CS4246 AI Planning and Decision Making - Project 2
Planning and Decision Making Automation on Depression

Team 01
Antoine Charles Vincent Garcia - A0159072A
Chan Jun Wei - A0112084U
Chen Tze Cheng - A0112092W
Eric Ewe Yow Choong - A0112204E
Han Liang Wee, Eric - A0065517A
Ho Wei Li - A0094679H

Abstract

Depression is a debiliatating mental illness that has good prognosis given early detection and treatment. However, detection is difficult with the various factors that raise the entry barriers and decrease the accuracy of a diagnostic test. Research have shown that accurate predictions of emotions can be made with Gaussian Process models. This study explores the novel use Gaussian Process in predicting depression severity using acoustic measures of voice. Our work has succesfully shown that Gaussian Process Dot Product trained using MFCC feature set is a good model for depression prediction and can predict PHQ-8 better than other state-of-the-art models at RMSE of 6.34.

Introduction

Depression has a severe, and at times long-term, negative impact on an individual's quality of life. Major depression is 3rd leading cause of disability worldwide with 65 million life years spent living with the disability or lost due to early death [World Health Organization, 2004]. Depression's annual toll on U.S businesses amounts to about $80 billion in medical expenditures, lost productivity and suicide. Among the costs, close to $10 billion accrued in lost workdays each year and more than $33 billion in other costs accrue from decreased productivity due to symptoms that sap energy, affect work habits, cause problems with concentration, memory, and decision-making [Greenberg et al., 2015].

Left unchecked, depression increases risk for morbidity, suicide, decreased cognitive and social functioning, self-neglect, and early death [Fiske, Wetherell, and Gatz, 2009]. Death from suicide is one of the top 10 causes of death, above the death rate for chronic liver disease, Alzheimer's, homicide, arteriosclerosis or hypertension [Jiaquan Xu et al., 2016].

Despite the severe consequences, depression is one of the most treatable mental illnesses but it is also one of the most under-diagnosed globally. In general health-care, 48.4% of patients suffering from depression go unrecognized [RC et al., 2003].

Motivation and Objective

The Personal Health Questionnaire depression scale (PHQ-8) is a self-administered, 8-question diagnostic test for depressive disorders that has proven to be an effective severity measure for depression in large clinical studies [Kroenke et al., 2008]. Nevertheless, one of the biggest obstacles to successful diagnosis of depression is the unwillingness of patients to admit their predisposition to depression by seeking help.

People often subscribe to the social stigma that being depressed reflects a weakness in their character, a permanent defect in their personality. This stigma manifests itself particularly in a phenomenon known as social distancing whereby people with mental issues are more isolated from others [Smith and Cashwell, 2011]. People suffering from depression hence tend to be ashamed of their condition and are generally convinced that denying and hiding it from others gives them a better shot at integrating with society and living a normal life [Wolpert, 2001]. Even if they do seek help, the accuracy of the PHQ-8 or just questionaires and surveys in general are often adversely affected by the Hawthorne Effect, a type of reactivity in which individuals modify or improve an aspect of their behavior in response to their awareness of being observed. [McCambridge, Witton, and Elbourne, 2014].

In the past decade, there have been research successfully correlating emotion with voice production and speech acoustics [Johnstone, 2001]. Corollary to that, active research of late into the use of voice acoustics as predictors of clinical depression scores has seen success, proving that it is an effective indicator of depression severity [Hashim et al., 2016].

In this paper, we investigate the applicability and feasibility of Gaussian Process (GP) models in predicting clinical ratings of depression severity on the PHQ-8 scale with

acoustic measures of voice from a sample of patients and compare their performance with current state-of-the-art machine learning models. Some of our preliminary studies have shown that despite consensus among the scientific community that Support Vector Machine (SVM) models have a very high predictive accuracy specifically in speech emotion recognition [Chavhan, Dhore, and Yesaware, 2010], GP models have been proven to consistently outperform SVM models on the task of music emotion recognition [Markov and Matsui, 2013]. Nevertheless, there are no studies on the use of GP in predicting depression severity. The findings of this paper will hopefully open up new frontiers and fuel further research interest on this topic.

Important Requirements

...

Modelling and Approach

Leveraging on the success of the modelling of depression prediction of PHQ-8 scores in project 1, we extend and apply the work done to solve the aforementioned problem of under-staffing. Our solution is to implement pre-screening and automate the process of the deciding if the person needs an appointment or otherwise. Moreover, our process allows prioritizing of patient's appointments by their PHQ-8 score. Henceforth, we would need to use 2 GPs: Gaussian Process Classifier (GPC) and the Gaussian Process Regressor (GPR). In this section, we will firstly describe the process that we are proposing and we will go into detail on each component of the process.
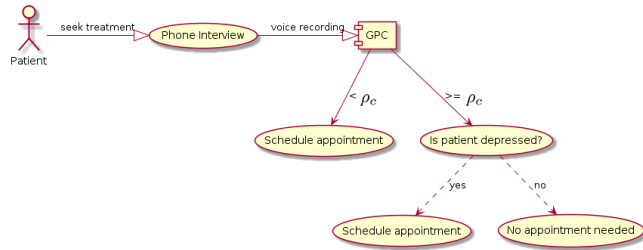
Automation Flow

Figure 1: Automation Flow

As shown in Figure 1, the model is the steps of an average user who is seeking for medical attention in the early stages of depression. Prior to medical treatment, the patient is required to call up a psychiatrist to arrange an appointment time and date. At that time, a short phone interview consisting of a fixed number of unbiased questions is conducted. The objective of the interview is to record the patient's voice pattern. The recordings will the serve as an input for the first GP, which is a classifier that will output both the confidence interval and a binary state to whether the patient is suffering from depression or not. This leads to two scenarios depending on the ρ_c, which is the probability estimate:

1. $>= \rho_c$
   The patient's affliction diagnosable. However, this is also dependent on the binary state i.e. whether the patient is depressed or not depressed. A depressed patient will be given further attention by the psychiatrist by scheduling an appointment whereas a non-depressed patient would be not be given attention by the psychiatrist.

2. $< \rho_c$
   The patient's affliction is uncertain. In this case, the patient can still schedule an appointment and is placed under continuous scrutiny until he/she can be diagnosed as depressed or vice versa. This outcome is not unusual as depression is disorder can be recurrent and often in episodes [Cesar and Chavoushi, 2013], a prediction generated over a time span therefore produces more accurate results.
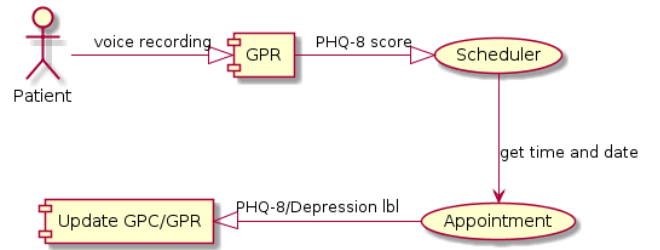

Figure 2: Flow for Scheduling an Appointment

Continuing from Figure 2 after the psychiatrist has agreed for an appointment, the voice of the patient will then be used as an input for the GPR which models the PHQ-8. This is to determine the scale of depression that the patient is suffering from which also helps the psychiatrist to select an appropriate appointment date. Similar the to the triage procedure in emergency services, a patient with a higher PHQ-8 value will be placed on the higher end of the priority list and vice versa. The patient's LBL and PHQ-8 scores will be observed over time during scheduled sessions to identify the patient's improvement in his/her emotional state. All the data gathered, will then be fed into the GPC and GPR until there is an improvement in the results. Undiagnosable patients undergo the same process until he/she can be given a proper classification of whether he/she is depressed or not. From here on, he will also undergo the process in 2 or is refuse an appointment if found that he/she is not depressed.

Insights to our model

Gaussian Processes : In our modelling, we used 2 gaussian processes(GP) for regression and classification tasks respectively. The GP models that we use here are modelled similarly to the modelling done in project 1. GPR predicts the (Personal Health Questionaire depression scale) PHQ-8 score, while GP classifier(GPC) predicts whether the person is depressed or not. Similar to project 1, we trained GPR and

GPC with data obtained from Audio/Visual Emotion Challenge and Workshop (AVEC 2016) [Valstar et al., 2016b], with PHQ-8 and depression labels respectively. We applied the same audio signal processing techniques to the audio files as per project 1. We only used the Mel Frequency Cepstral Coefficients (MFCC) feature subset as we have seen the good theorical and experimental results the feature subset had produced. Most importantly, we make the same two assumptions so that the GP model planning and decision making is suitable for depression prediction:

1. Depression prediction is an event-based recognition which provides a single depression estimate over a certain amount of time. [Valstar et al., 2016a]

2. The speech signals extracted from different people suffering from depression should share some similarities and thus admissible for prediction with the Gaussian Process models[Cummins et al., 2015].

Since the GP models recieve new data incrementally, we cannot use an offline GP as per project 1. Hence, we would need to tweak the GP to fit this problem. Since it is not possible for us to obtain a comprehensive dataset of all people coming to the clinic, we would then need a GP that learn dynamically, adapting to new data as it becomes avaliable. Online machine learning is a method that allows the data to be updated when it becomes available. We have noted that there are online variants of GP, namely: XXYY, XXYYX and XXHS. It is with this reasoning that we chose this particular online variant of GP.

Phone Interview : We note that regardless of the questions asked or the circumstance, a depressed person will still exhibit signs of depression in his speech[Kroenke et al., 2008]. Hence, the questions asked are irrelevant. The questions will be presented in the proof of concept section.

Decision making : The greatest advantage of using GP over other machine learning algorithms is that it provides us with a probability estimate representing confidence of the prediction. With that probability, we can decide if the predicted label can be relied on or that it cannot be trusted and needs to learn this data point. In our model, we exploit this property, unique to GPs. After the interview, GPC will predict the depression label with a confidence p_c on the audio recording. As mentioned in the modelling, we define a particular probability ρ_c. If GPC predicts with $p_c > \rho_c$, then we can trust the predicted label and continue to decide appropriately based on the label. We determine the probability ρ_c experimentally by observing the prediction quality of the labels in the training data, a summary is given in table 1.

Scheduling : For the cases when the person needs to be scheduled for an appointment with the psycharist, we can use GPR to determine the predicted PHQ-8 scores. With the predicted scores, we can piroritise certain higher risk individuals over the rest. Since the depression clinic is under-staffed and struggling to keep up with number of patients, it is wise to piroritize the individuals who are predicted with higher PHQ-8 scores, which indicates that they are likely to be more depressed. Hence, we optimize the scheduling piroritizing people with higher predicted PHQ-8 scores.

p_c(%)	Depression	No Depression
[0, 10)	-	-
[10, 20)	100%(1/1)	-
[20, 30)	100%(4/4)	-
[30, 40)	95	100%(1/1)
[40, 50)	6.326	7.548
[50, 60)	5.597	6.690
[60, 70)	100%(1/1)	31
[70, 80)	-	100%(1/1)
[80, 90)	-	100%(4/4)
[90, 100]	-	-
-	95	31

Table 1: Relationship between probability and accuraccy... Fix table later

Qualitative Advantages

The greatest advantage of using a GP is that it provides us with a probability estimate along with the predicted label. We rely on the probability to determine the reliability of the predicted label. As we are dealing with human beings, we would want to rely on the predicted label only if it is reliable. Additionally, we have read in medical literature regarding the dire consequences of misdiagnosis and/or inappropriate treatment[Nasrallah, 2015; Bowden, 2001; Dunner, 2003] in the area of depression. With the probability, we can be confident of the decisions that we make, that can potentially affect a person. Assume that we have a predicted label that can be trusted, then with the label, i.e. true for depressed and false for not depressed, we can decide if the person needs to come to the clinic for an appointment. Hence, reducing the number of appointments that are made, easing the workload of the staff in the clinic.

In addition to reducing the workload of the staff, the GP models can potentially improve as more patients go through the pre-screening process. It can be the case that the GP is not confident of its prediction, then we should not trust the label that the GP had predicted. Then, we would need to determine by the means of a physical examiniation if the person in question is depressed or not and administer the appropriate treatment. Through the appointment with the psychiatrist, he can determine if the patient is depressed or not. With the ground truth, we can now train the GPs with the new data point. We are also careful to only update the GP if the data point is predicted with low confidence.

In our model, we do not require a trained staff to administer the phone interview. GP will determine the depression label of the person objectively, not considering the content of the interview but relying on certain depression indicators in speech[of Health and Services, 2015]. This allows the removal of any bias(eg. gender, racial) in pre-screening, which can potentially cause misdiagnosis. Additionally, the clinic can save skilled-manpower as they can hire anyone or can use an automated system to perform the pre-screening task.

Hence, our model introduces an unbiased pre-screening process reaping the following benefits: leading to a reduction of the number of appointments, reduction of manpower required and a pre-screening process whose accuracy improves incrementally over time.

Evaluation

In order to test our proposed GP models, we conducted tests on data obtained from Audio/Visual Emotion Challenge and Workshop (AVEC 2016) [Valstar et al., 2016b]. The goal of AVEC is to weigh-in on the various approaches used to recognize emotions under unambiguous conditions. AVEC 2016 provided 2 pieces of data as input: visual and auditory data from each of the participants. However, we would be reducing the scope of the experiment, limiting the experiment to only the auditory data. Two Sub-Challenges were lised in AVEC 2016. We are only interested in the Depression Classification Sub-Challenge, which requires participants to classify inputs by the PHQ-8 score. In this experiment, we would be using the audio data along with their corresponding PHQ-8 scores to test our assumptions and confirm our hypothesis.
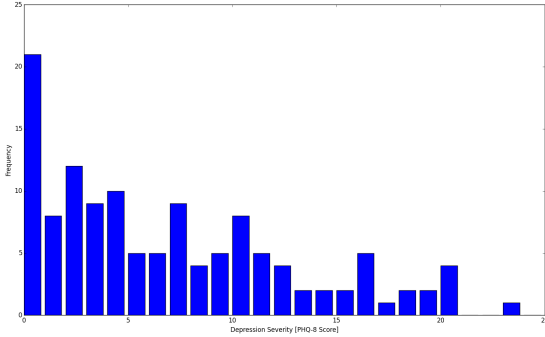
Data

Figure 3: Histogram of the PHQ-8 scores

The depression data used in AVEC 2016 was obtained from the benchmarking database, Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ). Data collected from DAIC-WOZ include raw audio and video recordings and the corresponsing PHQ-8 score (from 0 to 24) [Kroenke et al., 2008]. Hence, we would need to pre-process the auditory data before we use it in our experiment. The pre-processing is briefly discussed in the section below. The distribution of the depression severity scores in the dataset is given in Figure 3. The data provided are split into 2 sets: training and development. An overview of the data is given in Table 2.

Pre-processed data

Since the focus of this paper is the prediction of the PHQ-8 score, we will not describe the pre-processing step in detail.

	Training	Development	All
n	95	31	126
μ	6.326	7.548	6.626
σ	5.597	6.690	5.909

Table 2: Summary of Datasets provided

We used standard signal processing techniques to extract the 4 audio feature sets (Energy, MFCC, Magnitude Spectrum, Zero-crossing) as presented in the Modelling and Approach section. Each audio feature set comprises of several individual features and the breakdown of the actual number of feature columns is summarized in Table 3.

Audio Feature Sets	Number of features
Magnitude Spectrum	512
MFCC	12
Energy	1
Zero-Crossing Rate	1
Total	**526**

Table 3: Number of features extracted

Measure of Accuracy

AVEC 2016 provided a baseline classifier that consistently predicts the PHQ-8 score with RMSE $= 6.7418$ [Valstar et al., 2016b]. In order to provide a meaningful and consistent comparison to the baseline provided, we used the same Root Mean Square Deviation Error (RMSE) to measure the error rate on both Training and Development datasets. RMSE (Equation 1) is a commonly used in the machine learning community to measure the differences between the values predicted by a model and the ground truth [Dhanani et al., 2014].

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \tag{1}$$

Feature Selection

Feature selection is the process of selecting a subset of relevant features including variables or predictors to be used in a model for machine learning. The purpose of feature selection is to reduce the complexity of a model to more easily be interpreted. The benefit is three-fold: improving the prediction performance of the predictors, providing faster and most cost-effective predictors, and providing a better understanding of the underlying process that generated the data [Guyon and Elisseeff, 2003].

Since we have more features than data points, it tends to lead to overfitting [Smith and Somorjai, 2011]. Therefore feature selection is first performed on the data before applying machine learning. The feature selection algorithms used are popular and are taken from scikit-feature, a feature selection library [Li et al., 2016]: CIFE [Lin and Tang, 2006], Relief [Robnik-Šikonja and Kononenko, 2003], CFS [Hall

and Smith, 1999]. We will not go into detail as feature selection is not the main focus of the report.

Experimental Setup

We compared the proposed GP models against state-of-the-art machine learning models as mentioned in the previous section. For the ease of testing, all implementations of the algorithms except for GP ARD come from the popular machine learning library, Scikit Learn [Pedregosa et al., 2011]. We used the implementation of GP ARD from GPy, a Gaussian Processes framework in Python [GPy, since 2012]. The hyper-parameters are either determined by the defaults used in either libraries or some reasonable defaults were used. Each machine learning model is trained against the training set and thereafter tested against the development set using RMSE as the error metric. The entire experimental process is shown in Figure 4.
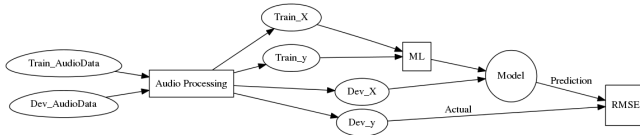

Figure 4: Experimental process

Results

We first ran the experiment across the dataset using all 526 features, without feature selection. As we would expect [Cawley and Talbot, 2014], the results are unacceptable as the ratio of the number of features to the number of data points is too high, resulting in possible overfitting. The results of the initial experiment is illustrated in Figure 5. We would expect the GP ARD would be able to theoritcally extract relevant features and improve prediction. However, we have observed experimentally that GP ARD performs poorly, along with other GPs.
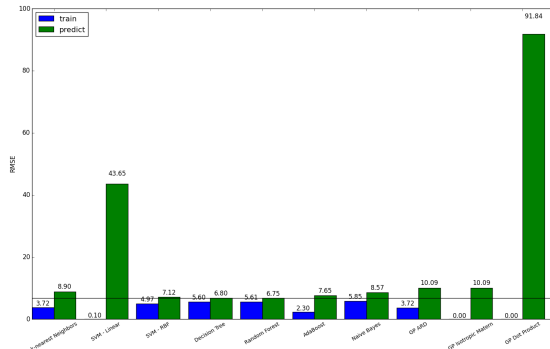

Figure 5: Results across all features

We repeated the experiment with feature selection and ran each of the feature subset gathered from the feature selection algorithms against each of the machine learning algorithms. We observed that Relief, CIFE and CFS selected a large number of MFCC features. The number of features in each feature subset is shown in Table 4. This confirms our assumption that MFCC gives the best predictive power in PHQ-8 depression severity prediction. Hence, we also ran the experiment using only MFCC features. The best results across all feature subsets are shown in Figure 6 and in Table 5. The line shown across the bar chart represents the baseline RMSE provided.

Feature Selection	Number of features
MFCC	12
CIFE	3
Relief	23
CFS	6
All	526

Table 4: Feature subsets
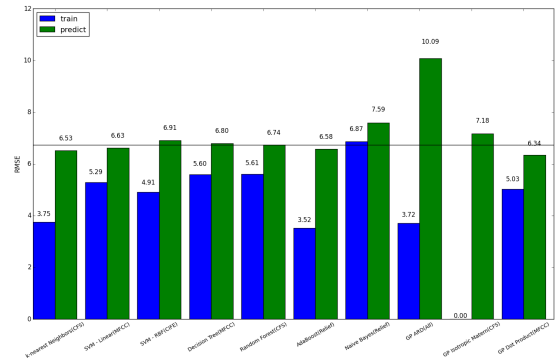
Figure 6: Best Results across all feature subsets

Algorithm	Subset	RMSE	
		Train	Dev
GP Dot Product	**MFCC**	**5.03**	**6.34**
AdaBoost	Relief	3.55	6.52
K-Nearest Neighbors	CFS	3.75	6.53
SVM - Linear	MFCC	5.29	6.63
Random Forest	CFS	5.61	6.75
Decision Tree	MFCC	5.60	6.80
SVM - RBF	CIFE	4.91	6.91
GP Isotropic Matern	CFS	0.00	7.18
Naive Bayes	Relief	6.87	7.59
GP ARD	All	3.72	10.09

Table 5: RMSE Results

As expected, the models perform better with the MFCC feature set. Unexpectedly, the simple GP dot product model,

trained with 12 features and 95 data points, outperforms all other machine learning models in our tests. Our results also confirms the initial assumption that MFCC is an appropriate feature set to be used in emotion and therefore depression prediction and that GP is applicable and feasible in predicting PHQ-8 scores.

Conclusion

Our work has succesfully shown that GP is a good model for this problem and can predict PHQ-8 better than state-of-the-art machine learning models. In addition to being on par or better at prediction, GP can inherently provide an estimate of prediction uncertainty. This allows the user to gauge the model's confidence of the prediction, and to make more informed decisions based on both the prediction and its uncertainty. We can also intelligently supplement more data to our training set based on the prediction uncertaintly. Therefore, after considering both results and GP's advantages, we conclude the GP Dot Product trained using MFCC feature set is a good model for depression prediction.

Further Work

For this experiment, we only used machine learning algorithms with their default parameters. An aspect that deserves further exploration is to perform automatic hyper-parameter optimization across all the machine learning algorithms to fine-tune each model's performance. In particular, we can try Hyperopt-sklearn [Komer, Bergstra, and Eliasmith, 2014] or GP based hyper-parameter tuner. We opine that with hyper-parameter tuning, we can predict PHQ-8 scores better and can have a more objective comparison of the different learning algorithms.

Contributions

- **Antoine Charles Vincent Garcia**: Scripting the program, setting up machine learning libraries, running tests and generation of the utility function.

- **Chan Jun Wei**: Scripting the program, setting up machine learning libraries, running tests and generation of the utility function.

- **Chen Tze Cheng**: Scripting the program, setting up machine learning libraries, running tests and generation of the utility function.

- **Eric Ewe Yow Choong**: Formatting the report as well as research and writing up of the technical approach section.

- **Han Liang Wee, Eric**: Retrieving data, testing as well as research and writing up of the technical approach section.

- **Ho Wei Li**: Research, vetting of the report and writing up of the motivation and introduction of the experiment.

References

Bowden, C. L. 2001. Strategies to reduce misdiagnosis of bipolar depression. *Psychiatric Services*.

Cawley, G. C., and Talbot, N. L. C. 2014. Kernel learning at the first level of inference. *Neural Networks* 53:69–80.

Cesar, J., and Chavoushi, F. 2013. Background paper 6.15, depression. *Priority Medicines for Europe and the World, "A Public Health Approach to Innovation"* BP 6.15.

Chavhan, Y.; Dhore, M. L.; and Yesaware, P. 2010. Speech Emotion Recognition using Support Vector Machine. *International Journal of Computer Applications* 1(20):8–11.

Cummins, N.; Scherer, S.; Krajewski, J.; Schnieder, S.; Epps, J.; and Quatieri, T. F. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication* 71:10–49.

Dhanani, A.; Lee, S. Y.; Phothilimthana, P.; and Pardos, Z. 2014. A comparison of error metrics for learning model parameters in bayesian knowledge tracing. Technical Report UCB/EECS-2014-131, EECS Department, University of California, Berkeley.

Dunner, D. L. 2003. Clinical consequences of under-recognized bipolar spectrum disorder. *Bipolar disorders* 5(6):456–463.

Fiske, A.; Wetherell, J. L.; and Gatz, M. 2009. Depression in older adults. *Annual Review of Clinical Psychology* 5:363–389.

GPy. since 2012. GPy: A gaussian process framework in python. `http://github.com/SheffieldML/GPy`.

Greenberg, P. E.; Fournier, A.-A.; Sisitsky, T.; Pike, C. T.; and Kessler, R. C. 2015. The economic burden of adults with major depressive disorder in the united states. *The Journal of Clinical Psychiatry* 76(2):155–162.

Guyon, I., and Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)* 3(3):1157–1182.

Hall, M. A., and Smith, L. A. 1999. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, 235–239. AAAI Press.

Hashim, N. W.; Wilkes, M.; Salomon, R.; Meggs, J.; and France, D. J. 2016. Evaluation of voice acoustics as predictors of clinical depression scores. *Journal of Voice* 0(0).

Jiaquan Xu, M.; Sheery L. Murphy, B.; Kenneth D. Kochanek, M.; and Brigham A. Bastian, B. 2016. Deaths: Final data for 2013. *National Vital Statistics Reports* Vol 64, No. 2.

Johnstone, T. 2001. *The effect of emotion on voice production and speech acoustics*. Ph.D. Dissertation, University of Western Australia.

Komer, B.; Bergstra, J.; and Eliasmith, C. 2014. Hyperopt-sklearn: Automatic hyper-parameter configuration for scikit-learn.

Kroenke, K.; Strine, T.; Spitzer, R. L.; Williams, J. B.; Berry, J. T.; and Mokdad, A. 2008. The phq-8 as a measure of current depression in the general popoulation. *Journal of Affective Disorders* 114(1-3):163–73.

Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Robert, T.; Tang, J.; and Liu, H. 2016. Feature selection: A data perspective. *arXiv:1601.07996*.

Lin, D., and Tang, X. 2006. *Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion*. Berlin, Heidelberg: Springer Berlin Heidelberg. 68–82.

Markov, K., and Matsui, T. 2013. Music Genre and Emotion Recognition Using Gaussian Processes. 2:2–3.

McCambridge, J.; Witton, J.; and 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.

Nasrallah, H. A. 2015. Consequences of misdiagnosis: Inaccurate treatment and poor patient outcomes in bipolar disorder. *The Journal of clinical psychiatry* 76(11):e1328–e1328.

of Health, U. D., and Services, H. 2015. Depression, what you need to know.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

RC, K.; P, B.; O, D.; and et al. 2003. The epidemiology of major depressive disorder: Results from the national comorbidity survey replication (ncs-r). *JAMA* 289(23):3095–3105.

Robnik-Šikonja, M., and Kononenko, I. 2003. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning* 53(1):23–69.

Smith, A. L., and 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.

Smith, I. C. P., and Somorjai, R. L. 2011. Deriving biomedical diagnostics from nmr spectroscopic data. *Biophysical Reviews* 3(1):47–52.

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

Valstar, M. F.; Gratch, J.; Schuller, B. W.; Ringeval, F.; Lalanne, D.; Torres, M.; Scherer, S.; Stratou, G.; Cowie, R.; and Pantic, M. 2016b. AVEC 2016 - depression, mood, and emotion recognition workshop and challenge. *CoRR* abs/1605.01600.

Wolpert, L. 2001. Stigma of depression - a personal view. *British Medical Bulletin* 57(1):221–224.

World Health Organization. 2004. The global burden of disease: 2004 update. Technical report, World Health Organization.