewresearch.org/internet/2015/10/29/the-demographics-of-device-ownership/>.

- [8] Canzian, L., & Musolesi, M. (2015). Trajectories of depression. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*.
<https://doi.org/10.1145/2750858.2805845>
- [9] Chicco, D., Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
<https://doi.org/10.1186/s12864-019-6413-7>
- [10] Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., Zhou, F., & De la Torre, F. (2009). Detecting depression from facial actions and vocal prosody. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. <https://doi.org/10.1109/acii.2009.5349358>
- [11] Dauria, E. (2019, December 8). *Accuracy, Recall & Precision*. Medium. Retrieved January 11, 2022, from <https://medium.com/@erika.dauria/accuracy-recall-precision-80a5b6cbd28d>
- [12] Dogrucu, A., Perucic, A., Ball, D., & Isaro, A. (2018). Sensing Depression. : Worcester Polytechnic Institute.
- [13] Dogrucu, A., Perucic, A., Isaro, A., Ball, D., Toto, E., Rundensteiner, E. A., Agu, E., Davis-Martin, R., & Boudreaux, E. (2020, April 18). *Moodable: On feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data*. Smart Health. Retrieved October 14, 2021, from <https://www.sciencedirect.com/science/article/pii/S2352648319300273>.
- [14] Gandhi, R. (2018, July 5). *Support Vector Machine - introduction to machine learning algorithms*. Medium. Retrieved January 8, 2022, from

<https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>

- [15] Guo, W., Yang, H., Liu, Z., Xu, Y., & Hu, B. (2021). Deep Neural Networks for depression recognition based on 2D and 3D facial expressions under emotional stimulus tasks. *Frontiers in Neuroscience*, 15. <https://doi.org/10.3389/fnins.2021.609760>
- [16] Gerych, W., Agu, E., & Rundensteiner, E. (2019). Classifying depression in imbalanced datasets using an autoencoder- based anomaly detection approach. *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*. <https://doi.org/10.1109/icosc.2019.8665535>
- [17] Jacob, T., & Kapnadak, I. (n.d.). Detecting depression through tweets. Retrieved March 30, 2022, from <https://ishankapnadak.github.io/Detecting-Depression-Through-Tweets/Final-Report.pdf>
- [18] Hatton, C. M., Paton, L. W., McMillan, D., Cussens, J., Gilbody, S., & Tiffin, P. A. (2019). Predicting persistent depressive symptoms in older adults: A machine learning approach to personalised mental healthcare. *Journal of Affective Disorders*, 246, 857–860. <https://doi.org/10.1016/j.jad.2018.12.095>
- [19] Mallick, S. (2016, December 6). *Histogram of oriented gradients explained using opencv*. LearnOpenCV. Retrieved December 26, 2021, from <https://learnopencv.com/histogram-of-oriented-gradients/>
- [20] Jayaswal, V. (2020, September 14). *Performance metrics: Confusion matrix, precision, recall, and F1 score*. Medium. Retrieved January 11, 2022, from

<https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262>

- [21] Kayastha, R., Caouette, H., Melican, V., Hernandez-Reisch, M., & Bruneau, C. (2021). *Machine Learning for Mental Health Screening*. : Worcester Polytechnic Institute.
- [22] Korstanje, J. (2021, August 31). *The F1 score*. Medium. Retrieved January 11, 2022, from <https://towardsdatascience.com/the-f1-score-bec2bbc38aa6>
- [23] Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9), 606–613. <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>
- [24] *Living well with major depressive disorder*. SAMHSA. (n.d.). Retrieved March 11, 2022, from <https://www.samhsa.gov/serious-mental-illness/major-depression>
- [25] Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9), 1315–1316. <https://doi.org/10.1097/jto.0b013e3181ec173d>
- [26] Maridaki, A. (2018, June 1). *Categorical assessment of depression based on low level features*. *Apothesis Αρχική*. Retrieved December 26, 2021, from <https://apothesis.lib.hmu.gr/handle/20.500.12688/8757>
- [27] Matteo, Daniel & Fotinos, Kathryn & Lokuge, Sachintha & Yu, Julia & Sternat, Tia & Katzman, Martin & Rose, Jonathan. (2020). The Relationship Between Smartphone-Recorded Environmental Audio and Symptomatology of Anxiety and Depression: Exploratory Study. *JMIR Formative Research*. 4. e18751. 10.2196/18751.

- [28] Mayo Foundation for Medical Education and Research. (2018, February 3). Depression (major depressive disorder). Mayo Clinic. Retrieved March 15, 2022, from <https://www.mayoclinic.org/diseases-conditions/depression/symptoms-causes/syc-20356007>
- [29] Morde, V. (2019, April 8). *XGBoost algorithm: Long may she reign!* Medium. Retrieved January 8, 2022, from <https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d>
- [30] Murphy-Chutorian, Erik and Trivedi, Mohan Manubhai, "Head Pose Estimation in Computer Vision: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 4, pp. 607-626, April 2009, doi: 10.1109/TPAMI.2008.106.
- [31] Narkhede, S. (2018, June 26). *Understanding AUC - roc curve*. Medium. Retrieved January 11, 2022, from <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
- [32] *Number of smartphone users in the U.S. 2025*. Statista. (2021, March 19). Retrieved October 14, 2021, from <https://www.statista.com/statistics/201182/forecast-of-smartphone-users-in-the-us/>.
- [33] Ray, A., Kumar, S., Reddy, R., Mukherjee, P., & Garg, R. (2019). Multi-level attention network using text, audio and video for Depression prediction. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop - AVEC '19*. <https://doi.org/10.1145/3347320.3357697>
- [34] Resom, A., Assan, J., Wu, Y., Gao, Y., & Flannery, M. (2019). *Machine Learning for Mental Health Detection*. : Worcester Polytechnic Institute.

- [35] Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., & Pantic, M. (2017). *Avec 2017. Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. <https://doi.org/10.1145/3133944.3133953>
- [36] Sensors and cellphones - stanford university. (n.d.). Retrieved March 30, 2022, from <https://web.stanford.edu/class/cs75n/Sensors.pdf>
- [37] Taye, Y., Pingal, N., Seifu, Y., Caltabiano, J., Thant, M., & Sargent, A. (2020). *Mental Health Sensing Using Machine Learning*. : Worcester Polytechnic Institute.
- [38] Ting K.M. (2011) Confusion Matrix. In: Sammut C., Webb G.I. (eds) *Encyclopedia of Machine Learning*. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_157
- [39] Y. . -I. Tian, T. Kanade and J. F. Cohn, "Recognizing action units for facial expression analysis," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97-115, Feb. 2001, doi: 10.1109/34.908962.
- [40] U.S. Department of Health and Human Services. (n.d.). *NIMH "depression*. National Institute of Mental Health. Retrieved October 14, 2021, from <https://www.nimh.nih.gov/health/topics/depression>.
- [41] Vogels, E. A. (2020, August 14). *About one-in-five Americans use a smart watch or fitness tracker*. Pew Research Center. Retrieved October 14, 2021, from <https://www.pewresearch.org/fact-tank/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/>.

- [42] Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014). StudentLife. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*.
<https://doi.org/10.1145/2632048.2632054>
- [43] Wu, Y., & Ji, Q. (2018). Facial Landmark Detection: A Literature Survey. *International Journal of Computer Vision*, 127(2), 115–142. <https://doi.org/10.1007/s11263-018-1097-z>
- [44] Yalamanchili, B., Kota, N. S., Abbaraju, M. S., Nadella, V. S., & Alluri, S. V. (2020). Real-time acoustic based depression detection using machine learning techniques. *2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE)*. <https://doi.org/10.1109/ic-etite47903.2020.394>
- [45] Yildirim, S. (2020, May 12). *Naive Bayes classifier-explained*. Medium. Retrieved January 8, 2022, from <https://towardsdatascience.com/naive-bayes-classifier-explained-50f9723571ed>
- [46] Yiu, T. (2019, June 12). *Understanding Random Forest*. Medium. Retrieved January 8, 2022, from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>