B.Comp. Dissertation

Depression Detection from Speech

By

Chan Jun Wei

A0112084U

Department of Computer Science

School of Computing

National University of Singapore

2016/17

B.Comp. Dissertation

Depression Detection from Speech

By

Chan Jun Wei

A0112084U

Department of Computer Science

School of Computing

National University of Singapore

2016/17

Project ID: H0201250

Advisor: Assoc Prof Ooi Wei Tsang

Deliverables:

Report: 1 Volume

Abstract

The lack of objective measures causes the most treatable metal illness, depression to be often under-diagnosed. Recent studies have shown that speech is a good indicator of depression, giving us a motivation to perform depression diagnosis using speech to create an objective measure. This project studies the use of state-of-the-art machine learning (ML) models including ensemble in predicting depression severity using audio features after optimizing the data. We obtain the audio data from Audio/Visual Emotion Challenge and Workshop 2016 (AVEC 2016) and aim to have a mean F1 of 0.8 on the development (dev) set. Our work has successfully shown that AdaBoost (AB) trained using the mean of Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, Mel Frequency Cepstral Coefficients (MFCCs) and Chroma Deviation is a good model for depression prediction, which is able to predict Personal Health Questionnaire eight-item depression scale (PHQ-8) at mean F1 of 0.82 and Root Mean Square Error (RMSE) of 6.43. The results are better than other state-of-the-art models including the baselines at mean F1 of 0.5 and RMSE of 6.74. It also gives a mean F1 of 1 in multi-class classification, which predicts the depression level of individuals. In the future, we aim to further verify the model correctness and create an autonomous agent that could help the depressed patients.

Subject Descriptors:

H3.3 Information Search and Retrieval

I2 Artificial Intelligence

J3 Life and Medical Sciences

J4 Social and Behavioral Sciences

J5 Arts and Humanities

Keywords:

Artificial Intelligence, Multimedia Systems

Implementation Software and Hardware:

Java 1.7, Python 2.7

Acknowledgement

I wish to express my sincere thanks to my advisor, Professor Ooi Wei Tsang, for providing me with all the professional and valuable guidance which is the key to my success. I am also grateful to Professor Bryan Kian Hsiang Low and Dr. Chua Tat-Seng from School of Computing for their guidance and sharing on their expertise. I take this opportunity to express gratitude to all faculty members of the Department for their help and support. I am extremely thankful to my CS2108 group mates and CS4246 group mates for agreeing to let me extend the work on depression studies. Also, I would like to thank the organiser of the Audio/Visual Emotion Challenge and Workshop 2016 (AVEC 2016) and the Audio/Visual Emotion Challenge and Workshop 2014 (AVEC 2014) for providing the depression corpus for us. Special thanks to the special ones, especially my family and friends, for their unceasing encouragement, support and attention. I also place on record, my gratitude to one and all, who directly or indirectly, have lent a hand in this venture.

List of Figures

Figure 5.1 An Overview Workflow of the Experiment. 32
Figure 5.2: Face-to-face interview setup (Left); Ellie, the virtual interviewer (Right) (Gratch et al., 2014). 33
Figure 5.3 PHQ8 Score Distribution in Train Set (Left); PHQ8 Score Distribution in Dev Set (Right) 34
Figure 5.4 Depression level distribution in Train Set (Left); Depression level distribution in Dev Set (Right) 34
Figure 5.5 The First 0.2 seconds of the Non-Depressed Audio Signal (Left); The First 0.2 seconds of the Depressed Audio Signal (Right) 35
Figure 5.6 ML Model Selection Workflow 39
Figure 6.1 Illustration of Different Normalization Methods. 43
Figure 6.2 The Regressor Performances based on RMSE where the solid line is the prediction RMSE baseline 51
Figure 6.3 The performance of all OVR or OVO models trained by chosen Audio Features. 52
Figure 0.1: The questions in PHQ-8 (Kroenke et al., 2009). 63
Figure 0.2: Distribution of depressive symptom severity and depressive disorders in BRFSS (Kroenke et al., 2009). 63
Figure 0.3 Algorithm Used to Extract MFCCs. 64

List of Tables

Table 2.1 Baseline Audio Features (Valstar et al., 2016). 4
Table 2.2 Baseline result for depression classification. Performance is measured in F1 score (Valstar et al., 2016)....4
Table 2.3 Baseline result for depression regression. Performance is measured in mean absolute (Valstar et al., 2016). 5
Table 2.4 Performance of Mel-scale filter bank features with different parameters for DepAudioNet (values of non-depression in brackets) (Ma et al., 2016). 5
Table 2.5 MITLL Audio Features (Williamson et al., 2016)....6
Table 2.6 Performance of the classifier proposed by MITLL compared to the Baseline. 6
Table 2.7 Classifier performance proposed by Pampouchidou compared to the Baseline. 7
Table 2.8 Performance of the classifier proposed by SCUBA compared to the Baseline.4 7
Table 2.9 Performance of the regressor proposed by SCUBA compared to the Baseline....8
Table 3.1 Audio Features implemented by pyAudioAnalysis (Giannakopoulos, 2015). 14
Table 3.2 Standard kernels used in this project. 23
Table 3.3 The definition of ML techniques that are used commonly in this project in both classification and regression task. 26
Table 3.4 Confusion Matrix for Depression Classification. 26
Table 5.1 Different types of interviews to collect audio and video recordings (Gratch et al., 2014). 33
Table 5.2 Sample of the transcripts for the audio recordings provided in DAIC-WOZ. 36
Table 6.1 Overview of experiments, whereby the term “all combinations” stated under audio features implies the use of Audio Feature Selection via Complete Search technique, thus it is not included in the feature selection column. Besides, the term “normalized” represents Audio Feature Standardization. 40
Table 6.2 Overview of tasks in this study. 41
Table 6.3 Result with features (MFCCs) for every audio segments in AY2016/17 Semester 1. The value which is not in the bracket is the value for depressed class and the value which is in the bracket is the value for not depressed class. The best result is bolded. 41
Table 6.4 Result produced using classifiers with features (MFCCs) with one feature row per participant in AY2016/17 Semester 1. The value which is not in the bracket is the value for depressed class and the value which is in the bracket is the value for not depressed class. The best result is bolded. 42

Table 6.5 The best 3 model performance based on mean F1 trained by 9 different combinations of mean audio features. (Partial, see full on Appendix E) 42
Table 6.6 The best 3 model performances based on mean F1 trained by mean and std of audio features. (Partial, see full on Appendix E).....43
Table 6.7 The Performances of AB Trained by Data Normalized by Different Methods.....44
Table 6.8 Effect of Audio Feature Standardization on individual Audio Feature based on AB performance. 44
Table 6.9 Effect of Audio Feature Standardization on individual Audio Feature based on performance of different classifiers. 45
Table 6.10 10 Audio Features Combinations include the best 8 classifiers performances, the combination of all positively correlated audio features and the combination without any negatively correlated audio features. Those combinations which contain audio features that would have negative effect after Audio Feature Standardization are not included. (Partial, see full on Appendix E) 45
Table 6.11 Effect of Feature Selection Techniques on different classifiers using Train Set where “All” means no feature selection method is applied. 46
Table 6.12 Inconsistent performance of AB trained by subset of train data extracted using Relief algorithms in different iterations. 46
Table 6.13 Effect of Feature Selection Techniques except Relief applied to the mean and std of audio features based on the performance of classifiers. (Partial, see full on Appendix E) .47
Table 6.14 Effect of Feature Selection Techniques except Relief applied to the normalized mean and std of audio features based on the performance of classifiers. (Partial, see full on Appendix E).....47
Table 6.15 Combination of non-normalized and normalized audio features.....48
Table 6.16 Combination of mean audio features and std audio features. 48
Table 6.17 Effect of Audio Feature Standardization on Mean Audio Features. (Partial, see full in Appendix E).....49
Table 6.18 Summary of all the important classification results.....49
Table 6.19 The performance of Regressor trained by suggested Audio Feature Subsets.....50
Table 6.20 The Performance of Regressors trained by the suggested Audio Feature Subset where the bolded values are the best result and the underlined results are the regressors which failed the baseline. 50
Table 6.21 Performances of Multi-class classifiers trained by suggested Audio Feature Set. 52

Table 6.22 Comparison of Depression Binary Classification Result with Related Studies using audio data from development set. The proposed method is bolded. 53
Table 6.23 Comparison of Depression Binary Classification Result with Related Studies on development set. The proposed method is bolded. 54
Table 6.24 Comparison of Depression Regression Result with Related Studies on development set. The proposed method is bolded and the lower RMSE and MAE score are underlined. 54
Table 6.25 Task summary of this paper. 54
Table 0.1: The representation of the PHQ-8 score (Kroenke et al., 2009). 62
Table 0.2 Classifiers Performances trained by mean MFCCs. The value which is not in the bracket is the value for depressed class and the value which is in the bracket is the value for not depressed class. The best result is bolded. 67

List of Equations

Equation 3-1 The definition of Zero Crossing Rate Zj where S is the number of samples in the frame and $sgnxi$ is the sign function defined in Equation 3-2 which will determine if the number is positive, negative or 0. 10
Equation 3-2 The definition of the sign function $sgnx$ which will determine if the number is positive, negative or 0. 11
Equation 3-3 Energy of Signal Processing Es(s) for signal (s) of discrete time (t)..... 11
Equation 3-4 Conversion of Energy of Signal Processing using magnitude Z to Energy in physics E. 11
Equation 3-5 Entropy H(x) of a random variable X. 11
Equation 3-6 Spectral Centroid fc is given in frequency f(i) * Magnitude M(i), with the Magnitude M(i) as the weight (Schubert & Wolfe, 2006). 12
Equation 3-7 The nth central moment μn of a discrete real-valued random variable X where E is the expectation operator. 12
Equation 3-8 The 2nd central moment $\mu 2$ or the (variance Var[X]) of a real-valued random variable X where E is the expectation operator. 12
Equation 3-9 The definition of Spectral Spectrum fs where M(i) is the magnitude of the signal, f(i) is the frequency, and fc is the Spectral Centroid (Bello, 2013)..... 12
Equation 3-10 The definition of Spectral Flux ft where Nt and $Nt-1$ are normalized magnitudes of time frame t and t – 1 respectively. 12
Equation 3-11 The use of Spectral Rolloff $RC(j)$ to concentrate in c% of the magnitude distribution Mj of frame j with S number of samples..... 13
Equation 3-12 The std σ of a finite data set $x1, x2, \dots, xn$ where μ is the mean 13
Equation 3-13 The Formula of Fisher Score where c is the number of classes, fi is the i-th feature, nj is the number of samples in class j, μi is the mean value of fi, $\mu i, j$ is the mean value of fi in class j and $\sigma i, j2$ is the variance of fi in class j (J. Li et al., 2016). 16
Equation 3-14 Relief Score which is measured using k randomly selected samples from n samples where fi is the i-th feature, NM(j) and NH(j) denotes the nearest sample from same class and other class respectively, d(.) is a distance function which is usually defined as Euclidean distance (J. Li et al., 2016). 16
Equation 3-15 The definition of Information Gain where S is the samples, A is the feature, Value(A) is possible values for A, Sv is the subset of S which has value v in A (Sugumaran et al., 2007) and Entropy(.) is defined in Equation 3-5. 17

Equation 3-16 Generalized version of information gain where X and Y are random variables where Entropy(.) is defined in Equation 3-5. 17
Equation 3-17 Feature Score based on MIFS where fi is the i-th feature, S is the selected feature set which is initially empty, β is an empirical parameter that is set to 1 and I(.) is defined in Equation 3-16 (J. Li et al., 2016). 17
Equation 3-18 Feature Score based on CIFE where fi is the i-th feature, S is the selected feature set which is initially empty and I(.) is defined in Equation 3-16. 18
Equation 3-19 Feature Score based on CMIM where fi is the i-th feature, S is the selected feature set which is initially empty and I(.) is defined in Equation 3-16 (J. Li et al., 2016). ..18
Equation 3-20 The formula for symmetric uncertainty (SU) where fS is the feature set, information gain I(.) defined in Equation 3-15 and Entropy H(.) defined in Equation 3-5 (J. Li et al., 2016). 19
Equation 3-21 The evaluation metric of CFS where S is the feature subset, k is the number of features, rcf is the average of feature-class correlation and rcf is mean feature-feature correlation (J. Li et al., 2016). 19
Equation 3-22 Formula used to calculate the score for Gini Index, provided that fi is the i-th feature with r different feature values where the dataset is split into samples set W and W using the j-th feature as the margin value. CS implies that s is the class label and p(.) means probability (J. Li et al., 2016). 19
Equation 3-23 The assumption made by Naïve Bayes where X is the feature vector and Xr is the value for each feature that forms the feature vector. 21
Equation 3-24 The classification method of Naïve Bayes (Kotsiantis et al., 2007). 21
Equation 3-25 Gaussian distribution property of the function vector f with the corresponding inputs X where μ is the mean vector, K is the covariance matrix (Rasmussen & Williams, 2006). 21
Equation 3-26 The definition of Multivariate Gaussian Distribution where f is the function vector, μ is the mean vector, K is the covariance matrix and D is the dimension of the function vector, mean vector and covariance matrix (Rasmussen & Williams, 2006). 21
Equation 3-27 The definition of covariance function or kernel (Snelson, 2007). 22
Equation 3-28 The decision rule for SVM or the line used to separate the 2 classes in SVM where w is the weight vector, b is the bias and $\xi i \geq 0$ is the slack variable used to account the cases of misclassification which implies that $i\xi i$is an upper bound of training error (Kotsiantis et al., 2007). 22

Equation 3-29 Algorithm used by SVM to optimize the margin where w is the weight vector and b is the bias (Kotsiantis et al., 2007). 22
Equation 3-30 The definition of Euclidean Distance (Kotsiantis et al., 2007). 23
Equation 3-31 The definition of Precision (Sokolova & Lapalme, 2009) where tp and fp are defined in Table 3.4. 27
Equation 3-32 The definition of Recall (Sokolova & Lapalme, 2009) where tp and fn are defined in Table 3.4. 27
Equation 3-33 The definition of Fscore (Sokolova & Lapalme, 2009) where β is a variable defined by the user, whereas tp, fp and fn are defined in Table 3.4. 28
Equation 3-34 The definition of F1, which is the case when the β of the Fscore (defined in Equation 3-33) is equal to 1, where Precision is defined in Equation 3-31 and Recall is defined in Equation 3-32, whereas tp, fp and fns are defined in Table 3.4. 28
Equation 3-35 The definition of accuracy (Sokolova & Lapalme, 2009) where tp, tn, fp and fn are defined in Table 3.4. 28
Equation 3-36 The definition of Root Mean Square Error (RMSE) (Chai & Draxler, 2014) where n is the number of samples and ei is the error of the model. 29
Equation 3-37 The definition of Mean Absolute Error (MAE) (Chai & Draxler, 2014) where n is the number of samples and ei is the error of the model. 29
Equation 0-1 Mel scale M(f) of a frequency f (Lyons, 2012). 64
Equation 0-2 The definition of L1-norm $|x|1$ (Sanjay, 2008). 65

List of Abbreviations

AB	AdaBoost
ANEW	Affective Norms for English Words ratings
AVEC 2016	Audio/Visual Emotion Challenge and Workshop
BDI-II	Beck Depression Inventory-II
BO	Bayesian optimization
CART	Classification and Regression Tree
CIFE	Conditional Infomax Feature Extraction
CMIM	Conditional Mutual Information Maximization
CNN	Convolutional Neural Network
csv	Comma-separated values
DAIC	Distress Analysis Interview Corpus
DAIC-WOZ	Distress Analysis Interview Corpus-Wizard of Oz
DCC	Depression Classification Sub-challenge
dMFCCs	delta MFCCs
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, fourth edition
DT	Decision Tree
F0	Fundamental Frequency
FAU	Facial Action Unit
FCBF	Fast Correlation Based Filter
fn	False Negatives
fp	False Positives
FYP	Final Year Project
GloVe	Global Vectors for word representation
GMM	Gaussian Mixture Model
GP	Gaussian Process
G-PDLA	Gaussian Probabilistic Linear Discriminant Analysis
H1H2	The difference of the first 2 harmonics of the differentiated glottal source spectrum
HMPDD0-12	Harmonic Model and Phase Distortion deviations
HMPDM0-24	Harmonic Model and Phase Distortion mean
HOG	Histogram of Oriented Gradients
ICAP	Interaction Capping
KNN	K-Nearest Neighbor
LMHI	Landmark Motion History Images
LMM	Landmark Motion Magnitude
LSTM	Long Short Term Memory
MAE	Mean Absolute Error
MCEP0-24	Mel cepstral coefficients
MDD	Major Depressive Disorder
MDQ	Maxima Dispersion Quotient
MFC	Mel Frequency Cepstral
MFCCs	Mel Frequency Cepstral Coefficients
MIFS	Mutual Information Feature Selection
MIM	Mutual Information Maximization

MITLL	MIT Lincoln Laboratory
ML	Machine Learning
MRMR	Minimum Redundancy Maximum Relevance
NAQ	Normalized Amplitude Quotient
NB	Naïve Bayes
OVO	One-Vs-One
OVR	One-Vs-The-Rest
PCA	Principle Component Analysis
peak-Slope	Spectral tilt/slope of wavelet responses
PHQ-8	Personal Health Questionnaire eight-item depression scale
PHQ-9	Patient Health Questionnaire nine-item depression scale
PSP	Parabolic Spectral Parameter
PTSD	Post-Traumatic Stress Disorder
QOQ	Quassi Open Quotient
RBF	Radial Basis Function
Rd	Shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics
RF	Random Forest
RMSE	Root Mean Square Error
SCID	Structured Clinical Interview for DSM-IV
SGD	Stochastic Gradient Descent
Std	Standard Deviation
SVM	Support Vector Machine
tn	True Negatives
tp	True Positives
UROP	Undergraduate Research Opportunity Project
US	United States
USC	University of Southern California
VT	Vocal Tract
VUV	Voicing
ZCA	Whitening

Table of Contents

Abstract ..... *ii*
Acknowledgement ..... *iii*
List of Figures ..... *iv*
List of Tables ..... *v*
List of Equations ..... *viii*
Chapter 1 Introduction *1*
1.1 Impact of Depression **1**
1.2 Difficulties in Depression Diagnosis **2**
1.3 Depression Diagnosis using Speech **2**
1.4 Study Background and Objective **3**
1.5 My Contributions **3**
Chapter 2 Related Studies *4*
2.1 AVEC2016 ..... **4**
2.1.1 Baseline 4
2.1.2 DepAudioNet 5
2.1.3 MIT Lincoln Laboratory (MITLL) 5
2.1.4 Pampouchidou's Study 6
2.1.5 SCUBA 7
2.2 Other Related Studies **8**
2.3 Research Gap **8**
2.4 Chapter Summary **9**
Chapter 3 Background Knowledge *10*
3.1 Audio Features **10**
3.1.1 Time-Domain Audio Features 10
3.1.2 Frequency-Domain Audio Features 11
3.1.3 Chroma-based Audio Features 13
3.1.4 Section Overview 14
3.2 Data Normalization Techniques **14**
3.2.1 Feature Warping 14
3.2.2 Min-max Scaling 15
3.2.3 Section Overview 15
3.3 Feature Selection Techniques **15**

3.3.1 Supervised Similarity Based 16
3.3.2 Supervised Informational Theoretical Based 17
3.3.3 Supervised Statistical Based 19
3.3.4 Section Overview 20
3.4 Machine Learning (ML) Models **20**
3.4.1 Probabilistic Machine Learning Model 20
3.4.2 Kernel-based Machine Learning Model 21
3.4.3 K-Nearest Neighbor (KNN) 23
3.4.4 Decision Tree 23
3.4.5 Ensemble 24
3.4.6 Multi-class Classification 24
3.4.7 Optimization 25
3.4.8 Section Overview 25
3.5 Evaluation Metrics **26**
3.5.1 Classification 26
3.5.2 Regression 28
3.6 Chapter Summary **29**
Chapter 4 Proposed Method ***30***
4.1 Audio Feature Standardization **30**
4.2 Audio Feature Selection via Complete Search **30**
4.3 Machine Learning Method **31**
4.4 Multi-class Classification Task **31**
4.5 Chapter Summary **31**
Chapter 5 Methodology ***32***
5.1 Selecting and Understanding Dataset **33**
5.1.1 Data Distribution 33
5.1.2 Depressed and Non-Depressed Samples 34
5.1.3 Section Overview 35
5.2 Audio Pre-processing **35**
5.2.1 Noise Reduction 35
5.2.2 Obtaining Speech Segments 36
5.2.3 Section Overview 37
5.3 Audio Features Extraction **37**

5.4 Data Preparation **37**
5.4.1 Data Normalization 38
5.4.2 Feature Selection 38
5.5 Model Training **38**
5.6 Chapter Summary **39**
Chapter 6 Result and Discussion ***40***
6.1 Depression Binary Classification **41**
6.1.1 Combination of speech segments 41
6.1.2 Audio Features of pyAudioAnalysis 42
6.1.3 Inclusion of the Std of Audio Features 42
6.1.4 Normalization Techniques Application 43
6.1.5 Feature Selection Techniques Application 46
6.1.6 Applying Feature Selection to Normalized Data 47
6.1.7 Combination of Normalized and Not Normalized Audio Features 48
6.1.8 Combination of Section 6.1.3 and Section 6.1.5 48
6.1.9 Normalization of Mean Audio Features 49
6.1.10 Section Overview 49
6.2 Depression Regression **50**
6.2.1 Regression Results 50
6.2.2 The Performance of All Models 50
6.2.3 Section Overview 51
6.3 Multi-class Classification **51**
6.3.1 Results 51
6.3.2 OVO or OVR Models 52
6.3.3 Section Overview 52
6.4 Chapter Summary **52**
Chapter 7 Conclusion ***55***
7.1 Summary **55**
7.2 Limitation **55**
7.3 Future Work **55**
References ***56***
Appendix A : PHQ-8 ..... ***62***
Appendix B : MFCCs Implementation ***64***
Appendix C : Feature Selection Techniques provided in Scikit ***65***
Appendix D : The definition of Norm ***66***

Appendix E : N-Layer Ensemble *67*
Appendix F : Subsets of Audio Features in pyAudioAnalysis *68*

Chapter 1 Introduction

Depression is an ambiguous word in psychiatry, it can either be described as sadness or people who deny feeling sad (Stratou, Scherer, Gratch, & Morency, 2015). It might be misunderstood as a part of aging, but it is actually a temporary mental issue characterized by "sadness, loneliness, despair, low self-esteem and self-reproach" (Fiske, Wetherell, & Gatz, 2009).

Clinical depression usually means Major Depressive Disorder (MDD) or unipolar depression (Stratou et al., 2015), which is the third leading cause of disorders globally with 65 million lived with the disability or lost due to early death (World Health Organization, 2008). Besides MDD, there are a few other depressive disorders include dysthymia or minor depression (Kroenke et al., 2009). There are also other mental illnesses that always co-occur with depression, such as Post-Traumatic Stress Disorder (PTSD) (Stratou et al., 2015).

When the individuals are depressed, they would:

1. feel sad, hopeless and tearful (Stratou et al., 2015).
2. have significantly lesser interest or pleasure (Stratou et al., 2015).
3. have their thoughts and physical actions slowed down (Stratou et al., 2015).
4. always feel tired (Stratou et al., 2015).
5. have difficulties in concentrating and thinking (Stratou et al., 2015).
6. be more indecisive (Stratou et al., 2015).

1.1 Impact of Depression

Depression impacts individuals negatively in either short or long time period. The affected individuals range from youth (Morey, Arora, & Stark, 2015) to elderly (Fiske et al., 2009). Depression is one of the most significant causes of disease worldwide in 2004 (World Health Organization, 2008) and is predicted to be the second leading cause by 2020 (MacPherson et al., 2013). A depressed individual has higher possibility in early death, morbidity, self-neglect, decreased social and cognitive functioning, as well as suicide (Fiske et al., 2009). While suicide imposes high socio-economic costs onto communities and is listed as one of the top 10 causes of death, which ranked above chronic liver disease, Alzheimer's, homicide, arteriosclerosis or hypertension (Xu, Murphy, Kochanek, Bastian, & Statistics, 2016), approximately 50% of them are caused by depression (Cummins et al., 2015).

Besides, depression is one of the economic burdens in the world. It costs the United States (US) about $80 billion in medical expenditures, lost productivity and suicides

(Greenberg, Fournier, Sisitsky, Pike, & Kessler, 2015). More than $33 billion has been consumed by lost productivity because of symptoms that drain energy, affect work habits and cause issues with concentration, memory and decision-making (Greenberg et al., 2015). Similarly, the European Union also spent €92 billion on the issue of depression in 2010, with €54 billion lost because of decreased work productivity (Cummins et al., 2015).

1.2 Difficulties in Depression Diagnosis

Despite being one of the most under-diagnosed illnesses globally with severe consequences, depression is in fact one of the most treatable illnesses (Kessler et al., 2003). In general health-care, 48.4% of patients suffering from depression are unrecognized (Kessler et al., 2003).

Depression diagnosis becomes more difficult as the patients are always unwilling to admit feeling depressed by seeking help. In their opinions, being depressed implies that they are being weak and depression is a personality defect. This stigma causes people with mental problems to separate themselves from others. This phenomenon is known as social distancing (Smith & Cashwell, 2011). Hence, individuals tend to feel ashamed of being depressed. They deny being depressed and hide their feelings from others in order to live normally without being treated differently (Wolpert, 2001).

On the other hand, even if they look for help, the accuracy of the depression assessments such as the Personal Health Questionnaire eight-item depression scale (PHQ-8), a self-administered, 8-question diagnostic test for depressive disorders (Kroenke et al., 2009) explained in Appendix A, is affected by the Hawthorne Effect. Under the Hawthorne Effect, individuals would improve or change an aspect of their behavior if they are being observed (McCambridge, Witton, & Elbourne, 2014). Thus, depression diagnosis using assessments requires a lot of training and time to generate an acceptable result (Cummins et al., 2015). Lacking of an objective measure for depression is a big issue, so researchers try to construct it based on behavioral, biological and physiological signals (Cummins et al., 2015).

1.3 Depression Diagnosis using Speech

As speech is easy to obtain non-offensively with low-cost, many researchers explore the possibility of depression prediction through speech (Cummins et al., 2015). In the past decade, predicting emotion using speech has been a success (Johnstone, 2001). Corollary to that, the proof of speech bring useful in predicting clinical depression scores implies that speech is an important indicator for depression (Hashim, Wilkes, Salomon, Meggs, & France, 2016). There are also reports stating that the verbal behavior of a depressed patients should have "decreased

verbal activity productivity, a diminished prosody and monotonous, 'lifeless' sounding speech" (Cummins et al., 2015). It is also proven that “diminished, prosodic and monotonous speech” always has a strong co-relationship with depression (Cummins et al., 2015). This paper aims to boost the depression diagnostics accuracy by using speech as an objective measure.

1.4 Study Background and Objective

This study started as an Undergraduate Research Opportunity Project (UROP) focusing on recognizing emotions before extending to the Final Year Project (FYP) to work on depression diagnosis. The experience of working on emotional speech has greatly benefited us in depressed speech analysis. Recently, the organizer of the Depression Classification Sub-challenge (DCC) from the Audio/Visual Emotion Challenge and Workshop (AVEC 2016) provided us a depression analysis corpus with the baseline performances at mean F1 of 0.5 and Root Mean Square Error (RMSE) of 6.74 (Valstar et al., 2016). Therefore, with the assumption of depressed patients from distinct cultures have similar verbal behavior, we would investigate the applicability and feasibility of different audio features and machine learning (ML) models in predicting clinical ratings of depression severity based on PHQ-8 scale using the depression analysis corpus provided by AVEC 2016.

Among all the works done for AVEC2016, the current best model performance using only audio produces a mean F1 score of 0.73 (Nasir, Jati, Shivakumar, Chakravarthula, & Georgiou, 2016). Meanwhile, the optimal model performance using features which are not limited to audio produces a mean F1 score of 0.81 (Williamson et al., 2016). However, the performance of model trained using only speech might be improved if given a more optimal set of audio features, better normalization method, better feature selection technique and ML algorithm. Thus, the objective of this study is to boost the mean F1 to at least 0.8.

1.5 My Contributions

Using Audio Feature Selection via Complete Search, the study found an optimal set of audio features, including the mean of Zero-crossing rate, entropy of energy, spectral spread, spectral entropy, Mel Frequency Cepstral Coefficients (MFCCs) and chroma deviation. After comparing many ML algorithms, AdaBoost (AB) is proven to be the best performed model by providing a mean F1 of 0.82 and a Root Mean Square Error (RMSE) of 6.43. It also gives a mean F1 of 1 in multi-class classification which predicts the depression level of an individual. Meanwhile, Audio Feature Standardization is proposed as the normalization method, which is useful but should be used with care.

Chapter 2 Related Studies

This chapter discusses and reviews related studies on depression classification or regression using speech, focusing on extensive study on AVEC2016 and brief discussions on other studies.

2.1 AVEC2016

These studies all use the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) Depression Database, which would be described further in section 5.1 to solve the DCC.

2.1.1 Baseline

The audio features used by the AVEC2016 organizer, as shown in Table 2.1, include prosodic, voice quality and spectral features. They are extracted using the COVAREP (v1.3.2) (Valstar et al., 2016). On the other hand, the video features including facial landmarks, histogram of oriented gradients (HOG), gaze direction estimate for both eyes and head pose, are extracted using OpenFace framework (Valstar et al., 2016). Emotion and facial action unit (FAU) continuous measures are also extracted using FACET software (Valstar et al., 2016).

Audio Feature Group	Audio Features
Prosodic	Fundamental Frequency (F0), voicing (VUV)
Voice Quality	Normalized amplitude quotient (NAQ), Quassi open quotient (QOQ), the difference of the first 2 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 glottal pulse dynamics (Rd)
Spectral	Mel cepstral coefficients (MCEP0-24), Harmonic Model and Phase Distortion mean (HMPDM0-24) and deviations (HMPDD0-12)

Table 2.1 Baseline Audio Features (Valstar et al., 2016).

For depression classification, they use a linear support vector machine (SVM) provided by Scikit-learn toolbox with stochastic gradient descent (SGD) (Valstar et al., 2016). Grid search technique has been used to optimize the hyper parameters. The baseline is shown in Table 2.2. Meanwhile, they compute the regression baseline that is shown in Table 2.3 using random forest regressor (Valstar et al., 2016).

Partition	Modality	F1 Score	Precision	Recall
Development	Audio	0.41 (0.58)	0.27 (0.94)	0.89 (0.42)
Development	Audio-Video	0.58 (0.86)	0.47 (0.94)	0.78 (0.79)
Test	Audio	0.46 (0.68)	0.32 (0.94)	0.86 (0.54)
Test	Audio-Video	0.50 (0.90)	0.60 (0.87)	0.43 (0.93)

Table 2.2 Baseline result for depression classification. Performance is measured in F1 score (Valstar et al., 2016).

Partition	Modality	RMSE	MAE
Development	Audio	6.7418	5.3566
Development	Audio-Video	6.6212	5.5222
Test	Audio	7.7758	5.7224
Test	Audio-Video	7.0467	5.6567

Table 2.3 Baseline result for depression regression. Performance is measured in mean absolute (Valstar et al., 2016).

2.1.2 DepAudioNet

This recent study that is working on DCC presents DepAudioNet, which classifies depression through audio-based-features including the batch normalized (which include feature warping) Mel-scale filter bank feature, using the combination of Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) (Ma, Yang, Chen, Huang, & Wang, 2016). They resolved the bias caused by uneven sample distribution by using a random-sampling strategy (Ma et al., 2016). They successfully improve the F1 score as shown in Table 2.4.

Partition	Time Window W	Max-pooling length l	F1 score	Precision	Recall
Baseline (Valstar et al., 2016)	-	-	0.41(0.58)	0.27(0.94)	0.89(0.42)
Development	120	3	**0.52(0.70)**	0.35(1.00)	1.00(0.54)

Table 2.4 Performance of Mel-scale filter bank features with different parameters for DepAudioNet (values of non-depression in brackets) (Ma et al., 2016).

2.1.3 MIT Lincoln Laboratory (MITLL)

This study uses a combination of audio, video and semantic features to tackle DCC (Williamson et al., 2016). The audios provided are segmented based on the time in the transcript and then performed min-max scaling (Williamson et al., 2016). Next, they are filtered with cut-off frequency of -15dB before normalizing to the range of -1 to 1 (Williamson et al., 2016). As shown in Table 2.4, the correlation structure of formant tracks and delta MFCCs as well as lower vocal tract physiology features and loudness variation features are extracted from the pre-processed audio (Williamson et al., 2016). Besides, the correlation structure of FAU are used as video features (Williamson et al., 2016). This study also includes semantic content features which are not discussed in other related studies. In the semantic context, global vectors for word representation (GloVe) is a way to represent a set of questions or answers in the transcript which embeds words in a high dimensional space (Williamson et al., 2016). Principle Component Analysis (PCA) and the whitening (ZCA) transform are then used to transform the sparse GloVe into PCA- and ZCA-transformed embedded space because the transformed space would usually improve the performance of ML model (Williamson et al., 2016). As shown in Table 2.6, the Gaussian Staircase classifier and regressor that are trained using all the features

mentioned perform the best by providing highest mean F1, lowest RMSE and MAE (Williamson et al., 2016).

Audio Feature Group	Audio Features	Optimization
Spectral	Correlation structure of formant tracks, Correlation structure of delta MFCCs (dMFCCs)	PCA, z-scoring
Lower Vocal Tract Physiology Features	Lower Vocal Tract (VT) Resonance Pattern	-
Loudness Variation Features	Mean, standard deviation, range of peak-to-rms	-

Table 2.5 MITLL Audio Features (Williamson et al., 2016).

Proposer	Partition	Modality	Mean F1	RMSE	MAE
Baseline	Development	Audio	0.50	6.74	5.36
Baseline	Development	Audio-Video	0.72	6.62	5.52
MITLL	Development	Audio	0.57	6.38	5.32
MITLL	Development	Audio-Video-Semantic	**0.81**	**5.31**	4.18

Table 2.6 Performance of the classifier proposed by MITLL compared to the Baseline.

2.1.4 Pampouchidou's Study

This study suggests a few models based on audio, visual, text, gender and their fusion (Pampouchidou et al., 2016). In audio modality, audios are segmented per the timestamps in the transcript file followed by removing the segments of silence, laughter, sigh, synch, scrubbed entry and short (≤ 5ms) (Pampouchidou et al., 2016). Audio features including F_0, NAQ, QOQ, H1H2, PSP, MDQ, peakSlope, Rd, Rd conf, MCEP 0-24, HMPDM 1-24, HMPDD 1-12 and Formants 1-3 are extracted using COVAREP toolbox at 10-ms intervals (Pampouchidou et al., 2016). To consider about speaker dependency, besides normalizing F_0 to the range of 0 to 1, the deltas and delta-deltas for F_0 and MFCCs are also extracted (Pampouchidou et al., 2016). Other than these low level audio features, 8 high level audio features are calculated, including the Pause Ratio, Voiced Segment Ratio, Speaking Ratio, Mean Laughter Duration, Mean Delay to answer the question, Mean Duration of Pauses, Maximum Duration of Pauses and Fraction of pauses in overall time (Pampouchidou et al., 2016). On the other hand, visual features such as Landmark Motion History Images (LMHI), Landmark Motion Magnitude (LMM), Head Motion, Blinking Rate, Face Emotions, AUs, Gaze and Pose are extracted, although AUs and Gaze are considered as redundant features while LMHIFaceHOG is an important feature (Pampouchidou et al., 2016). Other than that, significant amount of text features are extracted, including 93 features extracted by LIWC software which indicate about individuals' social processes, 7 features extracted by the Affective Norms for English Words ratings (ANEW) for pleasure, arousal, dominance ratings and word frequency, fifth text-based feature extracted for the consideration of depression related words and features that indicate about non-verbal

symptoms of depression (Pampouchidou et al., 2016). Feature selection technique which accesses the feature importance by the effect of feature removal on F1 score are developed to reduce the feature dimension (Pampouchidou et al., 2016). There are 4 different decision fusions model which are all implemented using Decision Tree algorithm (Pampouchidou et al., 2016). The model with Decision Fusion {(Video OR Gender Based Audio) AND (Video OR Gender based Text)} (recorded as audio-video-text-gender modality) provides the best average F1 score in development set, while using gender based audio only provides the best average F1 score in test set (recorded as audio-gender modality) (Pampouchidou et al., 2016). The results are recorded in Table 2.7.

Proposer	Modality	Development F1 Score	Test F1 Score
Baseline	Audio	0.41 (0.58)	0.46 (0.68)
Baseline	Audio-Video	0.58 (0.86)	0.50 (0.90)
Pampouchidou	Audio-Gender	0.59 (0.87)	**0.52 (0.81)**
Pampouchidou	Audio-Video-Text-Gender	**0.62 (0.91)**	0.23 (0.71)

Table 2.7 Classifier performance proposed by Pampouchidou compared to the Baseline.

2.1.5 SCUBA

SCUBA from University of Southern California (USC) uses a fusion of audio and video features in solving the DCC (Nasir et al., 2016). For the audio modality, the i-vector of MFCCs is proposed as the audio feature, while the Gaussian Probabilistic Linear Discriminant Analysis (G-PDLA) models are chosen as the ML model (Nasir et al., 2016). For the video features used in the proposed video model, their velocity and acceleration are included in addition to those given in the baseline (Nasir et al., 2016). Besides, the video features comprise of geometrical features which are the distance and area features extracted from the face, as well as Polyfit to gain insights on the temporal variation of facial expression (Nasir et al., 2016). To improve the result, Mutual Information Maximization (MIM) based feature selection technique is used to select the best video features. The proposed audio classifier performs better than the baseline by providing F1 score of 0.57 for depressed class and F1 score of 0.89 for non-depressed class, while the proposed audio-video ensemble classifier also outperforms the baseline by providing F1 score of 0.63 for depressed class and F1 score of 0.89 for non-depressed class, as shown in Table 2.8 (Nasir et al., 2016).

Proposer	Partition	Modality	F1 Score	Precision	Recall
Baseline	Development	Audio	0.41 (0.58)	0.27 (0.94)	0.89 (0.42)
Baseline	Development	Audio-Video	0.58 (0.86)	0.47 (0.94)	0.78 (0.79)
SCUBA	Development	Audio	0.57 (0.89)	0.57 (0.89)	0.57 (0.89)
SCUBA	Development	Audio-Video	**0.63 (0.89)**	-	-

Table 2.8 Performance of the classifier proposed by SCUBA compared to the Baseline.4

Besides, a linear regressor trained by i-vector PDLA log-likelihood scores improves the baseline regressor by giving lower RMSE as shown in Table 2.9 (Nasir et al., 2016).

Proposer	Partition	Modality	RMSE	MAE
Baseline	Development	Audio	6.7418	5.3566
SCUBA	Development	Audio	**6.7334**	5.8237

Table 2.9 Performance of the regressor proposed by SCUBA compared to the Baseline

2.2 Other Related Studies

The SVM and Gaussian Mixture Model (GMM) are two most famous ML methods that are applied in depression classification task, well known for their ability to handle sparse dataset robustly with relatively small computational cost and presence in many free third-party libraries (Cummins et al., 2015). By performing well in mental state classification (Cummins et al., 2015), both models are implied to be capable in classifying depression.

GMM has 77% accuracy when classifying depression using the Mel Frequency Cepstral Coefficients (MFCCs) feature only, whereas 79% accuracy is obtained when using the combination of MFCCs and formant feature (Cummins et al., 2015). On the other hand, SVM tends to perform stronger than GMM in depression classification task, but might perform worse if an unsuitable kernel is used (Cummins et al., 2015).

MFCCs, with reasonably good results, is the most popular feature used in these studies (Cummins et al., 2015). However, MFCCs performs worse after feature warping is applied (Cummins, Epps, Breakspear, & Goecke, 2011). Feature warping is a method to equalize the distribution of every feature by mapping them to a predefined distribution (Sethu, Ambikairajah, & Epps, 2007). If feature warping is used with care, it is proven to be very effective in emotion classification (Sethu et al., 2007; Sethu, Ambikairajah, Epps, Wales, & Nsw, 2009) as well as depression classification (Cummins et al., 2011). Moreover, the natural fundamental frequency F0 is a useful feature in classifying depression (Cummins et al., 2015). Meanwhile, this natural variation in speech would cause issues when performing normalization on prosodic features, which are often ignored in the literature (Cummins et al., 2015).

2.3 Research Gap

As shown in the related studies, researchers used various ML models, different sets of audio features, pre-processing solutions and feature selection methods in solving the depression tasks. However, it is difficult to find literatures that compares all these proposed methods. Although many studies have chosen SVM and GMM in the past, these methods are not chosen in the AVEC2016 literatures. In addition, AdaBoost (AB) is a strong ML model which is usually

proven to have greater performance than many single classifiers (Dietterich, 2000) but not found to be used in any related studies.

Normalization or standardization of datasets is mandatory for many ML models (Scikit-learn developers, 2010e). Nevertheless, most of the reviewed literatures that had chosen different normalizing methods did not explain the rationale behind. Besides, there are many state of the art feature selection techniques, but many of them are not examined in the depression context. As feature selection methods are important to make the model more representable and generalizable (J. Li et al., 2016), the technique that we choose to search the optimal subset should be considered carefully. Furthermore, there are a variety of audio features proposed to solve the DCC, but there are still audio features that are not being explored. For instance, Chroma-based features are audio features that are related to harmony (Müller & Ewert, 2011) and would provide information independent of the spectrum features (Ellis, 2007), thus might be useful in the depression context.

To address these research gaps, this study explores and discusses a variety of audio features, ML models, feature selection methods and normalizing methods. At the same time, existing normalization methods are also compared and explained.

2.4 Chapter Summary

This chapter reviews a few related studies and identifies the need of comparing representative audio features, established ML models, the use of normalization and feature selection techniques to provide a guideline for the researchers who are interested in this field of study. Background knowledge will be discussed in Chapter 3 to provide more insights about the topic.

Chapter 3 Background Knowledge

All the background knowledge such as audio features, normalization techniques, feature selection techniques and ML models, that would need to understand before proceeding to the content of this study, would be presented in this section.

3.1 Audio Features

Before learning depression from the audio recordings, the representation of the audio recordings in the computer should be chosen. These audio representations are called audio features. They would be the audio data that are needed to train the ML model.

As this study would use the terminology "audio features" and "features" extensively, it is important to clarify the difference of these terms. As mentioned before, an audio feature is an audio representation which comprises of one or more features. A feature will be one of the dimensions of the ML model and it is usually represented by a number. In other words, an audio feature is a set of multiple features whereas a feature is a numerical representation that belongs to a set of audio features.

Audio features are divided into: time-domain, frequency domain (Chu, Narayanan, & Kuo, 2009) and chroma-based audio features (Müller & Ewert, 2011). They would be explained in the following sections.

3.1.1 Time-Domain Audio Features

Time-Domain Audio Features (or temporal features) are audio features directly extracted from the sampled audio signal data, which is dependent on time including Zero-Crossing Rate, Energy and Entropy of Energy (Chu et al., 2009).

Zero-Crossing rate is the rate at which the signal changes from positive to negative or back. It is a temporal feature that is useful in speech recognition and music information retrieval. It can be calculated by the division of the number of the signal changes and the number of samples in the frame, as presented in Equation 3-1.

$$Z_j = \frac{1}{2\,(S-1)} \sum_{i=2}^{S} |sgn(x_i) - sgn(x_{i-1})|$$

Equation 3-1 The definition of Zero Crossing Rate Z_j where S is the number of samples in the frame and $sgn(x_i)$ is the sign function defined in Equation 3-2 which will determine if the number is positive, negative or 0.

$$sgn(x) = \begin{cases} -1 \; if \; x < 0 \\ 1 \; if \; x > 0 \\ 0 \; elsewise \end{cases}$$

Equation 3-2 The definition of the sign function $sgn(x)$ which will determine if the number is positive, negative or 0.

The energy of signal processing is slightly different from the one in physics as defined in Equation 3-3, and it can be converted to the energy in physics using Equation 3-4. To have a fair comparison between all audio signals, the energy of signal processing is normalized by the respective frame length. It is one of the commonly used temporal features that represents the variation of the amplitude over time (Chu et al., 2009).

$$E_s(s) = \sum_t |s(t)|^2$$

Equation 3-3 Energy of Signal Processing $E_s(s)$ for signal (s) of discrete time (t).

$$E = \frac{E_s}{Z}$$

Equation 3-4 Conversion of Energy of Signal Processing using magnitude Z to Energy in physics E.

Entropy is a measure of uncertainty as defined in Equation 3-5. Therefore, the Entropy of Energy can be viewed as a temporal feature that is used to measure abrupt changes (Giannakopoulos, 2015) . Its value would be small when there is a huge change in energy level in a frame (Giannakopoulos, Kosmopoulos, Aristidou, & Theodoridis, 2006).

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Equation 3-5 Entropy H(x) of a random variable X.

3.1.2 Frequency-Domain Audio Features

Spectral Entropy, Spectral Centroid, Spectral Spread, Spectral Flux, Spectral RollOff and MFCCs belong to Frequency-Domain Audio Features extracted from the FFT-transformed audio signal which is on the frequency domain (Chu et al., 2009).

Similar to Entropy of Energy, Spectral Entropy is a spectral feature that is used to measure the distribution of the spectrum (Powell & Percival, 1979). It can differentiate a speech signal from a noise signal based on the variance of spectral magnitude (Wu & Wang, 2005).

Spectral Centroid is the weighted mean of the line spectrum, which is a static spectral energy distribution (Grey, 1978) as expressed in Equation 3-6. It is a spectral feature that could be used to numerically characterize a spectral distribution (Grey, 1978). Therefore, Spectral Centroid measures the "brightness" of sound (Chu et al., 2009) with positive correlation

(Schubert & Wolfe, 2006), which is one of the sound qualities that is believed to be correlated with increased power at high frequencies (Schubert & Wolfe, 2006).

$$f_c = \frac{\sum_{i=1}^{N} f\,(i)M(i)}{\sum_{i=1}^{N} M(i)}$$

*Equation 3-6 Spectral Centroid f_c is given in frequency f(i) * Magnitude M(i), with the Magnitude M(i) as the weight* (Schubert & Wolfe, 2006).

Spectral Spread is defined as the second central moment of the spectrum. In statistics, a moment is a specific quantitative measure of the shape of a set of points. The central moment is a moment of a random variable about the random variable's mean or expected value. The nth central moment is expressed as Equation 3-7. The 2nd central moment is a special case because it is the definition of the variance, as shown in Equation 3-8. Therefore, Spectral Spread is also defined as the variance of the spectrum. As Spectral Centroid is the first central moment (or the mean) of the spectrum, Spectral Spread can be defined as a spectral feature that is used to measure the bandwidth of the spectrum, as provided in Equation 3-9 (Bello, 2013).

$$\mu_n = E[\,(X - E[X])^n\,] = \sum_{x \in X} (\,X - E[X]\,)^n\, f(x)$$

Equation 3-7 The nth central moment μ_n of a discrete real-valued random variable X where E is the expectation operator.

$$\mu_2 = E[\,(X - E[X])^2\,] = E[X^2] - E[X]^2 = Var[X]$$

Equation 3-8 The 2nd central moment μ_2 or the (variance Var[X]) of a real-valued random variable X where E is the expectation operator.

$$f_s = \frac{\sum_{i=1}^{N} (f(i) - f_c\,)^2 M(i)}{\sum_{i=1}^{N} M(i)}$$

Equation 3-9 The definition of Spectral Spectrum f_s where M(i) is the magnitude of the signal, f(i) is the frequency, and f_c is the Spectral Centroid (Bello, 2013).

Spectral Flux is the "squared difference between the normalized magnitudes of the spectra of the two successive frames" (Giannakopoulos, 2015), as presented in Equation 3-10. It is a spectral feature that measures the local spectral change (Giannakopoulos et al., 2006).

$$f_t = \sum_{i=1}^{n} (N_t(i) \; - \; N_{t-1}(i))^2$$

Equation 3-10 The definition of Spectral Flux f_t where N_t and N_{t-1} are normalized magnitudes of time frame t and t – 1 respectively.

Spectral Rolloff is defined as the frequency below which c% (c=90 in this case) of the magnitude distribution of the spectrum is concentrated in a frame (Giannakopoulos, 2015). The

concept is given in Equation 3-11. It calculates the skewness of the spectral shape and is positively correlated with the brightness of the sound (Giannakopoulos et al., 2006).

$$\sum_{i=1}^{R_C(j)} |M_j(i)| = \frac{c}{100} \sum_{i=1}^{S} |M_j(i)|$$

Equation 3-11 The use of Spectral Rolloff $R_C(j)$ to concentrate in c% of the magnitude distribution M_j of frame j with S number of samples.

As specified in section 2.2, MFCCs is a good audio feature often used in related studies. It can approximate human's auditory system responses more closely as stated in Appendix A.

3.1.3 Chroma-based Audio Features

Chroma Vector and Chroma Deviation are Chroma-based audio features which correlate to pitch and harmony (Müller & Ewert, 2011).

The features in a Chroma Vector are beat-synchronous chroma features (Ellis, 2007). It is a "twelve-element vector with each dimension representing the intensity associated with a particular semitone, regardless of octave" (Ellis, 2007). We can also view each dimension as the spectral energy which represents one of the 12 equal-tempered pitch classes of western-type music (Giannakopoulos, 2015). They are usually used to "reflect melodic and harmonic content and be invariant to instrumentation" (Ellis, 2007), thus is more commonly used in music audio classification. Despite being less informative, Chroma Vector provides information independent of the spectral features (Ellis, 2007), hence it might potentially improve either the depression classification result or depression regression result.

Chroma Deviation is the standard deviation (std) of the 12 features contained in Chroma Vector. Std is defined in Equation 3-12 and it is used to measure the variation of a set of data. In other words, Chroma deviation actually measures the variation of the 12 features in Chroma Vector. In speech segments, the variation of each feature in Chroma Vector is high over successive frames (Giannakopoulos, Pikrakis, & Theodoridis, 2008).

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \; Where \; \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Equation 3-12 The std σ of a finite data set $x_1, x_2, \ldots, x_n$ where μ is the mean

We extract these features by calling the command provided in the github repository of pyAudioAnalysis (link: https://github.com/tyiannak/pyAudioAnalysis) and the command parameters all use the default value provided by the author. It implies that the features are

extracted using a frame size of 50ms with 50% overlap. The command is written in a batch file called "usingPyAudioFeature.bat" to automate the process.

3.1.4 Section Overview

Index	Audio Feature Name	Description
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of sub-frames' normalized energies, it can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum.
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9~21	MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
22~33	Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
34	Chroma Deviation	The std of the 12 Chroma coefficients.

Table 3.1 Audio Features implemented by pyAudioAnalysis (Giannakopoulos, 2015).

Table 3.1 summarizes the audio features that are explained in this section. An open-source python library name "pyAudioAnalysis" is adapted to extract the audio features listed in this section and more details would be discussed in section 5.3.

3.2 Data Normalization Techniques

Data normalization including feature warping and min-max scaling is required for many ML algorithms (Scikit-learn developers, 2010e). This chapter reviews a few existing data normalization techniques.

3.2.1 Feature Warping

Feature warping is a method which treats each feature independently and equalizes the distribution of every feature by mapping them to a predefined distribution (Sethu et al., 2007). Standardization is a similar normalizing method implemented by Scikit, which maps the distribution of the data to Gaussian distribution which has zero mean and unit variance (Scikit-learn developers, 2010e). Standardization is mandatory to many ML models implemented in

Scikit (Scikit-learn developers, 2010e). Nevertheless, standardization might cause the prosody feature to lose its natural variation (Cummins et al., 2015).

Literature shows that the accuracy of classifier improves when feature warping applies on features such as the fundamental frequency F0, the short-term energy and Zero Crossing Rate but drops while applying on MFCCs and Spectral Centroid (Cummins et al., 2011). Therefore, feature warping is still useful when used with care. Related studies such as **DepAudioNet** has made use of this technique to normalize the features.

3.2.2 Min-max Scaling

Another existing normalization method is min-max scaling, which treats every feature independently and rescales them to the range in between the minimum and the maximum value, usually in the range of 0 to 1 (Scikit-learn developers, 2010e). Similarly, the feature can be rescaled to the range of 0 to maximum absolute value (Scikit-learn developers, 2010e).

Related studies such as **MITLL** and **Pampouchidou's Study** has adapted this technique in normalizing the features even though there are no justifications about the choice of this normalization method.

3.2.3 Section Overview

This section reviews normalization methods used in related studies include Feature Warping and Min-max Scaling. Further reviews of these methods will be explored in the next chapter.

3.3 Feature Selection Techniques

Feature selection techniques are useful in ML tasks by eliminating irrelevant features or selecting representable training data, so that a cleaner and more understandable ML model can be constructed (J. Li et al., 2016). Thus, feature selection is applied to the static dataset obtained in this study, which can be categorized as either generic data or heterogeneous data (J. Li et al., 2016). Feature selection techniques applied to generic data usually assume that every feature is independent from each other, whereas those proposed for heterogeneous data with multi-view treat data from various feature spaces (J. Li et al., 2016). Although our data is more suitable to be categorized as heterogeneous data as it contains a group of audio features, there is no algorithm implemented in Scikit-Feature targeting this type of data as it focuses on evaluating the performance of feature selection approach for generic data (J. Li et al., 2016). Thus in the current stage, famous feature selection mechanisms implemented for generic data are examined instead. While this study only uses supervised feature selection techniques, this

section discusses about a few techniques that are experimented to obtain a more accurate audio representation for depression classification.

3.3.1 Supervised Similarity Based

Supervised Similarity based Feature Selection Methods rate the importance of features to conserve the data similarity, which can be calculated using class information (J. Li et al., 2016). Among them, "Fisher Score" and "Relief" algorithms (J. Li et al., 2016) are implemented and evaluated in this study. Feature selection algorithms using data similarity are examined as they might be able to find the common pattern between the speech signals of depressed patients.

Fisher Score, which is measured using Equation 3-13, rates the features which have low feature values within the same class but high feature values from different classes higher and chooses the top k features which have the highest k Fisher scores (J. Li et al., 2016). As the original fisher algorithm is inadequate to discard irrelevant features caused by the independence of similarity calculation for each feature, a Generalized Fisher Score method which picks features concurrently and maximizes the lower bound of Fisher Score is proposed to obtain the optimal subset of features (J. Li et al., 2016). This algorithm is chosen due to its simplicity and comprehensiveness.

$$fisher\_score(f_i) = \frac{\sum_{j=1}^{c} n_j \, (\mu_{i,j} - \mu_i \,)^2}{\sum_{j=1}^{c} n_j \, \sigma_{i,j}^2}$$

Equation 3-13 The Formula of Fisher Score where c is the number of classes, f_i is the i-th feature, n_j is the number of samples in class j, μ_i is the mean value of f_i, $\mu_{i,j}$ is the mean value of f_i in class j and $\sigma_{i,j}^2$ is the variance of f_i in class j (J. Li et al., 2016).

Relief, which is a filter algorithm for binary classification, picks features to detach samples from distinct classes and measures the score using Equation 3-14 (J. Li et al., 2016). ReliefF extends the Relief algorithm to solve the multi-class classification tasks and selects features that keep an unique data similarity matrix which is derivable from the classes (J. Li et al., 2016). Relief is chosen because the algorithm is easily understandable.

$$relief\_score(f_i) = \; 0.5 \sum_{j=1}^{k} [d\big(X(j,i) - X(NM(j),i)\big) - \; d\big(X(j,i) - X(NH(j),i)\big)]$$

Equation 3-14 Relief Score which is measured using k randomly selected samples from n samples where f_i is the i-th feature, NM(j) and NH(j) denotes the nearest sample from same class and other class respectively, d(.) is a distance function which is usually defined as Euclidean distance (J. Li et al., 2016).

3.3.2 Supervised Informational Theoretical Based

These algorithms are usually used on discrete data only and are designed based on heuristic filter criteria to maximize feature relevance calculated using its correlation with the classes, but not constructed to find optimal set of features globally because it is NP-hard (J. Li et al., 2016). Feature selection techniques belong to this family include Mutual Information Feature Selection (MIFS), Minimum Redundancy Maximum Relevance (MRMR), Conditional Infomax Feature Extraction (CIFE), Conditional Mutual Information Maximization (CMIM), Interaction Capping (ICAP) and Fast Correlation Based Filter (FCBF) (J. Li et al., 2016). They are experimented in this study due to their simplicity and efficiency.

First of all, the concept of information gain is defined. As expressed in Equation 3-15, it accounts the separateness of training samples with a feature (Sugumaran, Muralidharan, & Ramachandran, 2007). Similarly, it can be generally formulated in Equation 3-16, interpreted as the mutual information of 2 variables computed by the difference of entropy and conditional entropy (J. Li et al., 2016). This section will measure information gain based on Equation 3-16.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Equation 3-15 The definition of Information Gain where S is the samples, A is the feature, Value(A) is possible values for A, S_v is the subset of S which has value v in A (Sugumaran et al., 2007) *and Entropy(.) is defined in Equation 3-5.*

$$I(X, Y) = Entropy(X) - Entropy(X|Y)$$

Equation 3-16 Generalized version of information gain where X and Y are random variables where Entropy(.) is defined in Equation 3-5.

By maximizing the feature correlation with classes and minimizing the correlation between each feature, MIFS accounts for both importance and redundancy of features and ranks them based on Equation 3-17 (J. Li et al., 2016). It follows the exploration-exploitation tradeoff framework by balancing the importance of adding a new feature, measured in the first term, and the redundancy of adding a new feature, calculated in the second term which penalizes the feature with high mutual information with the chosen features (J. Li et al., 2016). It is chosen as it is simple and reasonably heuristic.

$$J_{MIFS}(f_i) = I(f_i, Class) - \beta \sum_{f_j \in S} I(f_j, f_i)$$

Equation 3-17 Feature Score based on MIFS where f_i is the i-th feature, S is the selected feature set which is initially empty, β is an empirical parameter that is set to 1 and I(.) is defined in Equation 3-16 (J. Li et al., 2016).

MRMR is similar to MIFS, except that the empirical parameter β is set as the reverse of the selected feature set dimension to moderately reduce the penalization of feature redundancy if given more chosen features (J. Li et al., 2016). The algorithm makes sense as new feature is less likely to be redundant to the relevant feature in selected feature sets and it tends to be more and more pairwise independent when the dimension of the selected feature set arises (J. Li et al., 2016). It is implemented because it is a popular extension of MIFS.

Other than the feature relevance and redundancy, CIFE also maximizes the conditional redundancy between new unchosen features and the picked features which provided the classes, in order to penalize the features that are highly irrelevant to the latter, as formulated in Equation 3-18 (J. Li et al., 2016). It is also experimented as it is an extension of MIFS and MRMR.

$$J_{CIFE}(f_i) = I(f_i, Class) - \sum_{f_j \in S} I(f_j, f_i) + \sum_{f_j \in S} I(f_j, f_i \mid class)$$

Equation 3-18 Feature Score based on CIFE where f_i is the i-th feature, S is the selected feature set which is initially empty and I(.) is defined in Equation 3-16.

CMIM, which computes the feature score based on Equation 3-19, continuously picks features such that the mutual information with the classes provided the chosen features is maximized and discards features which is not distinctive from the chosen features even if they have potential to predict classes well (J. Li et al., 2016). This study tests this algorithm as its concept is different from MIFS, MRMR and CIFE.

$$J_{CMIM}(f_i) = \min_{f_j \in S}[\, I(f_i, class \mid f_j)\,] = I(f_i, class) - \max_{f_j \in S}\left[\, I(f_j, f_i) - I(f_j, f_i \mid class)\,\right]$$

Equation 3-19 Feature Score based on CMIM where f_i is the i-th feature, S is the selected feature set which is initially empty and I(.) is defined in Equation 3-16 (J. Li et al., 2016).

ICAP is similar to CMIM except that it would maximize the absolute value of the term $I(f_j, f_i) - I(f_j, f_i \mid class)$ (J. Li et al., 2016). It is tested as it is a variant of CMIM.

FCBF is a filter technique which is not able to be generalized to the unified conditional likelihood maximization framework (J. Li et al., 2016). It exploits both correlations of feature-class and feature-feature at the same time (J. Li et al., 2016). Firstly, it chooses a feature subset S that has high feature-class correlation with the symmetric uncertainty (SU) based on Equation 3-20 ≥ certain threshold δ (J. Li et al., 2016). Feature f_k is defined as predominant if and only if SU($f_k, class$) ≥ δ and feature $f_i \in$ S(i ≠ k) which satisfies $SU(f_i, f_k) \geq SU(f_k, class)$ do not exist (J. Li et al., 2016). If such feature f_i exists, then it must be a redundant feature to f_k (J. Li et al., 2016). Based on these definitions, the redundant feature set denoted as S_P is

constructed and can be further divided into 2 subsets : $S_P^+ \; and \; S_P^-$ which have redundant feature f_i to predominant feature f_k satisfied by $SU(f_i, f_k) > SU(f_k, class)$ and $SU(f_i, f_k) < SU(f_k, class)$ respectively (J. Li et al., 2016). By applying different heuristics on S_P, $S_P^+ \; and \; S_P^-$, redundant features can be eliminated while the most relevant features to classes are preserved (J. Li et al., 2016). However, as it operates by finding the optimal subset of features, the developers could not specify the number of features that they want to select (J. Li et al., 2016). This study will evaluate this special approach to the depression audio dataset.

$$SU(f_S, class) = 2\frac{I(f_S, class)}{H(f_S) + H(class)}$$

Equation 3-20 The formula for symmetric uncertainty (SU) where f_S is the feature set, information gain I(.) defined in Equation 3-15 and Entropy H(.) defined in Equation 3-5 (J. Li et al., 2016).

3.3.3 Supervised Statistical Based

This section reviews CFS and GINI feature selection techniques. They belong to a family of techniques named "Supervised Statistical based Feature Selection Methods" which compute the feature importance using statistical mathematical functions without considering feature redundancy by evaluating every feature independently (J. Li et al., 2016). Thus, this study examines these techniques to show the importance of not treating features individually.

CFS uses a correlation based statistical method, shown in Equation 3-21, to measure the importance of feature subset S (J. Li et al., 2016). It can be conceptualized as strong feature-class correlation, which is an important property for optimal feature subset (J. Li et al., 2016).

$$CFS\_score(S) = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

Equation 3-21 The evaluation metric of CFS where S is the feature subset, k is the number of features, $\overline{r_{cf}}$ is the average of feature-class correlation and $\overline{r_{cf}}$ is mean feature-feature correlation (J. Li et al., 2016).

Gini Index measures the capability of a feature to separate samples from different classes using statistical formula given in Equation 3-22 (J. Li et al., 2016). Unlike other measures, the higher the gini index, the more redundant a feature is (J. Li et al., 2016).

$$gini\_index\_score(f_i) = \min_W \left(p(W)\left(1 - \sum_{s=1}^{c} p(C_S|W)^2\right) + p(\overline{W})\left(1 - \sum_{s=1}^{c} p(C_S|\overline{W})^2\right)\right)$$

Equation 3-22 Formula used to calculate the score for Gini Index, provided that f_i is the i-th feature with r different feature values where the dataset is split into samples set W and $\overline{W}$ using the j-th feature as the margin value. C_S implies that s is the class label and p(.) means probability (J. Li et al., 2016).

3.3.4 Section Overview

This section introduces many established popular feature selection heuristics. However, most of them might not work well in the dataset used in this study as it contains a few audio feature sets. The evaluation of these feature selection techniques would be provided in Chapter 6. On a side note, the reason for not using some simple feature selection methods such as Tree-based Selection and L1-based feature selection are given in Appendix C for more insights.

3.4 Machine Learning (ML) Models

Choosing a suitable ML model that performs well in depression diagnosis or other real world problem is crucial (Kotsiantis, Zaharakis, & Pintelas, 2007). To make an informed decision, basic understanding of these models and their good performance based on evaluation metrics are both important. This study focuses on researching the effectiveness of various supervised ML models, which would be reviewed in this section.

Different kinds of commonly used supervised ML models are experimented on their feasibility on both depression classification and regression, including probabilistic model (Naïve Bayes, Gaussian Process), kernel-based model (Gaussian Process, Support Vector Machine), ensembles (Random Forest, AdaBoost) and other simple classifiers (K-Nearest Neighbor, Decision Tree). As we have very limited data with only 107 training points, some state of the art ML models that require a lot of data such as Neural Network (Dietterich, 2000) are not experimented.

3.4.1 Probabilistic Machine Learning Model

Probabilistic ML model aims to model the data uncertainty using probability, which is the mathematical method of predicting uncertainties (Ghahramani, 2015). These models are famous for being conceptually simple, although they can be computationally complex (Ghahramani, 2015). Probabilistic ML model can be more generalizable to all types of training data by defining them as nonparametric models (Ghahramani, 2015).

Naïve Bayes (NB) is a probabilistic ML model based on the assumption of all features are independent with each other (Kotsiantis et al., 2007) as shown in Equation 3-23. Nevertheless, as that assumption is not valid most of the time, NB is usually less accurate compared to other learners (Kotsiantis et al., 2007). However, it is chosen as one of our ML models because of its high computational speed and simplicity as illustrated in Equation 3-24 (Kotsiantis et al., 2007).

$$P(X|i) = \prod P\,(X_r|i)$$

Equation 3-23 The assumption made by Naïve Bayes where X is the feature vector and X_r is the value for each feature that forms the feature vector.

$$R = \frac{P(i\,|X)}{P(j|X)} = \frac{P(i)P(X|i)}{P(j)P(X|j)} = \frac{P(i)\prod P(X_r|i)}{P(j)\prod P(X_r|j)} \begin{cases} R > 1\ then\ predict\ i \\ Else\ predict\ j \end{cases}$$

Equation 3-24 The classification method of Naïve Bayes (Kotsiantis et al., 2007).

Gaussian Process (GP) defines a probability distribution for functions (Rasmussen & Williams, 2006; Snelson, 2007). In GP, any function f is in Gaussian distribution or Normal distribution (Snelson, 2007). As shown in Equation 3-25, the distribution of the conditional probability p(f|X) is modelled instead of p(x), thus GP is a conditional probabilistic model (Snelson, 2007). Besides, the multivariate Gaussian distribution has a joint probability density that can be calculated by Equation 3-26 (Rasmussen & Williams, 2006). GP is widely used to solve hard problems as it is non-parametric, easy to compute and provides an acceptable accuracy in uncertainties predictions (Seeger, 2004). As GP also provides error bars on prediction (Rasmussen & Williams, 2006), it is applicable in depression classification to provide the psychiatrist a brief overview about the prediction accuracy and the depression severity of the patient. Thus, it is chosen as an important ML technique to test on.

$$p(f)\ \equiv p(f|X) = N(\mu, \mathrm{K})$$

Equation 3-25 Gaussian distribution property of the function vector f with the corresponding inputs X where μ is the mean vector, K is the covariance matrix (Rasmussen & Williams, 2006).

$$P(f|\mu, K) = (2\pi)^{-\frac{D}{2}}\,|K|^{-\frac{1}{2}} e^{-\frac{1}{2}(f-\mu)^T K^{-1}(f-\mu)}$$

Equation 3-26 The definition of Multivariate Gaussian Distribution where f is the function vector, μ is the mean vector, K is the covariance matrix and D is the dimension of the function vector, mean vector and covariance matrix (Rasmussen & Williams, 2006).

3.4.2 Kernel-based Machine Learning Model

Another powerful technique to model uncertainty is by encoding experiences or training data into a covariance function or kernel (Seeger, 2004). The models that make use of this technique are kernel-based ML models. However, the selection of kernels should be done with care as it is an important factor that affects the generalizability of the model (Seeger, 2004).

GP is also a kernel based ML model. Its covariance matrix is an important quantity which is constructed by using a covariance function or kernel, K(x, x') as defined in Equation 3-27 and has to be positive semidefinite (Snelson, 2007). The kernel is a similarity measure of different points (Snelson, 2007). Thus, choosing different kernels would lead to different conclusions.

$$K_{ij} = K(x_i, x_j) = \varepsilon[f(x_i), f(x_j)]$$

Equation 3-27 The definition of covariance function or kernel (Snelson, 2007).

Another supervised ML model that uses kernel is the support vector machine (SVM) (Kotsiantis et al., 2007; Seeger, 2004). SVM is experimented because it is one of the most famous modeling and classification techniques used for depression prediction (Cummins et al., 2015). In addition, SVM focuses on the line that divides the data into two classes to improve the accuracy of the prediction, as shown in Equation 3-28 (Kotsiantis et al., 2007). When the data is linearly separable, it is possible to find the optimum margin that maximizes the distance between the separating planes by minimizing the squared norm of the separating plane, as shown in Equation 3-29 (Kotsiantis et al., 2007). Appendix C provides more explanation about norm. On the other hand, when the data is not linearly separable, the data is mapped to a higher dimensional space, which is called the transformed feature space, through function f to be linearly separated (Kotsiantis et al., 2007). Fortunately, the training algorithm of SVM depends only on the dot products of the mapped data in the transformed feature space (Kotsiantis et al., 2007), which can be defined as kernel provided in Equation 3-27. Therefore, in practice, SVM works for any diverse type of prediction problem.

$$w \cdot x_i - b \begin{cases} \geq +1 - \xi_i \ for\ y_i = +1\ (when\ the\ data\ i\ belongs\ to\ positive\ class) \\ \leq -1 + \xi_i\ for\ y_i = -1\ (when\ the\ data\ i\ belongs\ to\ negative\ class) \end{cases}$$

Equation 3-28 The decision rule for SVM or the line used to separate the 2 classes in SVM where w is the weight vector, b is the bias and $\xi_i \geq 0$ is the slack variable used to account the cases of misclassification which implies that $\sum_i \xi_i$ is an upper bound of training error (Kotsiantis et al., 2007).

$$Minimize_{w,b} \Phi(w) = \frac{1}{2} ||w||^2$$

Equation 3-29 Algorithm used by SVM to optimize the margin where w is the weight vector and b is the bias (Kotsiantis et al., 2007).

The choice of kernel for SVM is important as the transformed feature space is defined by the kernel function (Kotsiantis et al., 2007). In fact, choosing a wrong kernel for SVM would lead to a poor result (Cummins et al., 2015). Thus, selecting an appropriate kernel is important for both GP and SVM. It is a common practice to apply model selection techniques such as cross-validation to select the best kernel after choosing a few potential kernels for both GP and SVM (Kotsiantis et al., 2007; Seeger, 2004). However, the main drawback is the decreased efficiency of training a model as this approach would slow down GP and SVM. Another method is to select or design kernel carefully based on the known characteristic of the problem (Seeger, 2004). Therefore, standard kernels shown in Table 3.2 are chosen to be the kernel for GP or SVM or both to save time. Dot Product (DP) Kernel and Linear Kernel are chosen for

GP and SVM respectively because they are simple and fast to compute, whereas RBF is chosen for both GP and SVM because of its high degree of smoothness. Even though RBF unrealistically assumes that many physical processes are very smooth and it has unreasonably small predictive variances when used for time series prediction, it is experimented because it is recommended for high-dimensional input (Seeger, 2004).

Kernel	Description	Equation
Dot Product (DP)	DP Kernel depends only on x and x' through their dot product. It is invariant to rotation of the coordinates about the origin, but not translations (Rasmussen & Williams, 2006).	$k(x, x') = \sigma_0^2 + x \cdot x'$
Linear	Linear Kernel is a special case of dot product kernel when $\sigma_0^2 = 0$ (Rasmussen & Williams, 2006).	$k(x, x') = x \cdot x'$
Radial Basis Function (RBF)	RBF kernel is an isotropic kernel that depends only on x and x' as w > 0 is a length scale constant for all combinations of x, x' (Seeger, 2004). It can also be anisotropic, named squared-exponential kernel when W is a diagonal matrix that has dimension as the number of data and length scale for each dimension (Seeger, 2004). As RBF is infinitely differentiable, it has mean square derivatives of all orders, thus it is very smooth (Seeger, 2004).	$k(x, x') = e^{-\frac{w}{2}\|\|x-x'\|\|^2}$

Table 3.2 Standard kernels used in this project.

3.4.3 K-Nearest Neighbor (KNN)

Based on the assumption of the similar data should be nearer to each other, K-Nearest Neighbor (KNN) predicts an unlabeled data depending on majority of the K-Nearest data points labels (Kotsiantis et al., 2007). The relative distance of the data points is given by a distance metric (Kotsiantis et al., 2007). As shown in Equation 3-30, the common distance metric used by Scikit is Euclidean distance (Scikit-learn developers, 2010d). KNN is experimented in this project as it is useful in many real domains (Kotsiantis et al., 2007). Although the choice of distance metric and k will affect KNN performance (Kotsiantis et al., 2007), they are not the main focus in this project and thus the default option would be used. However, feature selection that could improve KNN performance (Kotsiantis et al., 2007) is performed in this project.

$$Euclidean\ D(x, y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$

Equation 3-30 The definition of Euclidean Distance (Kotsiantis et al., 2007).

3.4.4 Decision Tree

Decision tree (DT) is a tree that predicts the label of the data based on sorted features (Kotsiantis et al., 2007). In DT, each node is a feature, each branch is a possible value whereas each leaf is a label (Kotsiantis et al., 2007; Sugumaran et al., 2007). Although many methods

can be used to search the features to split the data such as information gain and gini index, no single best method exists (Kotsiantis et al., 2007). In Scikit, DT is implemented using Classification and Regression Tree (CART) algorithm which splits the data based on the largest information gain, as defined in Equation 3-15 (Scikit-learn developers, 2010a). Although it is recommended to compare different methods of data splitting (Kotsiantis et al., 2007), Scikit implementation of DT is directly used in this project due to time constraints. DT is selected in this project with its good comprehensibility (Kotsiantis et al., 2007).

3.4.5 Ensemble

The ML models mentioned above are single models which operate by selecting and optimizing a function named as "hypothesis" to approximate the uncertainties (Dietterich, 2002). Ensemble is a ML algorithm that models uncertainties through collecting the hypotheses generated by a set of normal ML models and selecting the best hypothesis using a kind of voting algorithm (Dietterich, 2002). Generally, ensemble provides a better model performance and higher accuracy than models that use single hypothesis (Dietterich, 2000), as it reduces the risk of choosing the wrong hypothesis, better estimates the correct hypothesis and has a larger possible space of representable hypothesis given a finite dataset (Dietterich, 2000, 2002).

Random Forest (RF) is an ensemble of DTs, making it a stronger ML model (Touw et al., 2013). Every DT in RF is trained by a random subset of train data to ensure low correlation between DTs and prevent overfitting (Touw et al., 2013). Its prediction is based on majority of the unweighted vote of DTs (Touw et al., 2013). RF is experimented in this project as it is efficient in training, non-parametric, relatively tolerant to noise and outliers (Touw et al., 2013), as well as being selected as the baseline model in AVEC 2016 (Valstar et al., 2016).

AdaBoost (AB) is an ensemble of ML models which is trained by different random subset of train data such as RF (Dietterich, 2000). However, a set of weights for different models are calculated by minimizing the weight errors based on the prediction accuracy of the training data (Dietterich, 2000). The unknown data is predicted depending on the weighted vote of different learners in AB (Dietterich, 2000). Although AB is prone to overfitting in high-noise cases, it generally gives good performance (Dietterich, 2000) and thus it is experimented.

3.4.6 Multi-class Classification

This section discusses the common methods to generalize a ML model which is designed solely for binary classification to use as a multi-class classifier. These methods include One-Vs-The-Rest (OVR) or One-Vs-One (OVO) strategy (Scikit-learn developers, 2010b).

OVR, also named as “One-Vs-All”, trains a binary classifier which takes in a class as a positive input and others as negative inputs (Scikit-learn developers, 2010b; Galar, Fernández, Barrenechea, Bustince & Herrera, 2011). Hence, each classifier represents a class while its positive output indicates the corresponding class as a predicted class (Scikit-learn developers, 2010b; Galar et al., 2011). Using this approach, the information about a class can be learned by investigating the corresponding classifier (Scikit-learn developers, 2010b).

Among all the classes, OVO constructs a binary classifier for each pair to learn about the differences between them (Scikit-learn developers, 2010b; Galar et al., 2011). The prediction of an OVO model is selected depending on the number of votes for each class and the confidence level (Scikit-learn developers, 2010b). While its main drawback is consuming more memory and time, it is useful for techniques which do not scale well with large samples of data (Scikit-learn developers, 2010b).

Other than that, there are newer approaches to convert a binary classifier to a multi-class classifier such as probability estimates and binary-tree based strategies (Galar et al., 2011). However, there are not explored in this study.

3.4.7 Optimization

As the choice of the hyper-parameters might affect the performance of the respective ML model, they might have to be tuned using optimization technique. Nowadays, Bayesian optimization (BO) is getting famous for hyper-parameter tuning (Melorose, Perroy, & Careas, 2015). BO has two important building blocks which are the probabilistic surrogate model and the loss function (Melorose et al., 2015). The model captures the belief about the characteristic of the unknown objective function and observation function related to data generation, whereas the loss function describes the optimality of the sequence of queries (Melorose et al., 2015). It is an useful and popular technique for joint optimization of design choices (Melorose et al., 2015). Therefore, it is experimented using the implementation of GPyOpt (The GPyOpt authors, 2016) to tune the hyper-parameters of scikit ML models.

3.4.8 Section Overview

Table 3.3 summarizes the definitions of different ML models and provides the Scikit library which can be used in the experiment implementations of this study. Some optimization methods are also reviewed in this section.

Names	Scikit library	Description
Naïve Bayes (NB)	GaussianNB	NB estimates based on the assumption that all features are independent with each other (Kotsiantis et al., 2007).
Gaussian Process (GP)	GaussianProcessClassifier ,GaussianProcessRegressor	GP focuses on stochastic processes which generalize the probability distribution for functions (Rasmussen & Williams, 2006).
Support Vector Machine (SVM)	SVC, SVR	SVM tries to maximize the linear margin between the two separating hyperplane and thus focus on creating the largest possible distance between the two hyperplanes (Kotsiantis et al., 2007).
K-Nearest Neighbours (KNN)	KNeighboursClassifier, KNeighborsRegressor	KNN is constructed based on the assumption of all the similar data will generally exist closer to each other (Kotsiantis et al., 2007).
Decision Tree (DT)	DecisionTreeClassifier, DecisionTreeRegressor	DT contains a decision tree which performs the classification task by sorting them based on feature values (Kotsiantis et al., 2007).
Random Forest (RF)	RandomForestClassifie, RandomForestRegressor	RF is an ensemble of decision trees trained with different subsets of the training examples (Touw et al., 2013).
AdaBoost (AB)	AdaBoostClassifier, AdaBoostRegressor	AB is an ensemble of learners with different hypothesis by making changes to training examples. It maintains a set of weights over the learners (Dietterich, 2000).

Table 3.3 The definition of ML techniques that are used commonly in this project in both classification and regression task.

3.5 Evaluation Metrics

Before performing a meaningful comparison for ML models, we would need to define our evaluation metrics. This section will discuss several common evaluation metrics used for either Classification or Regression in this study, which the former includes Precision, Recall and Fscore and the later includes the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE).

3.5.1 Classification

Correctness of depression classification can be measured by calculating the amount of correctly recognized depressed individuals (true positives), the number of correctly recognized individuals who are not depressed (true negatives) and examples that either were incorrectly predicted as depressed individuals (false positives) or that were not recognized as depressed individuals (false negatives) (Sokolova & Lapalme, 2009). As shown in Table 3.4, these four can be organized as the confusion matrix for Depression Classification, which is constructed based on the common confusion matrix for binary classification (Sokolova & Lapalme, 2009).

		True	False
Is he/she depressed?	Yes (Positive)	True Positives (tp)	False Positives (fp)
	No (Negative)	True Negatives (tn)	False Negatives (fn)

Table 3.4 Confusion Matrix for Depression Classification.

Precision, Recall and Fscore (or F measures) are frequently-used evaluation metrics in Text Classification (Sokolova & Lapalme, 2009). The formulas of these metrics focus on measuring the importance of positive samples retrieval, ignoring the true negatives (Sokolova & Lapalme, 2009). They focus on true positive classification as positive samples retrieval is more important in some contexts (Sokolova & Lapalme, 2009) such as depression classification.

Precision, as shown in Equation 3-32, is defined as correctly recognized positive examples divided by the total number of examples recognized as positive by the classifier (Sokolova & Lapalme, 2009). In other words, precision focuses in evaluating the correctness of the total examples recognized as positive by the classifier (Sokolova & Lapalme, 2009). High Precision would show that the model retrieves a lot of correct examples.

$$Precision = \frac{tp}{tp + fp} = \frac{correctly\ recognized\ positive\ examples}{total\ examples\ recognized\ as\ positive}$$

Equation 3-31 The definition of Precision (Sokolova & Lapalme, 2009) *where tp and fp are defined in Table 3.4.*

Recall is the number of correctly recognized positive examples divided by the total number of actual positive examples in the data (Sokolova & Lapalme, 2009), as expressed in Equation 3-32. Recall focuses on how effective the classifier can recognize positive examples (Sokolova & Lapalme, 2009) by measuring the relevance of the total positive labelled examples. High Recall would show that the classifier is very effective in classifying the positive label.

$$Recall = \frac{tp}{tp + fn} = \frac{correctly\ recognized\ positive\ example}{total\ actual\ positive\ example}$$

Equation 3-32 The definition of Recall (Sokolova & Lapalme, 2009) *where tp and fn are defined in Table 3.4.*

Recall is negatively correlated with Precision. Generally, the higher the Precision, the lower the Recall (Tan, Wang, & Lee, 2002). Although high Precision and high Recall are favored, both of them could not be high at the same time. Furthermore, if nothing is classified as positive, the Precision will be infinite, which is very high but not meaningful. Similarly, Recall would be infinite but meaningless if there is no positive example. Thus, we would need some metrics that strike the balance. Some studies use break-even point (BEP) as a measure to compare result, which is the point at which Recall equals Precision (Tan et al., 2002). However, as BEP does not exist in some cases, another useful measure would be Fscore (Tan et al., 2002).

Fscore is the combination of Precision and Recall (Sokolova & Lapalme, 2009), as shown in Equation 3-33. It demonstrates the relationship between the actual positive examples

and the examples labelled as positive by the classifier (Sokolova & Lapalme, 2009). F1 score is one of the common variants of Fscore when $\beta = 1$, it is also the harmonic mean of Precision and Recall as defined in Equation 3-34. As the baseline paper measures the performance of the classifiers using F1 score for both depressed and not depressed classes (Valstar et al., 2016), same performance metric will be chosen for a meaningful and consistent result comparison. Other than that, the mean of the F1 score for both classes is also calculated and recorded to compare the performance of classifiers with the model proposed in related studies. Precision and Recall are also shown to provide a clearer picture about the performance of the classifiers.

$$Fscore = \frac{(\beta^2 + 1)\, tp}{(\beta^2 + 1)\, tp + \beta^2\, fn + fp}$$

Equation 3-33 The definition of Fscore (Sokolova & Lapalme, 2009) where β is a variable defined by the user, whereas tp, fp and fn are defined in Table 3.4.

$$\mathrm{F1} = \frac{2\, tp}{2tp + fn + fp} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Equation 3-34 The definition of F1, which is the case when the β of the Fscore (defined in Equation 3-33) is equal to 1, where Precision is defined in Equation 3-31 and Recall is defined in Equation 3-32, whereas tp, fp and fns are defined in Table 3.4.

In addition, we provide the Accuracy for the prediction of the classifiers to demonstrate the overall effectiveness of the classifier in predicting a group of data (Sokolova & Lapalme, 2009). It is defined as the total correctly recognized examples divided by the total examples in the data. Accuracy is a robust measure of classifier performance because it is invariant to the performance of a specific class (Sokolova & Lapalme, 2009). Performance of the classifier can be visualized more easily by knowing its Accuracy in predicting certain group of data.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} = \frac{\text{total correctly recognized examples}}{\text{total examples in the data}}$$

Equation 3-35 The definition of accuracy (Sokolova & Lapalme, 2009) where tp, tn, fp and fn are defined in Table 3.4.

In short, F1 score, Precision, Recall and Accuracy are provided as performance metrics in this study to provide a full picture of the performance of classifier. However, only F1 score is used in the comparisons of the performances of different classifiers.

3.5.2 Regression

The evaluation measures for regression must be different from the evaluation metrics for classification because the regressors estimate a real number but not a discrete label. Therefore, other evaluation metrics that are defined in this section are RMSE and MAE (Chai & Draxler,

2014). As they are also used by the baseline to evaluate the regressor performance (Valstar et al., 2016), they will be our evaluation metrics to provide meaningful and consistent comparison.

As defined in Equation 3-36, RMSE is a better evaluation metric than MAE as model usually has normally distributed error (Chai & Draxler, 2014). Although sensitive to outliers (Willmott & Matsuura, 2005), RMSE can reconstruct the error distribution closely (Chai & Draxler, 2014). As sum of squared errors is usually the cost function that should be minimized, penalizing large errors using the least-square terms can greatly improve the model performance. Thus, RMSE is able to present model performance better than MAE (Chai & Draxler, 2014).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (predicted\ value - real\ value)^2}$$

Equation 3-36 The definition of Root Mean Square Error (RMSE) (Chai & Draxler, 2014) *where n is the number of samples and e_i is the error of the model.*

MAE is presented in Equation 3-37. Different from RMSE, MAE is better in presenting uniformly distributed errors (Chai & Draxler, 2014). As opposed to RMSE, MAE does not penalize large errors, so it is not sensitive to outliers (Willmott & Matsuura, 2005). It can be more easily illustrated as it represents the average absolute error rate. If the model estimates a score, the actual score ranges between the difference of the score and MAE and their sum.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |e_i| = \frac{1}{n}\sum |predicted\ value - \ real\ value|$$

Equation 3-37 The definition of Mean Absolute Error (MAE) (Chai & Draxler, 2014) *where n is the number of samples and e_i is the error of the model.*

As large error is not preferred in the context of depression regression, RMSE is a better evaluation metric than MAE. However, any evaluation metric only represents the error data partially (Chai & Draxler, 2014). While RMSE is used as the evaluation metric to measure the model performance, MAE is also given to provide different insights of the error distribution. Both RMSE and MAE should be presented to ensure a fair comparison with the baseline model.

3.6 Chapter Summary

This chapter reviews a number of technical knowledges required to understand about this study. Based on the related background knowledge, some new methods will be proposed in the next chapter.

Chapter 4 Proposed Method

This chapter proposes a few techniques, including Audio Feature Standardization and Audio Feature Selection via Complete Search. This study also proposes the use of AB in depression classification and regression. The evaluations of these methods will be provided in Chapter 6.

4.1 Audio Feature Standardization

The use of normalization methods is usually not justified in related studies. Thus, this study aims to provide a review on the effects of the different normalization methods mentioned in section 3.2. Besides, Audio Feature Standardization is proposed as an alternative normalization method. While normalization should be mandatory to all the ML models from Scikit adapted in Section 3.4 (Scikit-learn developers, 2010e), some audio features such as MFCCs are not suitable to be feature warped (Cummins et al., 2011),. Thus, audio feature standardization is proposed to normalize these audio features that are not suitable for feature warping. The concept is simple and similar to the concept of feature warping, except that this proposed method is not treating every feature independently but only treating each audio feature independently. The rationale of only treating each audio feature independently is that the features contained in an audio feature should be somehow dependent to each other and thus should be treated as a whole. If the audio feature contains only one feature, Audio Feature Standardization behaves like Feature Warping, thus it also inherits some good properties of Feature Warping. If the mean and std of these audio features are available, Audio Feature Standardization is applied to them independently.

4.2 Audio Feature Selection via Complete Search

One of the major flaws of the feature selection techniques discussed in Section 3.3 is that they treat every features individually. While all the features contained in an audio feature might be highly dependent on one another, applying them to the data might not be effective as it possibly selects only a subset of the features in the audio feature but not all of them, which consequently reduces the data representation. Thus, features contained in an audio feature should be treated as a whole and removed if and only if the audio feature is found redundant.

As there are 11 audio features which would be explored in this study, if the complete search technique (or commonly named as “brute force method”) is employed, there are only $(2^{11} - 1)$ combinations or subsets of the audio features that have to be experimented, excluding the situation where there is no feature chosen. In other words, as the number of

combinations are still quite few, the complete search technique can be adapted to demonstrate the potential and strength of audio features selection.

4.3 Machine Learning Method

In this study, we propose the use of AB as our ML model because AB is a famous and well-established ML model. However, we have also tried out a method named N-Layer ensemble, which yields a very bad result. Thus, it will be documented in Appendix E and not discussed in detail in this paper.

4.4 Multi-class Classification Task

This study aims to solve a new problem: how to predict the depression level of a patient using speech. This approach benefits the model in real world applications for being more objective and not harsh by simply labeling a person as depressed. It also generalizes the result as it can now be tested by depression audio corpus constructed by using any assessment methods, which properly defines the depression level of a person as its ground truth.

4.5 Chapter Summary

This chapter proposes some new methods such as Audio Feature Standardization and Audio Feature Selection via Complete Search to boost depression prediction accuracy. Next, Chapter 5 would demonstrate the application of background knowledges and the proposed methods.

Chapter 5 Methodology

In this study, we not only predict which person is depressed, but also estimate the severity of depression. Thus, both the classification task and regression task are being addressed. The classification task is performed to predict who the depressed individual is, while the regression task is done to estimate the severity of depression. In the provided dataset, the depression binary classification of an individual (is depressed / not depressed) is given based on the severity of depression which is measured using PHQ-8, as explained in Section 2.2. These predictions are performed under the following assumptions:

1. Depression regression and depression classification are event-based predictions of the level of depression and the severity of depression remains constant over a certain period of time rather than changing at every moment in time (Valstar et al., 2016).
2. The speech signals extracted from different people suffering from depression should share some similarities. For example, diminished, prosodic and monotonous speech is often strongly correlated with depression (Cummins et al., 2015).

This study is conducted under these assumptions. As we are also using the data provided for DCC in AVEC2016, the results of the baseline and related studies are also compared to demonstrate the performance effectiveness of the proposed model (Valstar et al., 2016).

Figure 5.1 provides a brief overview of the study workflow. Firstly, the audio is pre-processed to remove noises in the audios provided and obtain a more accurate representation of the participants' speech in AVEC2016 dataset. Next, audio features are extracted to obtain a numerical representation of the speech and find a pattern within the numerical representation in depression context. Both the necessity of the data normalization and the use of features selection are examined in this study. A list of ML models is experimented with different sets of features, various methods of data normalization and different techniques in features selection. The generated prediction results are then compared, analysed and recorded in this paper. This workflow is generalizable to other ML tasks such as emotion classification and music recognition.

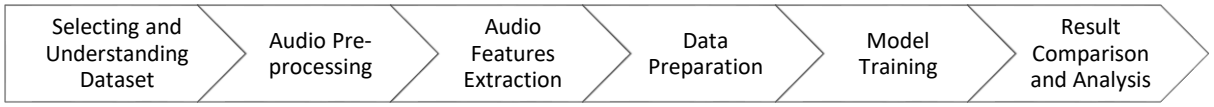

Figure 5.1 An Overview Workflow of the Experiment.

This study is implemented in python using the famous and effective python open-source libraries including numpy, scipy, matplotlib, Scikit-learn, Scikit-feature, GPy and GPyOpt.

5.1 Selecting and Understanding Dataset

Inspired by the AVEC2016 challenge, we decided to work on the database suggested by the organizer, known as the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) Depression Database. DAIC-WOZ is part of a larger corpus, which is the Distress Analysis Interview Corpus (DAIC) (Valstar et al., 2016). Created by USC, DAIC contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression and PTSD (Gratch et al., 2014).

Figure 5.2: Face-to-face interview setup (Left); Ellie, the virtual interviewer (Right) (Gratch et al., 2014).

The corpus data comprises of audio recordings, video recordings, PHQ-8 scores and the ground truth indicating whether the participant is depressed (Gratch et al., 2014). The audio and video recordings are collected through 4 different types of interview, as shown in Table 5.1.

Interview Type	Description
Face-to-face	A human interviewer interviews the participants face-to-face with the setup shown in Figure 5.2 (left).
Teleconference	A human interviewer interviews the participants through phone calls.
Wizard-of-Oz	A human-controlled agent, Ellie, interviews the participants face-to-face, as shown in Figure 5.2 (right).
Automated	An autonomous agent, Ellie, interviews the participants face-to-face, as shown in Figure 5.2 (right).

Table 5.1 Different types of interviews to collect audio and video recordings (Gratch et al., 2014).

In the dataset provided by AVEC2016, audio and video features are computed using methods described in Section 2.1.1 before being distributed to the participants in AVEC2016.

5.1.1 Data Distribution

There are 107 audios provided in the training (train) set, 35 audios given in development (dev) set and 47 audios contained in the test set. As the ground truth of the audios in test set are not provided, they are not used in this study. Among the 107 train samples, there are 86 non-

depressed samples and 21 depressed-samples. In the dev set, there are 28 non-depressed samples and 7 depressed samples. The PHQ-8 score distribution in the train set is illustrated in Figure 5.3 (left) whereas the distribution in the dev set is shown in Figure 5.3 (right).

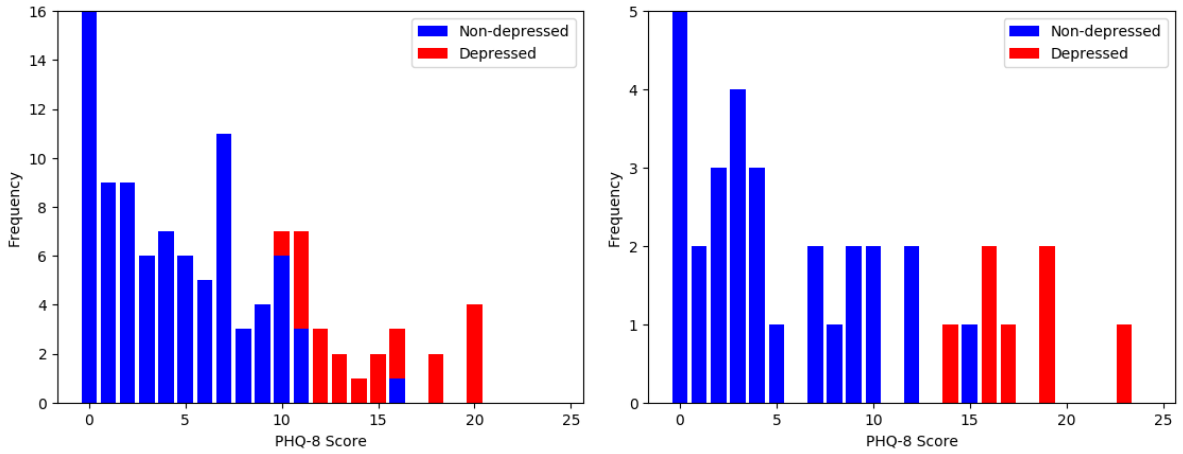

Figure 5.3 PHQ8 Score Distribution in Train Set (Left); PHQ8 Score Distribution in Dev Set (Right)

On the other hand, the depression level distribution in the train set is illustrated in Figure 5.4 (left) whereas the distribution in the dev set is shown in Figure 5.4 (right).

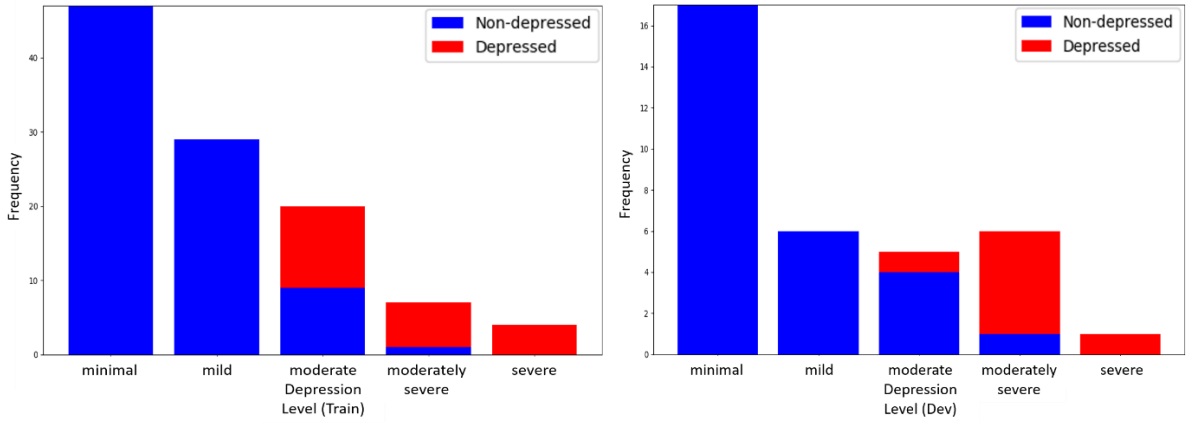

Figure 5.4 Depression level distribution in Train Set (Left); Depression level distribution in Dev Set (Right)

5.1.2 Depressed and Non-Depressed Samples

As the depression data is highly correlated to the accuracy of the ML model constructed, a deep understanding of the data is necessary. To further understand the difference of depressed and non-depressed data, a depressed audio sample (participant 303) and a non-depressed audio sample (participant 319) in the train set are selected and analyzed. Their audios are pre-processed using the method that will be explained in section 5.2 before plotting the first 0.2 seconds of the audio signals. The non-depressed audio signal is plotted in Figure 5.5 (left) whereas the depressed audio signal is displayed in Figure 5.5 (right). Obviously, the depressed

audio sample shown in Figure 5.5 (right) is more monotonous than the non-depressed audio sample illustrated in Figure 5.5 (left). Thus, monotonous speech is an indicator of depression as mentioned in Chapter 1 and detecting depression using speech is definitely possible.

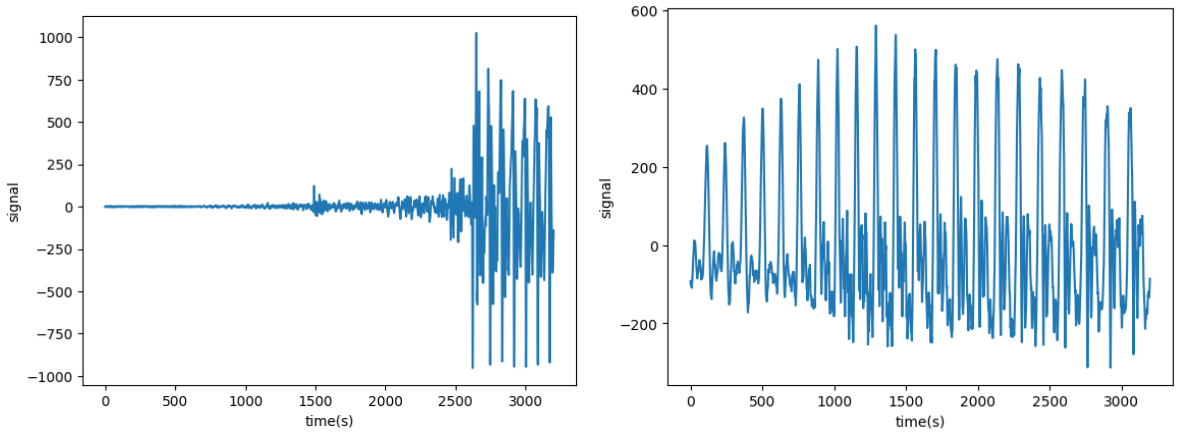

Figure 5.5 The First 0.2 seconds of the Non-Depressed Audio Signal (Left); The First 0.2 seconds of the Depressed Audio Signal (Right)

5.1.3 Section Overview

This section explains the reason of choosing the depression analysis corpus, discusses about the data collection method, illustrates the distribution of the data in the AVEC2016 dataset and demonstrates the difference between depressed speech signal and non-depressed speech signal.

5.2 Audio Pre-processing

The original audio contains many silence intervals caused by the equipment set up time for the interview. Besides, the original audio includes the voice of the interviewer, Ellie. There are also some noises in the environment that should be removed. Therefore, before extracting the audio features, the audio are pre-processed first to obtain a more accurate and reasonable result.

5.2.1 Noise Reduction

The performance of the ML system could be improved by removing background noises of the audio recordings (Pohjalainen, Fabien Ringeval, Zhang, & Schuller, 2016). There are also many third-party tools that provide the noise reduction feature, such as Sound eXchange (SoX) and Audacity. FFmpeg (Bodecs et al., 2016) and SoX (Bagwell, 2014) are tools that provide a way to automate the process of reducing the background noise in audios (Babin, 2011).

In this project, FFmpeg is used to extract a silence segment of the audio recording which contains noise (Babin, 2011). This type of segment is ideal for SoX to use in the noise reduction process, which can be easily found at the start or the end of the audio recordings (Bagwell,

2014). In fact, the first second of the audio recordings we obtained from the DAIC-WOZ is this type of segment, which is extracted by FFmpeg to be the noise sample.

Next, the noise sample is input into SoX to generate a noise profile (Babin, 2011), which is a representation of the audio recording that contains data of the consistent background noises, such as hiss or hum. While SoX is used to reduce such noises to produce a cleaner audio, the amount of noise removed is chosen to be 0.21 as suggested by Babin (2011). However, the noise reduction method provided by SoX is only moderately effective in removing consistent background noises, implying that more noises could be reduced if a better approach is available. Since most of the related studies that used the data provided by DAIC-WOZ did not perform noise removal, we can reasonably assume that the quality of the original audio is actually acceptable. Thus, the quality of the cleaned audio recordings after noise removal is deemed acceptable and a more effective noise reduction technique is not required. A Windows batch file named cleanaudio.bat is written to automate the noise reduction process.

5.2.2 Obtaining Speech Segments

After reducing some noises from the original audio recordings, the cleaned recordings still consist of the speech segments of both the participants and the interviewer, Ellie. Since Ellie's voice does not relate to the severity of the depression of the participants, her speech segments would add noise to the audio recordings. Fortunately, the transcripts of all audio recordings are provided in Comma-separated values (csv) files, which the speech segments of different speakers are provided with a time frame. Therefore, we could obtain the speech segments of the participants based on the time frames given in the transcripts, as shown in the sample in Table 5.2. FFmpeg is used to split the audio segments and a python script named "extractUserSpeech.py" is written to automate this process.

Start time	Stop time	Speaker	Value
60.028	61.378	Ellie	How are you doing today?
62.328	63.178	Participant	Good.

Table 5.2 Sample of the transcripts for the audio recordings provided in DAIC-WOZ.

After extracting speech segments of the participants into different audio recordings, the speech segments of the same participant are then combined using the functions provided by SoX. A python script named "combineAudio.py" is written to automate this process.

5.2.3 Section Overview

This chapter demonstrates the use of SoX and FFmpeg to pre-process the audios provided in the AVEC2016 dataset to remove the noise and unrelated speech segments. At the end, only the speech segments of the participants in AVEC2016 are kept and merged into a speech audio.

5.3 Audio Features Extraction

We extract MFCCs only using a Java MFCCs feature extractor, which is implemented based on the extraction algorithm attached in Appendix A. However, an open-source python library named "pyAudioAnalysis" that could perform features extraction is used to try out more audio features commonly used by others. As it claimed that it has been used in depression classification (Giannakopoulos, 2015), we decide to apply the features extracted by "pyAudioAnalysis" to the audio recordings from DAIC-WOZ. The detailed explanation of all audio features are provided in Section 3.1 and summarized in Section 3.1.4.

Audio features are extracted at a frame size of 50ms with 50% overlap and automated by a windows batch file named "usingPyAudioFeature.bat". A feature file containing audio features for a signal window each line is generated by pyAudioAnalysis. According to our assumption stating that depression severity remains constant over a certain period of time rather than changing at every moment in time, we assume that the speech signal should not vary a lot for depression and average of the features in different time frames can be taken. Thus, we take the mean of audio features from the feature file to obtain only one line of audio feature per file.

As we had split the audios provided in DAIC-WOZ into speech segments of the participants using the transcript provided, features for different speech segments are extracted to train the classifiers. To achieve a better efficiency, we also combine the audios of different audio segments of a participant into one audio, then average the features in different time frames to get only one data per participant. The results would be shown in the following sections. Different optimization techniques are also applied to improve the result. However, having only the mean of the feature would discard the changes of the audio features in the whole interview, causing some audio features such as Energy to lose meaning. Thus, in the later experiments, the std of audio features are also included.

5.4 Data Preparation

This process prepares the training data to be optimal to train the ML models. Both Data Normalization and Feature Selection are part of the Data Preparation. In this study, we perform

Audio Feature Selection via Complete Search to find the optimal subset of audio features. Then, we try to improve the result by normalizing the data and removing more redundant features.

5.4.1 Data Normalization

For data normalization, **Feature Warping** can be implemented using the "scale" function in the "preprocessing" module provided by Scikit, while **Min-max Scaling** can be implemented using "MinMaxScaler" or "MaxAbsScaler" in the same module (Scikit-learn developers, 2010e). Meanwhile, the proposed normalization method, Audio Feature Standardization, is implemented using python in classifierWithFS.py as a function called "Standardized". Audio features passed in the function would be standardized to normal distributed audio features.

5.4.2 Feature Selection

This study utilizes feature selection algorithms from an open-source feature selection library named "Scikit-Feature", built on top of famous python libraries such as Scikit and implemented based on many recent state of the art feature selection algorithms (J. Li et al., 2016). However, the feature selection techniques provided by Scikit are not used given the reason provided in Appendix C. For feature selection techniques designed for generic data and implemented by Scikit-Feature, their effectiveness is investigated when applied to depression audio features dataset (J. Li et al., 2016). These techniques include Relief, Fisher, CIFE, CMIM, MRMR, MIFS, ICAP, FCBF, CFS and GINI. The feature selection methods are set up in the preprocessFeatSelection(.) function in classifierWithFS.py. Meanwhile, Audio Feature Selection via Complete Search is also implemented using an exhaustive recursive search method named "automaticChooseFeatures(.)" implemented in classifierGenerateTable.py.

5.5 Model Training

This study implements the reviewed classifiers in python using Scikit-learn toolbox because it provides a large variety of supervised and unsupervised ML algorithms (Pedregosa et al., 2012), including a lot state-of-the art classifiers. It is also easy to use and focuses on computational efficiency (Pedregosa et al., 2012). In this project, all existing hyper-parameters or parameters for the implemented ML models are usually set as the default value provided by Scikit. Every Scikit classifier is capable in performing multi-class classification by default, although OVR and OVO are also applied to them to observe the effect (Scikit-learn developers, 2010b). In Scikit, OVR and OVO could be applied to the models by wrapping them around OneVsRestClassifier and OneVsOneClassifier respectively (Scikit-learn developers, 2010b).

Figure 5.6 shows the methods adapted to choose the best ML models, either classifiers or regressors. ML model is trained using the data contained in the train set and its performance is evaluated using the data extracted from the dev set. The performances of the best ML model trained using certain subsets of audio features with or without normalization and feature selection are tabulated and compared. The model which provides a good performance based on evaluation metrics and is justifiable by the observations in depression context would be chosen and proposed to be used as the prototype that helps diagnosing depression using speech. Similarly, this workflow could be extended for the use of other ML tasks.

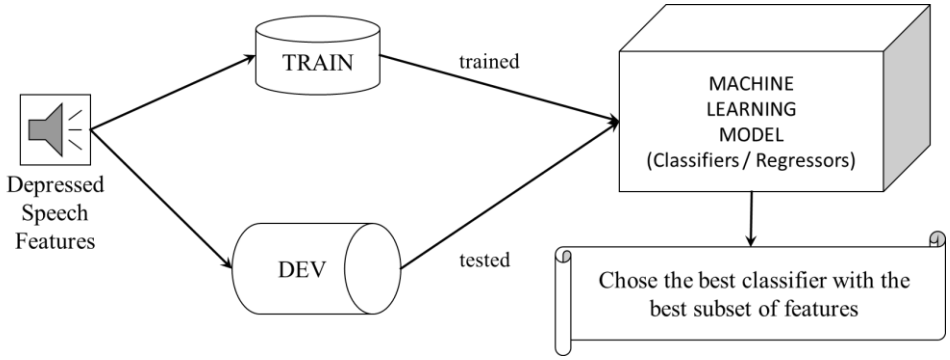

Figure 5.6 ML Model Selection Workflow

5.6 Chapter Summary

This chapter discusses the process to set up and run the experiment conducted in this study. After understanding the selected dataset, audios are pre-processed before extracting the audio features. Then, the extracted features are normalized while applying feature selection methods to them. The processed data is then passed into ML models to evaluate the models performance based on the metrics chosen. The result and discussion would be provided in Chapter 6.

Chapter 6 Result and Discussion

The depression binary classification and regression task are the main focuses in this study. In the former task, we aim to accurately predict whether an individual is depressed, whereas the latter task estimates the severity of depression. Both proposed classifier and regressor manage to surpass the benchmark and most related studies. After Audio Pre-processing from section 5.2, we will explore how many audio features should we extract and whether should we perform **Data Normalization** and **Feature Selection** to identify the best ML models. Most of the ML models that generate the results in this chapter, either classifiers of regressors, are trained by audio features extracted using pyAudioAnalysis. However, the result in Section 6.1.1 is generated using MFCCs extracted by Java Extractor. Using the supervised ML techniques provided in Section 3.4, the good results are analyzed and discussed in this section. The performance of ML model is not stated if it is unable to classify depression instances correctly. Overviews of experiments and tasks are given in Table 6.1 and Table 6.2 respectively.

Section	Audio Features	Statistical Descriptors	Normalization	Feature Selection	Summary
6.1.1	MFCCs	Mean	None	None	Better to use one feature row to represent each participant
6.1.2	All Combinations	Mean	None	None	We need more audio features. Best result for mean audio feature are provided.
6.1.3	All Combinations	Mean + Std	None	None	Std included to capture the variation of audio features.
6.1.4	All Combinations	Mean + Std	Normalized	None	Normalizing is mandatory for many ML models.
6.1.5	All Combinations	Mean + Std	None	All except Relief	Feature Selection Techniques help identifying redundant features.
6.1.6	All Combinations	Mean + Std	Normalized	All except Relief	FS is unable to improve the performance of normalized audio features.
6.1.7	Best of Normalized and not Normalized	Mean + Std	Both	All except Relief	Combination of mean and mean + std is found to be useless.
6.1.8	Best of mean and mean + std	Mean, Mean + Std	None	All except Relief	Combination of mean and mean + std is found to be useless.
6.1.9	All Combinations	Mean	Normalized	None	Best normalized mean result is also obtained.
Main deliverables: Optimal audio feature set for mean, mean + std, normalized mean, normalized mean + std are obtained.					

Table 6.1 Overview of experiments, whereby the term "all combinations" stated under audio features implies the use of Audio Feature Selection via Complete Search technique, thus it is not included in the feature selection column. Besides, the term "normalized" represents Audio Feature Standardization.

Section	Tasks	Objective	Baseline	Best (Audio)	Best (All)
6.1	Binary Classification	To identify if a person is depressed.	Mean F1 of 0.5	Mean F1 of 0.73	Mean F1 of 0.81
6.2	Regression	To predict the PHQ-8 score (depression severity) of individuals.	RMSE of 6.74	RMSE of 6.38	RMSE of 5.31
6.3	Multi-class Classification	To predict the depression level of an individual.	None	None	None

Table 6.2 Overview of tasks in this study.

6.1 Depression Binary Classification

In the depression binary classification task, we aim to predict whether an individual is depressed accurately. The depression binary classification task is a binary classification task which classifies the participants' speech into two classes: depressed and not depressed. This section will describe the efforts done to improve the classification result based on mean F1. The Precision, Recall and F1 score for each class and its Accuracy are also given as references.

6.1.1 Combination of speech segments

Previously, MFCCs are extracted using our own algorithm mentioned in Appendix A to train our classifiers. In the following sections, audio features are extracted using pyAudioAnalysis. The rationale for starting the study by using only MFCCs is because it is a good audio feature used in many related studies. Thus, only the mean of MFCCs across the time is examined in this section. Besides, the data is not normalized nor feature selected to reduce the factors affecting the result. As details of audio feature extraction are given in Section 5.3, only the necessity of combining speech segments are discussed here.

As every audio segment is short (ranging from 0.43 seconds to 7.73 seconds) and the features of different time frames of a speech segment should not greatly vary, a feature row can be used to represent an audio segment. Table 6.3 shows the result achieved using the supervised ML techniques without data normalization. The best classifier is AB with the highest F1 score for depressed class (0.16). Classifiers with the Recall value of 0 for depressed class are omitted.

Classifiers	F1	Precision	Recall	Accuracy
AB	**0.16(0.83)**	**0.11(0.93)**	**0.34(0.75)**	**71%**
KNN	0.15(0.83)	0.10(0.95)	0.40(0.75)	73%
Decision Tree	0.02(0.83)	0.01(0.96)	0.07(0.73)	71%

Table 6.3 Result with features (MFCCs) for every audio segments in AY2016/17 Semester 1. The value which is not in the bracket is the value for depressed class and the value which is in the bracket is the value for not depressed class. The best result is bolded.

After combining all the speech segments of a participant into one audio, we have only one feature row per participant. According to our assumption, this might produce a better result

as more data might lead to overfitting. This section uses classifiers and optimization technique mentioned in Section 3.4. The results produced by classifiers are recorded in Table 6.4. The results for classifier and ensembles with the Recall value of 0 for depressed class are omitted.

AB gives the best result in F1 score for depressed class (0.29), which slightly improved from the previous result. This proves that our assumption holds and works well for the depression corpus we used. Nevertheless, the result of classifier with optimization is not better than AB. Thus, the optimization technique is not used again in the later stage of the project.

Classifier	F1	Precision	Recall	Accuracy
AB	**0.29(0.82)**	**0.29(0.82)**	**0.29(0.82)**	**71%**
RBF SVM (With Bayesian Optimization)	0.13(0.76)	0.14(0.75)	0.13(0.78)	63%

Table 6.4 Result produced using classifiers with features (MFCCs) with one feature row per participant in AY2016/17 Semester 1. The value which is not in the bracket is the value for depressed class and the value which is in the bracket is the value for not depressed class. The best result is bolded.

6.1.2 Audio Features of pyAudioAnalysis

Starting from this section, all audio features are extracted by pyAudioAnalysis. This section starts by using only the mean MFCCs extracted by pyAudioAnalysis to train the classifiers to compare the result provided in Section 6.1.1 and then trying the combinations of different audio features to improve the result. As AB performs consistently with F1 score of 0.29(0.82), it makes sense to try out all audio features provided by pyAudioAnalysis.

Audio Feature Selection via Complete Search technique is also used to explore combinations of audio features that contribute to better model performance. Table 6.5 shows the best 3 combinations which produce the best F1 mean after fitting into one of the classifiers.

Audio Feature	No. of Audio Features	No. of Features	Best F1 mean	Best F1 score	Best ML Model
Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation	**6**	**18**	**0.82**	**0.71(0.93)**	**AB(All)**
Zero-crossing Rate, Energy, Entropy of Energy, Spectral Rolloff, MFCCs, Chroma Deviation	6	18	0.80	0.67(0.93)	AB(All)
Energy, Entropy of Energy, Spectral Rolloff, MFCCs, Chroma Deviation	5	17	0.80	0.67(0.93)	AB(All)

Table 6.5 The best 3 model performance based on mean F1 trained by 9 different combinations of mean audio features. (Partial, see full on Appendix E)

6.1.3 Inclusion of the Std of Audio Features

In the previous sections, this study explored only mean audio features of the audio feature files generated by pyAudioAnalysis. This section would discuss the necessity to include the std of audio features by exploring the result provided using each combination of audio features.

While the performance of mean audio features is very promising, some of them such as Energy might become redundant if only the mean is taken. Intuitively, capturing mean of the Energy only records the mean of the loudness of individuals' voice, which is meaningless in the depression context. Likewise, taking only the mean of the audio feature would remove the consideration about the changes of the features which are important in this context. To address this issue, std of the audio features can be added because it captures the changes of the audio features. Hence, the classifiers performances trained by the mean and std of audio features are generated and recorded in Table 6.6. AB trained by "Spectral Centroid, Spectral Spread and Spectral Flux" performs the best by giving the F1 mean of 0.80, but it performs worse than the best classifier in Table 6.5. More factors that should be considered would be discussed in the following sections, such as the need of other feature selection techniques and normalization.

Audio Feature Used	No. of Audio Features	No. of Features	Best F1 mean	Best F1 score	Best ML Model
Spectral Centroid, Spectral Spread, Spectral Flux	**3**	**6**	**0.80**	**0.67(0.93)**	**AB(All)**
Spectral Centroid, Spectral Spread, Chroma Deviation	3	6	0.79	0.67(0.91)	AB(All)
Energy, Spectral Spread, Spectral Flux, MFCCs	4	32	0.73	0.57 (0.89)	AB(All)

Table 6.6 The best 3 model performances based on mean F1 trained by mean and std of audio features. (Partial, see full on Appendix E)

6.1.4 Normalization Techniques Application

Input data which contains only the first and second bin of MFCCs are extracted using pyAudioAnalysis, normalized using audio feature standardization, mix-max scaling and feature warping technique, then plotted into Figure 6.1. The data distribution is better preserved after performing audio feature standardization compared to min-max scaling or feature warping. Data distribution is a key factor in choosing normalization method as it is very important for some audio features.

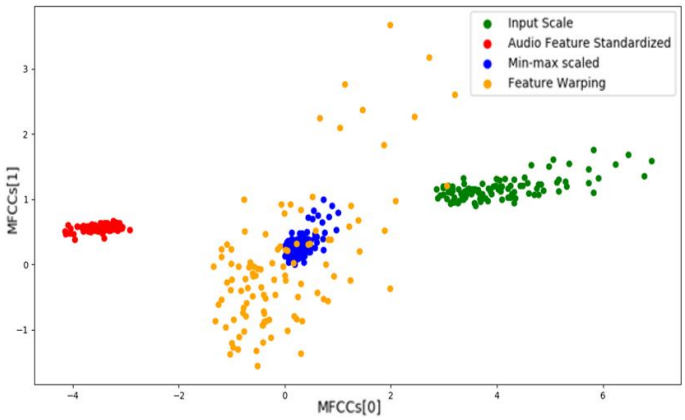

Figure 6.1 Illustration of Different Normalization Methods.

The data which contains only MFCCs is also normalized and fit into the AB classifier as it provides consistently good result. The performances of the classifier correspond to the data which is normalized by different normalization techniques are recorded and tabulated in Table 6.7. The performance of AB improves after applying Audio Features Standardization to the training data, thus this normalization technique is proven useful for MFCCs.

Normalization Technique	Mean F1	F1	Precision	Recall	Accuracy
Input Data without normalization	0.53	0.2 (0.87)	0.33 (0.81)	0.14 (0.93)	77%
Audio Features Standardized Data	0.62	0.36 (0.88)	0.5 (0.84)	0.29 (0.93)	80%
Min-max scaled Data	0.46	0.14 (0.79)	0.14 (0.79)	0.14 (0.79)	66%
Features Warped	0.51	0.18 (0.85)	0.25 (0.81)	0.14 (0.89)	74%

Table 6.7 The Performances of AB Trained by Data Normalized by Different Methods.

The proposed normalization method preserves the data distribution better than other existing methods and inherits the good property of feature warping if the audio feature is only one dimensional. Proven to be useful in standardizing MFCCs, its effect can be tried out on other audio features. As shown in Table 6.8, Audio Feature Standardization has negative effects on Zero-Crossing Rate, Spectral Centroid, Spectral Spread, Spectral Flux and Spectral Rolloff based on the performance of AB. Nonetheless, if multiple classifiers are used to measure the effect of Audio Feature standardization as demonstrated in Table 6.9, the technique no longer has effect on Zero-crossing Rate and Spectral Rolloff. This suggests the need of using multiple classifiers in classifying multiple audio features, but this possibility is not explored in this study.

Audio Features	F1 Score (Without Normalization)	F1 (With Audio Feature Standardization)
Zero-Crossing Rate	**0 (0.87)**	0 (0.85)
Energy	0 (0.87)	**0 (0.88)**
Entropy of Energy	0 (0.77)	**0.36 (0.88)**
Spectral Centroid	**0.18 (0.85)**	0 (0.79)
Spectral Spread	**0.18 (0.85)**	0.14 (0.78)
Spectral Entropy	0 (0.83)	**0.25 (0.78)**
Spectral Flux	**0.33 (0.86)**	0.18 (0.85)
Spectral Rolloff	**0 (0.85)**	0 (0.81)
MFCCs	0.2 (0.87)	**0.36 (0.88)**
Chroma Vector	0.2 (0.87)	0.2 (0.87)
Chroma Deviation	0 (0.79)	**0.13 (0.76)**

Table 6.8 Effect of Audio Feature Standardization on individual Audio Feature based on AB performance.

As suggested in Table 6.9, Entropy of Energy, Spectral Entropy, MFCCs, Chroma Vector and Chroma Deviation are more beneficial to the classifiers after applying Audio Feature Standardization.

Audio Features	F1 Score (Without Normalization)	Classifier	F1 (With Audio Feature Standardization)	Classifier
Zero-Crossing Rate	0 (0.89)	GP-DP	0 (0.89)	GP-DP
Energy	0 (0.89)	GP-DP	0 (0.89)	GP-DP
Entropy of Energy	0 (0.89)	GP-DP	**0.36 (0.88)**	**AB**
Spectral Centroid	**0.18 (0.85)**	**AB**	0 (0.89)	GP-DP
Spectral Spread	**0.18 (0.85)**	**AB**	0.14 (0.78)	AB
Spectral Entropy	0 (0.89)	GP-DP	**0.25 (0.78)**	**AB**
Spectral Flux	**0.33 (0.86)**	**AB**	0.25 (0.90)	KNN
Spectral Rolloff	0 (0.89)	GP-DP	0 (0.89)	GP-DP
MFCCs	0.2 (0.87)	AB	**0.36 (0.88)**	**AB**
Chroma Vector	0.2 (0.87)	AB	**0.25 (0.90)**	**GP-DP**
Chroma Deviation	0 (0.89)	GP-DP	**0.13 (0.76)**	**AB**

Table 6.9 Effect of Audio Feature Standardization on individual Audio Feature based on performance of different classifiers.

Besides, we examined the effect of Audio Feature Standardization on different audio features combinations and tabulated the classifiers performances are in Table 6.10. The best classifier performance is given by F1 score of 0.53 (0.87), which is a drop from the previous best classifier performance. Interestingly, the subsets with all positively correlated audio features and the subsets without any negatively correlated audio features are not providing the best performance across all of the combinations. To further improve the result, the effect of feature selection techniques on these combinations of audio features would be investigated.

Audio Feature Used	No. of Audio Features	No. of Features	Best F1 mean	Best F1 score	Model
Entropy of Energy, Spectral Centroid, Spectral Spread, Spectral Flux	4	8	0.70	0.53 (0.87)	AB
MFCCs, Chroma Vector	2	50	0.68	0.44 (0.92)	GP-DP
MFCCs, Chroma Vector, Chroma Deviation	3	52	0.68	0.44 (0.92)	GP-DP
Spectral Entropy, Spectral Rolloff, MFCCs	3	30	0.68	0.44 (0.92)	KNN
Energy, MFCCs, Chroma Vector	3	52	0.68	0.44 (0.92)	GP-DP
Energy, MFCCs, Chroma Vector, Chroma Deviation	4	54	0.68	0.44 (0.92)	GP-DP
Zero-crossing Rate, Energy, Chroma Vector, Chroma Deviation	4	30	0.68	0.44 (0.92)	Linear SVM
Zero-crossing Rate, Energy, MFCCs, Chroma Vector, Chroma Deviation	5	56	0.68	0.44 (0.92)	Linear SVM
Zero-crossing Rate, Energy, Entropy of Energy, Spectral Entropy, Spectral Rolloff, MFCCs, Chroma Vector, Chroma Deviation	8	62	0.58	0.25 (0.90)	AB
Entropy of Energy, Spectral Entropy, MFCCs, Chroma Vector, Chroma Deviation	5	56	0.55	0.22 (0.89)	Linear SVM

Table 6.10 10 Audio Features Combinations include the best 8 classifiers performances, the combination of all positively correlated audio features and the combination without any negatively correlated audio features. Those combinations which contain audio features that would have negative effect after Audio Feature Standardization are not included. (Partial, see full on Appendix E)

6.1.5 Feature Selection Techniques Application

As mentioned in section 5.4.2, feature selection methods other than Audio Feature Selection via Complete Search are investigated. The rationale of applying these techniques is because after using mean and std of audio features, we are not sure if all of them are useful. In fact, some of them might only be useful when std of audio feature is provided and vice versa. Thus, these techniques should be tried even though they do not treat audio features as a whole.

The effects of these feature selection techniques are tested on different classifiers using test set and are tabulated in Table 6.11. Interestingly, these states of the art ML techniques manage to improve the performance of classifiers even though they might not be able to treat every audio feature differently. The best performance is given by AB (Relief) with F1 score of 0.36 (0.88), followed by KNN (CIFE) with F1 score of 0.25 (0.90).

	F1 of Optimal Feature Subset										
Classifier	**All**	**Fisher**	**Relief**	**MIFS**	**MRMR**	**CIFE**	**CMIM**	**ICAP**	**FCBF**	**CFS**	**GINI**
GP	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)
Linear SVM	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)
RBF SVM	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)
KNN	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.85)	0 (0.85)	0.25 (0.90)	0 (0.85)	0 (0.85)	0.25 (0.90)	0 (0.89)	0 (0.89)
DT	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)
RF	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)	0 (0.89)
AB	0.14 (0.79)	0.14 (0.79)	**0.36 (0.88)**	0 (0.79)	0 (0.79)	0.18 (0.85)	0 (0.79)	0 (0.79)	0 (0.87)	0.33 (0.77)	0.14 (0.79)

Table 6.11 Effect of Feature Selection Techniques on different classifiers using Train Set where "All" means no feature selection method is applied.

However, as Relief searches for the optimal subset using randomly picked samples as mentioned in section 3.3.1, it provides inconsistent performance as shown in Table 6.12. Consequently, the most optimal feature set achieved by applying Relief should be recorded and it is not suitable to be directly compared with other feature selection techniques.

		Relief Iteration					
Classifier	**Audio Feature**	**1**	**2**	**3**	**4**	**5**	**6**
AB	**All**	0.36 (0.88)	0.25 (0.90)	0 (0.87)	0.2 (0.87)	0.22 (0.89)	0.33 (0.86)

Table 6.12 Inconsistent performance of AB trained by subset of train data extracted using Relief algorithms in different iterations.

In contrast to Audio Feature Selection via Complete Search, these feature selection techniques perform worse as they fail to find the most optimal audio feature subset even though

they are significantly faster. This proves the need of extensive search for the optimal subset of Audio Features. Yet, it is still useful to employ them to further eliminate redundant features.

Table 6.13 shows the best 8 classifier performances trained by the data applied to different feature selection techniques. As the audio feature used in ID 1 is the subset of audio feature combinations ID 2 to 8 and their good result is achieved using the CFS technique, the audio features in ID 1 are the most important among all of the mean and std of audio features.

ID	Crucial Audio Feature	Redundant Audio Feature	Best F1 mean	Best F1 score	Model (Feature Selection)
1	**Spectral Centroid, Spectral Spread, Spectral Flux**	**-**	**0.80**	**0.67 (0.93)**	**AB(All)**
2		Entropy of Energy	0.80	0.67 (0.93)	AB(CFS)
3		Energy	0.80	0.67 (0.93)	AB(CFS)
4		Energy, Entropy of Energy	0.80	0.67 (0.93)	AB(CFS)
5		Zero-crossing Rate	0.80	0.67 (0.93)	AB(CFS)
6		Zero-crossing Rate, Entropy of Energy	0.80	0.67 (0.93)	AB(CFS)
7		Zero-crossing Rate, Energy	0.80	0.67 (0.93)	AB(CFS)
8		Zero-crossing Rate, Energy, Entropy of Energy,	0.80	0.67 (0.93)	AB(CFS)

Table 6.13 Effect of Feature Selection Techniques except Relief applied to the mean and std of audio features based on the performance of classifiers. (Partial, see full on Appendix E)

6.1.6 Applying Feature Selection to Normalized Data

Similarly, feature selection methods are applied to normalized mean and std of audio features from Section 6.1.4, but Table 6.14 demonstrates that the results are not improving.

Audio Feature Used	No. of Audio Features	No. of Features	Best F1 mean	Best F1 score	Model
Entropy of Energy, Spectral Centroid, Spectral Spread, Spectral Flux	4	8	0.70	0.53 (0.87)	AB (All)
MFCCs, Chroma Vector	2	50	0.68	0.44 (0.92)	GP-DP (All)
MFCCs, Chroma Vector, Chroma Deviation	3	52	0.68	0.44 (0.92)	GP-DP (All)
Spectral Entropy, Spectral Rolloff, MFCCs	3	30	0.68	0.44 (0.92)	KNN (All)
Energy, MFCCs, Chroma Vector	3	52	0.68	0.44 (0.92)	GP-DP (All)
Energy, MFCCs, Chroma Vector, Chroma Deviation	4	54	0.68	0.44 (0.92)	GP-DP (All)
Zero-crossing Rate, Energy, Chroma Vector, Chroma Deviation	4	30	0.68	0.44 (0.92)	Linear SVM (All)
Zero-crossing Rate, Energy, MFCCs, Chroma Vector, Chroma Deviation	5	56	0.68	0.44 (0.92)	Linear SVM (All)
Entropy of Energy, Spectral Entropy, MFCCs, Chroma Vector, Chroma Deviation	5	56	0.62	0.36 (0.88)	Linear SVM (MRMR)
Zero-crossing Rate, Energy, Entropy of Energy, Spectral Entropy, Spectral Rolloff, MFCCs, Chroma Vector, Chroma Deviation	8	62	0.58	0.25 (0.90)	AB (All)

Table 6.14 Effect of Feature Selection Techniques except Relief applied to the normalized mean and std of audio features based on the performance of classifiers. (Partial, see full on Appendix E)

6.1.7 Combination of Normalized and Not Normalized Audio Features

As shown in section 6.1.5, Spectral Centroid, Spectral Spread and Spectral Flux are the most important audio features across the mean and std audio features. However, their performance would reduce if it is Audio Feature Standardized, as shown in section 6.1.4. This inspires us to think about the effects when all the audio features except the three features are normalized or when they are combined with the normalized positively correlated audio features. Audio Feature Combinations mentioned in Table 6.14 with good performance which do not include the three features are also tested. As shown in Table 6.15, these model performances are capped at the Mean F1 of 0.80, which is not improving and is the same performance of model using only Spectral Centroid, Spectral Spread and Spectral Flux. Thus, they are not proposed.

Audio Feature Without Normalization	Normalized Audio Feature	Best F1 mean	Best F1 score	Model
Spectral Centroid, Spectral Spread, Spectral Flux	Zero-crossing Rate, Energy, Entropy of Energy, Spectral Entropy, Spectral Rolloff, MFCCs, Chroma Vector, Chroma Deviation	0.80	0.67 (0.93)	AB (CFS)
	Entropy of Energy, Spectral Entropy, MFCCs, Chroma Vector, Chroma Deviation	0.80	0.67 (0.93)	AB (CFS)
	MFCCs, Chroma Vector	0.80	0.67 (0.93)	AB (CFS)
	MFCCs, Chroma Vector, Chroma Deviation	0.80	0.67 (0.93)	AB (CFS)
	Spectral Entropy, Spectral Rolloff, MFCCs	0.80	0.67 (0.93)	AB (CFS)
	Energy, MFCCs, Chroma Vector	0.80	0.67 (0.93)	AB (CFS)
	Energy, MFCCs, Chroma Vector, Chroma Deviation	0.80	0.67 (0.93)	AB (CFS)
	Zero-crossing Rate, Energy, Chroma Vector, Chroma Deviation	0.80	0.67 (0.93)	AB (CFS)
	Zero-crossing Rate, Energy, MFCCs, Chroma Vector, Chroma Deviation	0.80	0.67 (0.93)	AB (CFS)

Table 6.15 Combination of non-normalized and normalized audio features.

6.1.8 Combination of Section 6.1.3 and Section 6.1.5

The best combination of mean Audio Features are Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation, whereas the best combination of mean and std of audio features are Spectral Centroid, Spectral Spread and Spectral Flux. They are combined to check if better result could be obtained. However, as displayed in Table 6.16, the result does not improve as expected. Hence, we are not using this audio features set.

Mean And Std Audio Features	Mean Audio Feature	Best F1 mean	Best F1 score	Model
Spectral Centroid, Spectral Spread, Spectral Flux	Zero-crossing Rate, Entropy of Energy, Spectral Entropy, MFCCs, Chroma Deviation	0.80	0.67 (0.93)	AB (CFS)
Spectral Spread	Zero-crossing Rate, Entropy of Energy, Spectral Entropy, MFCCs, Chroma Deviation	0.76	0.62 (0.91)	AB (All)
Spectral Centroid, Spectral Flux	Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation	0.68	0.44 (0.92)	KNN (FCBF)

Table 6.16 Combination of mean audio features and std audio features.

6.1.9 Normalization of Mean Audio Features

The best combination of mean Audio Features are normalized and fitted into AB. The performance given by this AB is mean F1 of 0.76 and F1 score of 0.62 (0.91). All other combinations of audio features are also normalized and fitted into the reviewed classifiers and the results are tabulated in Table 6.17. Clearly, option 1 and 2 provide the best performance among all of the combinations. However, as option 1 is a subset of option 2, option 1 would be stronger and thus is proposed as the best audio feature set for normalized mean audio features. As we are only using Mean Audio Features, no feature selection techniques are applied.

Option	Feature Used	Mean F1	F1 score	Model
1	Spectral Centroid, Spectral Entropy, Spectral Rolloff, MFCCs, Chroma Vector, Chroma Deviation	0.77	0.6 (0.93)	AB
2	Entropy of Energy, Spectral Centroid, Spectral Entropy, Spectral Rolloff, MFCCs, Chroma Vector, Chroma Deviation	0.77	0.6 (0.93)	AB
3	Zero-crossing Rate, Spectral Spread, MFCCs, Chroma Deviation	0.76	0.62 (0.91)	AB
4	Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation	0.76	0.62 (0.91)	AB

Table 6.17 Effect of Audio Feature Standardization on Mean Audio Features. (Partial, see full in Appendix E)

6.1.10 Section Overview

Table 6.18 summarizes all the important classification results in Section 6.1. The best result is given by AB trained with mean of Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs and Chroma Deviation. Thus, they will be our proposed set of audio features. AB is also our proposed model. Besides, Audio Feature Normalization improves feature warping, but still needs to be used with care. Moreover, Feature Selection Techniques are useful in identifying redundant audio features. Furthermore, as the train set has limited train data, many classifiers could not predict the depressed dev samples correctly. Thus, AB is chosen as the classifier as it is able to perform consistently.

Audio Features	Model	Mean F1	F1
Mean of Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation	**AB (All)**	**0.82**	**0.71 (0.93)**
Normalized Mean of Spectral Centroid, Spectral Entropy, Spectral Rolloff, MFCCs, Chroma Vector, Chroma Deviation	AB (All)	0.77	0.6 (0.93)
Mean and Std of Spectral Centroid, Spectral Spread, Spectral Flux	AB (All)	0.80	0.67 (0.93)
Normalized Mean and Std of Entropy of Energy, Spectral Centroid, Spectral Spread, Spectral Flux	AB (All)	0.70	0.53 (0.87)

Table 6.18 Summary of all the important classification results

6.2 Depression Regression

The regression task is to estimate the participants' PHQ-8 score as mentioned in Appendix A. In this section, we will describe the efforts done to improve the result of the regression result using RMSE as the evaluation measure while providing MAE as a reference.

6.2.1 Regression Results

Table 6.19 shows the regression results using the optimal audio features subset in classification. Although the lowest RMSE score is given by Option 4, we choose option 1 to be the audio feature subset for the regression task to stay consistent with classification task.

Option	Audio Features	Model	RMSE	MAE
1	**Mean of Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation**	**AB**	**6.43**	**5.32**
2	Normalized Mean of Spectral Centroid, Spectral Entropy, Spectral Rolloff, MFCCs, Chroma Vector, Chroma Deviation	Linear SVR	6.42	5.16
3	Mean and Std of Spectral Centroid, Spectral Spread, Spectral Flux	AB	6.51	5.45
4	Normalized Mean and Std of Entropy of Energy, Spectral Centroid, Spectral Spread, Spectral Flux	AB	6.36	5.35

Table 6.19 The performance of Regressor trained by suggested Audio Feature Subsets.

6.2.2 The Performance of All Models

Using RMSE as the evaluation measure, most of the regressors performances surpass the baseline as shown in Figure 6.2. On the contrary, Table 6.20 shows both the MAE and RMSE of all the reviewed regressors. KNN and NB are not performing well because their models are too simplistic. SVR also performs consistently well, which is the reason why it is used in many related studies. Besides, RF which is used in the **Baseline** gives good performance when it is trained with the proposed subsets. Also, DT is a simple algorithm used in **Pampouchidou's Study** that surprisingly performs well and can be considered in depression studies. Moreover, GP-DP is a great and flexible ML model which performs sufficiently well and might perform better after optimizing its hyper-parameters. Last but not least, AB performs consistently well in both classification and regression tasks, thus is proposed to use as a ML model in the depression context.

Regressor	AB	RBF SVR	RF	DT	Linear SVR	GP-DP	KNN	NB
RMSE	**6.43**	6.54	6.55	6.57	6.68	6.74	7.16	8.54
MAE	**5.32**	5.31	5.50	5.50	5.25	5.80	6.0	6.54

Table 6.20 The Performance of Regressors trained by the suggested Audio Feature Subset where the bolded values are the best result and the underlined results are the regressors which failed the baseline.

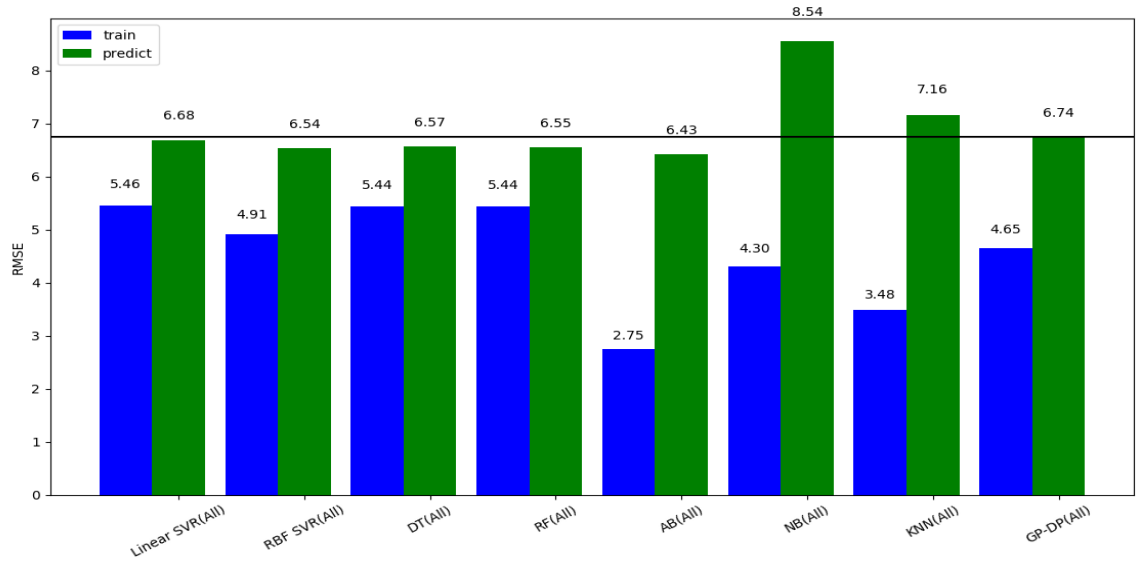

Figure 6.2 The Regressor Performances based on RMSE where the solid line is the prediction RMSE baseline.

6.2.3 Section Overview

The proposed audio features are Mean of Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation, whereas the proposed regressor is AB. AB provides a RMSE of 6.43 and MAE of 5.32.

6.3 Multi-class Classification

As depression level is common among different depression assessments, this study also performs multiclass classification on the depression level based on Appendix A to provide a more objective and general measure. In other words, there are 5 classes of depression which are minimal (0), mild (1), moderate (2), moderately severe (3) and severe (4). This approach enables the model to accept audio data from other depression corpus constructed using different assessment methods.

6.3.1 Results

Table 6.21 shows the multi-class classification results. By extending AB to multi-class classifier using either OVO or OVR implementation, AB performs consistently well by giving a mean F1 of 1.0. No matter AB is trained by which option, it still gives a mean F1 of 1, proving the strength of the model and our audio features set. To be consistent with our binary classification result, we choose the first option as our audio feature set.

Option	Audio Features	OVO Model	OVO Mean F1	OVR Model	OVR Mean F1
1	**Mean of Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation**	**AB**	**1.0**	**AB**	**1.0**

2	Normalized Mean of Spectral Centroid, Spectral Entropy, Spectral Rolloff, MFCCs, Chroma Vector, Chroma Deviation	AB	1.0	AB	1.0
3	Mean and Std of Spectral Centroid, Spectral Spread, Spectral Flux	AB	1.0	AB	1.0
4	Normalized Mean and Std of Entropy of Energy, Spectral Centroid, Spectral Spread, Spectral Flux	AB	1.0	AB	1.0

Table 6.21 Performances of Multi-class classifiers trained by suggested Audio Feature Set.

6.3.2 OVO or OVR Models

Figure 6.3 shows the performance of all OVR or OVO models. Obviously, AB performs consistently well in both. At the same time, the top 4 performances are given by OVR models, hinting that OVR might be dominating OVO models in this context. In fact, constructing OVO model might face the underfitting issue due to the shortage of audio data. Therefore, we suggest using OVR AB as our model in multi-class classification.

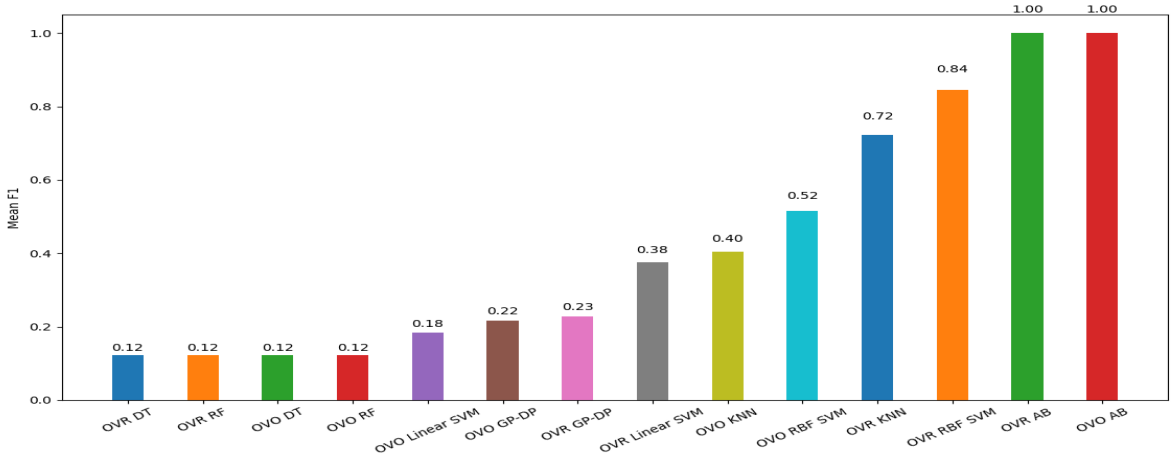

Figure 6.3 The performance of all OVR or OVO models trained by chosen Audio Features.

6.3.3 Section Overview

The proposed audio features are Mean of Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation, whereas the multi-class classifier is OVR AB. It provides a mean F1 of 1.

6.4 Chapter Summary

After doing the experiments specified in Section 6.1 and Section 6.2, the audio features that will be used are Mean of Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation, whereas the classifier and the regressor would be AB. In this section, we will compare the proposed model with related studies.

In the depression binary classification task discussed in Section 6.1, the AB classifier achieves 88.57% accuracy. For depressed class, the AB has F1 of 0.71, precision of 0.71 and recall of 0.71 while it obtains F1 of 0.93, precision of 0.93 and recall of 0.93 for not depressed

class. Table 6.22 and Table 6.23 show the comparison of the performance of the AB classifier and 5 other related studies. Mean F1 is included for the comparison of some related studies. Table 6.22 compares the difference of this study with other related study in terms of audio features, model, normalization technique and optimization method. We achieve good result using the simplest way even though we also tried a lot of different techniques. Particularly, Table 6.23 includes the ML model which is not only trained by the audio features but also trained by video, semantic and text features. Nevertheless, the proposed AB classifier outperforms all other related studies.

Related Study	Model	Audio Features	Normalization	Optimization	F1	Mean F1
Proposed Method	**AB**	**Mean of Zero-crossing Rate, Entropy of Energy, Spectral Spread, Spectral Entropy, MFCCs, Chroma Deviation**	**None**	**None**	**0.71 (0.93)**	**0.82**
Baseline (Valstar et al., 2016)	Linear SVM with SGD	F0, VUV, NAQ, QOQ, H1H2, PSP, MDQ, peak-Slope, Rd, MCEP0-24, HMPDM0-24, HMPDD0-12	None	Grid Search for hyper-parameter optimization	0.41 (0.58)	0.50
DepAudio Net (Ma et al., 2016)	CNN and LSTM	Mel-scale filter bank feature	Batch Normalization (uses Feature Warping)	Random Sampling	0.52 (0.70)	0.61
MITLL (Williamson et al., 2016)	Gaussian Staircase Model	Correlation structure of formant tracks, Correlation structure of dMFCCs, Lower Vocal Tract (VT) Resonance Pattern, peak-to-rms	Min-max Scaling	Z-scoring, PCA	-	0.57
Pampouchidou (Pampouchidou et al., 2016)	Decision Fusion Model which is implemented using DT	F_0, NAQ, QOQ, H1H2, PSP, MDQ, peakSlope, Rd, Rd conf, MCEP 0-24, HMPDM 1-24, HMPDD 1-12, Formants 1-3, the deltas and delta-deltas for F_0 and MFCCs, Pause Ratio, Voiced Segment Ratio, Speaking Ratio, Mean Laughter Duration, Mean Delay to answer the question, Mean Duration of Pauses, Maximum Duration of Pauses and Fraction of pauses in overall time	Min-max Scaling	Feature Selection accesses based on the improvement of modal by removing features.	0.59 (0.87)	0.73
SCUBA (Nasir et al., 2016)	G-PLDA	Ivector (MFCC)	None	MIM-based feature selection	0.57 (0.89)	0.73

Table 6.22 Comparison of Depression Binary Classification Result with Related Studies using audio data from development set. The proposed method is bolded.

On the other hand, the AB regressor obtains a RMSE of 6.43 and MAE of 5.32. As there are significantly lesser people that works on the depression regression task, Table 6.24

shows a comparison of AB regressor performance with 3 other related studies. Similar to Table 6.23, Table 6.24 also includes model that is trained with both audio and video, text and semantic features. Other than the ensemble model and audio model suggested by Williamson et al. (2016), AB beats the other related studies.

Related Study	Modality	F1	Mean F1	Precision	Recall	Accuracy
Proposed Method (AB)	**Audio**	**0.71 (0.93)**	**0.82**	**0.71 (0.93)**	**0.71 (0.93)**	**88.57%**
Baseline (Valstar et al., 2016)	Audio	0.41 (0.58)	0.50	0.27 (0.94)	0.89 (0.42)	-
Baseline (Valstar et al., 2016)	Audio-Video	0.58 (0.86)	0.72	0.47 (0.94)	0.78 (0.79)	-
DepAudioNet (Ma et al., 2016)	Audio	0.52 (0.70)	0.61	0.35 (1.00)	1.00 (0.54)	-
MITLL (Williamson et al., 2016)	Audio	-	0.57	-	-	-
MITLL (Williamson et al., 2016)	Audio-Video-Semantic	-	0.81	-	-	-
Pampouchidou (Pampouchidou et al., 2016)	Audio-Gender	0.59 (0.87)	0.73	-	-	-
Pampouchidou (Pampouchidou et al., 2016)	Audio-Video-Text-Gender	0.62 (0.91)	0.77	-	-	-
SCUBA (Nasir et al., 2016)	Audio	0.57 (0.89)	0.73	0.57 (0.89)	0.57 (0.89)	
SCUBA (Nasir et al., 2016)	Audio-Video	0.63 (0.89)	0.76	-	-	-

Table 6.23 Comparison of Depression Binary Classification Result with Related Studies on development set. The proposed method is bolded.

Related Study	Modality	RMSE	MAE
Proposed Method (AB)	**Audio**	**6.43**	**5.32**
Baseline (Valstar et al., 2016)	Audio	6.7418	5.3566
Baseline (Valstar et al., 2016)	Audio-Video	6.6212	5.5222
MITLL (Williamson et al., 2016)	Audio	6.38	5.32
MITLL (Williamson et al., 2016)	Audio-Video-Semantic	5.31	4.18
SCUBA (Nasir et al., 2016)	Audio	6.7334	5.8237

Table 6.24 Comparison of Depression Regression Result with Related Studies on development set. The proposed method is bolded and the lower RMSE and MAE score are underlined.

If the modality of Table 6.23 and Table 6.24 are limited to audio only, AB algorithm would definitely beat other related studies. Thus, it is sufficient to show that our proposed method works well for depression binary classification and regression using speech. A task summary of this paper is provided in Table 6.25.

Section	Tasks	Baseline	Best (Audio)	Best (All)	Proposed Result
6.1	Binary Classification	Mean F1 of 0.5	Mean F1 of 0.73	Mean F1 of 0.81	Mean F1 of 0.82
6.2	Regression	RMSE of 6.74	RMSE of 6.38	RMSE of 5.31	RMSE of 6.43
6.3	Multi-class Classification	None	None	None	Mean F1 of 1

Table 6.25 Task summary of this paper.

Chapter 7 Conclusion

This chapter summarizes the findings, contributions of the study and the future work that can be done.

7.1 Summary

In this study, an existing powerful machine learning (ML) model, AdaBoost (AB), trained by a new optimal audio feature set has been proposed as the model for depression analysis. Based on the assessment using Audio/Visual Emotion Challenge and Workshop 2016 (AVEC2016) data, the performance of the proposed approach surpassed related studies in both classification and regression tasks.

The optimal audio feature set comprises the mean of Zero-crossing rate, entropy of energy, spectral spread, spectral entropy, Mel Frequency Cepstral Coefficients (MFCCs), and chroma deviation. It is chosen using Audio Feature Selection via Complete Search and is proven to allow most of the ML models to perform robustly in depression task analysis. The ML model AB is the best and achieved a mean F1 of 0.82 and a Root Mean Square Error (RMSE) of 6.43 compared to the baseline audio models which give a mean F1 of 0.5 and a RMSE of 6.74. In fact, AB also gives a mean F1 of 1 in multi-class classification which predicts the depression level of an individual.

A standardized normalization method is also proposed, which maps all audio features instead of individual features. The approach is able to better preserve the original distribution of audio data. However, it should still be used with care.

7.2 Limitation

It is difficult to obtain other depression dataset to further validate the result because many institutions or clinics are not sharing them due to privacy reasons. The result of this study is limited to the assumption of cultures would not affect the result of depression prediction through speech. In addition, depression dataset constructed based on PHQ-8 score is even harder to obtain.

7.3 Future Work

The multi-class classifier could be tested using the depression level provided in other depression corpus. Ultimately, based on the proposed model, we hope to create an autonomous agent that could help the depressed patients.

References

Babin, Z. (2011). *How To Do Noise Reduction Using ffmpeg and sox.* Retrieved December 16, 2016, from http://www.zoharbabin.com/how-to-do-noise-reduction-using-ffmpeg-and-sox/

Bagwell, C. (2014). *SoX − Sound eXchange, the Swiss Army knife of audio manipulation.* Retrieved from http://sox.sourceforge.net/sox.pdf

Bello, J. P. (2013). *Low-level features and timbre. Lecture Notes*, 1–31. Retrieved from http://www.nyu.edu/classes/bello/MIR_files/timbre.pdf

Bodecs, B., Rapp, T., Niedermayer, M., Cadhalpun, A., Comeau, J., Liu, S., … Tumer, T. (2016). *FFmpeg*. Retrieved from http://ffmpeg.org/ffmpeg.html

Chai, T., & Draxler, R. R. (2014). *Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7*(3), 1247–1250. http://doi.org/10.5194/gmd-7-1247-2014

Chu, S., Narayanan, S., & Kuo, C.-C. J. (2009). *Environmental Sound Recognition with Time Frequency Audio Features.* IEEE Transactions on Audio, Speech, and Language Processing, 17(6), 1142–1158. http://doi.org/10.1109/TASL.2009.2017438

Cummins, N., Epps, J., Breakspear, M., & Goecke, R. (2011). An investigation of depressed speech detection: Features and normalization. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2997–3000.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). *A Review of Depression and Suicide Risk Assessment Using Speech Analysis.* Speech Communication, 71, 10–49. http://doi.org/10.1016/j.specom.2015.03.004

Dhingra, S. S., Kroenke, K., Zack, M. M., Strine, T. W., & Balluz, L. S. (2011*). PHQ-8 Days: a measurement option for DSM-5 Major Depressive Disorder (MDD) severity*. Population Health Metrics, 9(1), 11. http://doi.org/10.1186/1478-7954-9-11

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems*, 1–15. http://doi.org/10.1007/3-540-45014-9_1

Dietterich, T. G. (2002). Ensemble Learning.pdf. *The Handbook of Brain Theory and Neural Networks*. Retrieved from http://www-vis.lbl.gov/~romano/mlgroup/papers/hbtnn-ensemble-learning.pdf

Ellis, D. P. W. (2007). Classifying Music Audio with Timbral and Chroma Features. *Int Symp on Music Information Retrieval ISMIR, 199*(34), 339–340. http://doi.org/10.1.1.137.9005

Fiske, A., Wetherell, J. L., & Gatz, M. (2009). *Depression in Older Adults*. Annual Review of Clinical Psychology, 5(1), 363–389. http://doi.org/10.1146/annurev.clinpsy.032408.153621

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). *An Overview of Ensemble Methods for Binary Classifiers in Multi-Class Problems: Experimental Study on One-Vs-One and One-Vs-All Schemes.* Pattern Recognition, 44(8), 1761–1776. https://doi.org/10.1016/j.patcog.2011.01.017

Ghahramani, Z. (2015). *Probabilistic Machine Learning and Artificial Intelligence*. *Nature*, *521*(7553), 452–459. https://doi.org/10.1038/nature14541

Giannakopoulos, T., Kosmopoulos, D., Aristidou, A., & Theodoridis, S. (2006). Violence content classification using audio features. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3955 LNAI*, 502–507. http://doi.org/10.1007/11752912_55

Giannakopoulos, T., Pikrakis, A., & Theodoridis, S. (2008). Music tracking in audio streams from movies. *Proceedings of the 2008 IEEE 10th Workshop on Multimedia Signal Processing, MMSP 2008*, 950–955. http://doi.org/10.1109/MMSP.2008.4665211

Giannakopoulos, T. (2015). PyAudioAnalysis: *An Open-Source Python Library for Audio Signal Analysis*. PLoS ONE, 10(12), 1–17. http://doi.org/10.1371/journal.pone.0144610

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherere, S., Nazarian, A., Morency, L.-P. (2014). *The Distress Analysis Interview Corpus of human and computer interviews.* Proceedings of Language Resources and Evaluation Conference, 3123–3128. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf

Greenberg, P. E., Fournier, A., Sisitsky, T., Pike, C. T., & Kessler, R. C. (2015). *The Economic Burden of Adults with Major Depressive Disorder in the United States (2005 and 2010).* The Journal of Clinical Psychiatry, 2010(February), 155–162. http://doi.org/10.4088/JCP.14m09298

Grey, J. M. (1978). Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America*, *63*(5), 1493. http://doi.org/10.1121/1.381843

Hashim, N. W., Wilkes, M., Salomon, R., Meggs, J., & France, D. J. (2016). *Evaluation of Voice Acoustics as Predictors of Clinical Depression Scores. Journal of Voice.* http://doi.org/10.1016/j.jvoice.2016.06.006

Idri, A., Hosni, M., & Abran, A. (2016). *Systematic literature review of ensemble effort estimation.* Journal of Systems and Software, 118, 151–175. http://doi.org/10.1016/j.jss.2016.05.016

Johnstone, T. (2001). *The Effect of Emotion on Voice Production and Speech Acoustics*.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., … Wang, P. S. (2003). *The Epidemiology of Major Depressive Disorder.* JAMA, 289(23), 3095. http://doi.org/10.1001/jama.289.23.3095

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). *Supervised Machine Learning: A Review of Classification Techniques*, *31*, 249–268. Retrieved from

http://books.google.com/books?hl=en&lr=&id=vLiTXDHr_sYC&oi=fnd&pg=PA3&dq=Supervised+Machine+Learning+:+A+Review+of+Classification+Techniques&ots=CXnrvy3Kkk&sig=IIaoAHenDor69TQpzd4DRs7Pp5Q%5Cnhttp://books.google.com/books?hl=en&lr=&id=vLiTXDHr_sYC&oi=fnd&p

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). *The PHQ-8 as a measure of current depression in the general population*. Journal of Affective Disorders, 114(1–3), 163–173. http://doi.org/10.1016/j.jad.2008.06.026

Li, Y., & Zhu, J. (2008). L_1-Norm Quantile Regression. *Journal of Computational and Graphical Statistics*, *17*(1), 163–185. http://doi.org/10.1198/106186008X289155

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2016). Feature Selection: A Data Perspective. Journal of Machine Learning Research, 1–73. Retrieved from http://arxiv.org/abs/1601.07996

Lyons, J. (2012). *Mel Frequency Cepstral Coefficient (MFCC) tutorial*. Retrieved January 4, 2017, from http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

Ma, X., Yang, H., Chen, Q., Huang, D., & Wang, Y. (2016*). DepAudioNet*. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16 (pp. 35–42). New York, New York, USA: ACM Press. http://doi.org/10.1145/2988257.2988267

MacPherson, H., Richmond, S., Bland, M., Brealey, S., Gabe, R., Hopton, A., … Watt, I. (2013). *Acupuncture and Counselling for Depression in Primary Care: A Randomised Controlled Trial*. PLoS Medicine, 10(9), e1001518. http://doi.org/10.1371/journal.pmed.1001518

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). *Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects*. Journal of Clinical Epidemiology, 67(3), 267–277. http://doi.org/10.1016/j.jclinepi.2013.08.015

Melorose, J., Perroy, R., & Careas, S. (2015). Taking the Human Out of the Loop: A Review of Bayesian Optimization, *1*(1). http://doi.org/10.1017/CBO9781107415324.004

Morey, M. E., Arora, P., & Stark, K. D. (2015). *Multiple-Stage Screening of Youth Depression in Schools.* Psychology in the Schools, 52(8), 800–814. http://doi.org/10.1002/pits.21860

Müller, M., & Ewert, S. (2011). Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features. *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, (Ismir), 215–220. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:CHROMA+TOOLBOX+:+MATLAB+IMPLEMENTATIONS+FOR+EXTRACTING+VARIANTS+OF+CHROMA-BASED+AUDIO+FEATURES#0

Nasir, M., Jati, A., Shivakumar, P. G., Chakravarthula, S. N., & Georgiou, P. (2016). *Multimodal and Multiresolution Depression Detection from Speech and Facial Landmark Features*, 43–50. http://doi.org/10.1145/2988257.2988261

Pampouchidou, A., Simantiraki, O., Fazlollahi, A., Pediaditis, M., Manousos, D., Roniotis, A., … Tsiknakis, M. (2016*). Depression Assessment by Fusing High and Low Level Features from Audio, Video and Text*, (October), 27–34. http://doi.org/10.1145/2988257.2988266

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2012). *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research*, 12, 2825–2830. http://doi.org/10.1007/s13398-014-0173-7.2

Pohjalainen, J., Fabien Ringeval, F., Zhang, Z., & Schuller, B. (2016). *Spectral and Cepstral Audio Noise Reduction Techniques in Speech Emotion Recognition*. Proceedings of the 2016 ACM on Multimedia Conference - MM '16, 670–674. http://doi.org/10.1145/2964284.2967306

Powell, G. E., & Percival, I. C. (1979). A spectral entropy method for distinguishing regular and irregular motion of Hamiltonian systems. *Journal of Physics A: Mathematical and General*, *12*(11), 2053–2071. http://doi.org/10.1088/0305-4470/12/11/017

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press. Retrieved from http://www.gaussianprocess.org/gpml/chapters/RW.pdf

Sahidullah, M., & Saha, G. (2012). *Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition*. Speech Communication, 54(4), 543–565. http://doi.org/10.1016/j.specom.2011.11.004

Sanjay, L. (2008). Norms and Vector Spaces. *Information Systems Laboratory*, (i), 1–6. Retrieved from http://lall.stanford.edu/svn/engr207c_2010_to_2011_autumn/data/norms_2008_10_07_01.pdf

Schubert, E., & Wolfe, J. (2006). Does timbrai brightness scale with frequency and spectral centroid? *Acta Acustica United with Acustica*, *92*(5), 820–825.

Scikit-learn developers. (2010a). *1.10. Decision Trees*. Retrieved February 1, 2017, from http://scikit-learn.org/stable/modules/tree.html

Scikit-learn developers. (2010b). *1.12. Multiclass and multilabel algorithms*. Retrieved April 4, 2017, from http://scikit-learn.org/stable/modules/multiclass.html

Scikit-learn developers. (2010c). 1.*13. Feature selection*. Retrieved January 28, 2017, from http://scikit-learn.org/stable/modules/feature_selection.html

Scikit-learn developers. (2010d). *1.6. Nearest Neighbors*. Retrieved February 1, 2017, from http://scikit-learn.org/stable/modules/neighbors.html

Scikit-learn developers. (2010e). *4.3. Preprocessing data*. Retrieved February 2, 2017, from http://scikit-learn.org/stable/modules/preprocessing.html

Seeger, M. (2004). Gaussian Processes for Machine Learning. *International Journal oNeural Systems*, *14*(2), 69–106. http://doi.org/10.1142/S0129065704001899

Sethu, V., Ambikairajah, E., & Epps, J. (2007). Speaker normalisation for speech-based emotion detection. *2007 15th International Conference on Digital Signal Processing, DSP 2007*, 611–614. https://doi.org/10.1109/ICDSP.2007.4288656

Sethu, V., Ambikairajah, E., Epps, J., Wales, S., & Nsw, S. (2009). *Speaker Dependency Of Spectral Features and Speech Production Cues* The School of Electrical Engineering and Telecommunications, National Information Communication Technology Australia (NICTA), Australian Technology Park , Eveleigh 1430 , Australia. *Electrical Engineering*, 4693–4696.

Smith, A. L., & Cashwell, C. S. (2011). *Social Distance and Mental Illness: Attitudes Among Mental Health and Non-Mental Health Professionals and Trainees.* The Professional Counselor: Research and Practice \, 1(1), 13–20. http://doi.org/10.15241/als.1.1.13

Snelson, E. L. (2007). Flexible and efficient Gaussian process models for machine learning. *ACM SIGKDD Explorations Newsletter*, *7*(2001), 1–135. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.4041&rep=rep1&type=pdf%5Cnhttp://portal.acm.org/citation.cfm?id=1117456

Sokolova, M., & Lapalme, G. (2009). *A systematic analysis of performance measures for classification tasks.* Information Processing and Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Stratou, G., Scherer, S., Gratch, J., & Morency, L. P. (2015). *Automatic nonverbal behavior indicators of depression and PTSD: the effect of gender.* Journal on Multimodal User Interfaces, 9(1), 17–29. http://doi.org/10.1007/s12193-014-0161-4

Sugumaran, V., Muralidharan, V., & Ramachandran, K. I. (2007). Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing. *Mechanical Systems and Signal Processing*, *21*(2), 930–942. http://doi.org/10.1016/j.ymssp.2006.05.004

Tan, C. M., Wang, Y. F., & Lee, C. Do. (2002). *The use of bigrams to enhance text categorization. Information Processing and Management*, 38(4), 529–546. https://doi.org/10.1016/S0306-4573(01)00045-0

The GPyOpt authors. (2016). GPyOpt: *A Bayesian Optimization framework in python.* Retrieved February 1, 2017, from http://github.com/SheffieldML/GPyOpt%7D

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & Sacha van Hijum, A. F. T. (2013). Data mining in the life science swith random forest: A walk in the park or lost in the jungle? *Briefings in Bioinformatics*, *14*(3), 315–326. http://doi.org/10.1093/bib/bbs034

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., … Pantic, M. (2014). Avec 2014: 3d dimensional affect and depression recognition challenge. *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC '14*, 3–10. http://doi.org/10.1145/2661806.2661807

Valstar, M., Gratch, J., Ringeval, F., Torres, M. T., Scherer, S., & Cowie, R. (2016). *AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge.*

Webb, G. I., & Zheng, Z. (2004). *Multistrategy ensemble learning: reducing error by combining ensemble learning techniques*. IEEE Transactions on Knowledge and Data Engineering, 16(8), 980–991. http://doi.org/10.1109/TKDE.2004.29

Williamson, J. R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., … Quatieri, T. F. (2016). Detecting Depression using Vocal, Facial and Semantic Communication Cues. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, (October), 11–18. http://doi.org/10.1145/2988257.2988263

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. http://doi.org/10.3354/cr030079

Wolpert, L. (2001). *Stigma of depression--a personal view*. British Medical Bulletin, 57, 221–224.

World Health Organization. (2008). *The Global Burden of Disease: 2004 update*. 2004 Update, 146. http://doi.org/10.1038/npp.2011.85

Wu, B. F., & Wang, K. C. (2005). Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 762–774. http://doi.org/10.1109/TSA.2005.851909

Xu, J., Murphy, S. L., Kochanek, K. D., Bastian, B. A., & Statistics, V. (2016). *National Vital Statistics Reports Deaths : Final Data for 2013*, 64(2).

Appendix A : PHQ-8

The Patient Health Questionnaire nine-item depression scale (PHQ-9) is a depression measure which is short and valid for both diagnostic and severity measure (Kroenke et al., 2009). It is used in clinical and public settings as it can detect depressive symptoms of individuals from diverse cultures. It contains the nine items for depression from the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV) (Kroenke et al., 2009).

PHQ-8 is a duplicate of PHQ-9, without the ninth question of PHQ-9 which is used to access suicidal or self-injurious thoughts (Kroenke et al., 2009). As the removal is due to ethical reason and has only a minor impact on scoring, PHQ-8 can diagnose depression as well as PHQ-9 (Kroenke et al., 2009). The other questions are shown in Figure 0.1.

As represented in Table 0.1, each item of the PHQ-8 is assigned 0~3 points, making a total maximum points of 24 (Kroenke et al., 2009). The meaning of the PHQ-8 score is important as we will perform a depression regression task with output being the PHQ-8 score. PHQ-8 severity score is highly correlated with MDD diagnostic status, as shown in Figure 0.2. As 100% of the MDD patients obtain a PHQ-8 score ≥ 10, a PHQ-8 score ≥ 10 is proven to be able to represent MDD (Kroenke et al., 2009).

Total Severity Score (PHQ-8)	Depression Level
0 ~ 4	Minimal Depression
5 ~ 9	Mild Depression
10 ~ 14	Moderate Depression
15 ~ 19	Moderately Severe Depression
20 ~ 24	Severe Depression

Table 0.1: The representation of the PHQ-8 score (Kroenke et al., 2009).

Nowadays, the clinician-administered Hamilton Rating Scale for Depression is the state of the art depression severity assessment, whereas the state of the art diagnosis assessment is the Structured Clinical Interview for DSM-IV (SCID) (Valstar et al., 2016). Beck Depression Inventory-II (BDI-II) is also commonly used in depression diagnosis (Valstar et al., 2014).

Over the last 2 weeks, how often have you been bothered by any of the following problems? (Use "✓" to indicate your answer)	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. Troubleconcentratingonthings,suchas reading thenewspaperorwatchingtelevision.	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
(For office coding: Total Score ____ = ___ + ___ + ___)				

Figure 0.1: The questions in PHQ-8 (Kroenke et al., 2009).

Depressive symptom severity [PHQ-8 Score]	Depressive disorder by diagnostic algorithm				Total
	Major	Other	Any	None	
	(*n*=8476)	(*n*=9577)	(*n*=18,053)	(*n*=180,625)	(*n*=198,678)
0–4	0	200	200	149,921	150,121
5–9	0	5297	5297	26,220	31,517
10–14	1944	3978	5922	4046	9968
15–19	4296	102	4398	428	4826
20–24	2236	0	2236	10	2246
≥10	8476	4080	12,556	4484	17,040
<10	0	5497	5497	176,141	181,638

Figure 0.2: Distribution of depressive symptom severity and depressive disorders in BRFSS (Kroenke et al., 2009).

Appendix B : MFCCs Implementation

As MFCCs is one of the most commonly used spectral features in related studies, it is chosen to be the main audio feature during AY2016/17 Semester 1. The definition is provided below.

Mel Frequency Cepstral (MFC) is a representation of short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. The formula to convert from frequency f to Mel scale M(f) is given in Equation 0-1.

$$M(f) = 1125\ln(1 + \frac{f}{700})$$

Equation 0-1 Mel scale M(f) of a frequency f (Lyons, 2012).

MFCCs are the coefficients that forms the MFC. The greatest benefit of using MFCCs is that the scale approximates to the human's auditory system response more closely, hence it allows for a better representation of sound. In order to obtain MFCCs, Discrete Fourier Transform (DFT) is performed on the sound signals to gain insights on the spectral features. The algorithm used to extract MFCCs is shown in Figure 0.3.

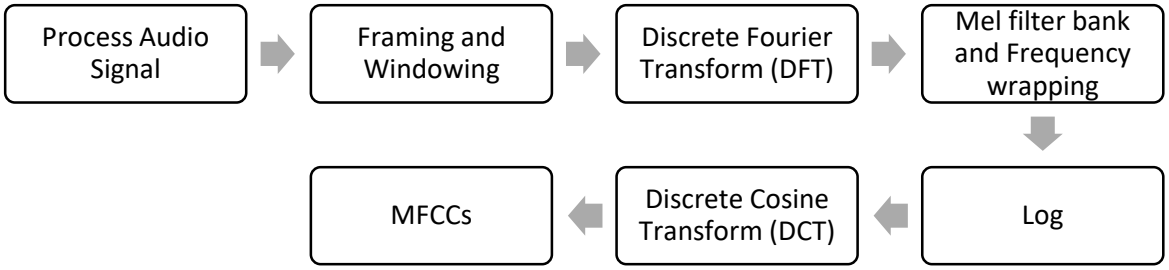

Figure 0.3 Algorithm Used to Extract MFCCs.

Firstly, we process the audio signal from wav files. Then, we need to divide the audio signals to smaller frames such as 20~40ms frames so that we have enough samples to have a reasonable spectral estimate while signal changes are almost statistically stationary (Lyons, 2012). Next, we apply DFT to calculate the periodogram estimate of the power spectrum. This estimate is to simulate the way the human ear, or more specifically, the way the human cochlea works (Lyons, 2012). The next step is to apply the mel filterbank to the power spectra, sum the energy in each filter to learn about the amount of energy in various frequency regions (Lyons, 2012). Thenafter, we take the logarithm of the energy amount in various frequency regions to simulate human hearing, as human does not hear loudness in linear scale (Lyons, 2012). Lastly, we perform DCT to increase the robustness and decorrelate the energy amount in various frequency regions (Sahidullah & Saha, 2012). Only the first 13 coefficients are kept because they are the most effective features for speech recognition (Lyons, 2012). A Java Program is written to extract MFCCs from the pre-processed audio recordings.

Appendix C : Feature Selection Techniques provided in Scikit

There are a few feature selection techniques provided in Scikit-learn toolbox (Pedregosa et al., 2012), which are implemented to improve the model performance on datasets with many features (Scikit-learn developers, 2010c). They are not examined in this study because they are wrapper methods which are inefficient and dependent on the ML model performance (J. Li et al., 2016).

Firstly, as tree-based estimators such as the decision tree are constructed based on the importance of features (Kotsiantis et al., 2007), it could be adapted to calculate the importance of features (Scikit-learn developers, 2010c). This type of feature selection method is called Tree-based Selection which can be implemented using ExtraTreesClassifier provided by Scikit (Scikit-learn developers, 2010c).

Besides, Scikit provides L1-based feature selection which has the LASSO model in Scikit as its regression sparse estimator as well as predict in classification space using Logistic Regression and Linear Support Vector Classifier (Scikit-learn developers, 2010c). It selects features based on the non-zero estimated coefficients or the importance of features for the linear models penalized with the L1-norm (Scikit-learn developers, 2010c). Some of the fitted coefficients for the linear model could be shrinked to exactly zero by L1-norm penalty (Y. Li & Zhu, 2008; Scikit-learn developers, 2010c). The basis pursuit model and the LASSO model are examples which make use of L1-norm penalty for least squares regression (Y. Li & Zhu, 2008).

On the other hand, the definition of norm can be found in Appendix D and L1-norm is defined in Equation 0-2. In ML, the model's complexity is penalized using the L1-norm of the coefficient of the features (Y. Li & Zhu, 2008). As the coefficients of the features are their weights, the coefficients can also be interpreted as the importance of the features. A regularization parameter $\lambda > 0$ is used to balance the quantile loss and the penalty (Y. Li & Zhu, 2008). When λ is large enough, some of the coefficients would be exactly 0 (Y. Li & Zhu, 2008).

$$||x||_1 = \sum_{i=1}^{n} |x_i|$$

Equation 0-2 The definition of L1-norm $||x||_1$ (Sanjay, 2008).

Appendix D : The definition of Norm

A norm ||x|| is a function f in the vector space V which has the following properties (Sanjay, 2008):

1. $f(x) \geq 0 \; for \; \forall x \; \in V$ (Sanjay, 2008)
2. $f(x + y) \; \leq f(x) + f(y) for \; \forall x, y \; \in V$ (Sanjay, 2008)
3. $f(\lambda x) = |\lambda| \, f(x) \; for \; \forall \lambda \; \in \; C \; \& \, x \; \in V$ (Sanjay, 2008)
4. $f(x) = 0 \; \leftrightarrow \; x = 0$ (Sanjay, 2008)

Appendix E : N-Layer Ensemble

As ensemble tends to perform better than single classifier (Dietterich, 2000), we propose the use of N-Layer ensemble besides the ML models mentioned in Section 3.4. N-layer ensemble has N-1 layers, which each of them contain M ML models described above, and the last layer contains only one of the M models mentioned. The first layer of the ensemble is trained by the training data, whereas the subsequent layers are trained by the output of the previous layer. The output of the last layer would be the prediction of N-layer ensemble. While this model seems to make sense, it could not improve the result.

The results produced by 2-layer Ensemble are recorded in Table 0.2. As the results produced by 3-layer Ensemble tend to have a similar F1 score which is 0.29 for depressed class and 0.82 for non-depressed class, the results are omitted. Linear SVM 2-layer Ensemble and AB 2-layer Ensemble performed the best among all of them. As the performance of n-layer ensemble tends to be capped at the best performance of the normal classifiers (AB), it is not used in the study.

Classifiers	F1	Precision	Recall	Accuracy
AB	**0.29(0.82)**	**0.29(0.82)**	**0.29(0.82)**	**71%**
Linear SVM 2-layer Ensemble	**0.29(0.82)**	**0.29(0.82)**	**0.29(0.82)**	**71%**
AB 2-layer Ensemble	**0.29(0.82)**	**0.29(0.82)**	**0.29(0.82)**	**71%**
RBF SVM 2-layer Ensemble	0.27(0.8)	0.29(0.79)	0.25(0.81)	69%

Table 0.2 Classifiers Performances trained by mean MFCCs. The value which is not in the bracket is the value for depressed class and the value which is in the bracket is the value for not depressed class. The best result is bolded.

Appendix F : Subsets of Audio Features in pyAudioAnalysis

As the tables that contain all the different subsets of Audio Features is too large, they are attached as a pdf file. In the tables, the rows are sorted according to the Best F1 Score provided by the ML model which performs the best given a combination of audio features. There are 2048 rows including the header row. The tables contain the information of the subsets of the audio features used, the number of audio features, the number of features and the best F1 score with the best ML model.

There are 5 different tables which contain all the subsets of audio features. Firstly, “classifierWithFeatureMean.csv” contains all the classifiers performances trained by different mean audio features subsets. Secondly, “classifierWithFeatureMeanStd.csv” consists of classifiers performances trained by different mean and std of audio features subsets. Next, “classifierProposedWithFeatureMeanStd.csv” records the classifiers performances trained by different normalized mean and std of audio features subsets. Furthermore, “classifierWithFeatureMeanStdFS.csv” includes classifiers performances trained by different mean and std of audio features subsets after applying feature selection techniques. Moreover, “classifierProposedWithFeatureMean.csv” records the classifiers performances trained by different normalized mean of audio features subsets. Finally, “classifierProposedWithFeatureMeanStdFS.csv” comprises of classifiers performances trained by different mean and std of normalized audio features subset which the redundant features are removed.