

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 debilitating mental illness that has good prognosis given early detection and treatment. In recent years, youth suicide rates in Singapore have increased at an alarming rate. Given that suicide is often a byproduct of depression, it is preventable. General practitioners are the most accessible source of help but lack of time, attention, skills and the presence of social stigma prevent accurate diagnosis of depression. In our project, we propose a general framework to pre-screen patients before they make appointments in general practice for depression. Outputs from the Gaussian Process will be used to determine if the patient requires medical attention and if so, the urgency of the case. In addition, we produced a proof-of-concept prototype illustrating our framework. We also further note that the framework can be extended to other psychiatric disorders.

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.

General practitioners (GPs) are often dubbed as the first line in the provision of healthcare as they form the interface between specialized healthcare providers and patients. Compared to specialists, GPs are considered to be more accessible, less stigmatising and more comprehensive, thus they are expected to be a depression patient's best chance at getting initial help [Rothman and Wagner, 2003]. Indeed, in the US, almost 75% of depression patients who seek help, do so with a general practitioner [Goldman, Nielsen, and Champion, 1999].

However, in general practice, 48.4% of patients suffering from depression go unrecognized [Kessler et al., 2003]. During a visit to the GP, a patient's symptoms that are not physical in nature may not be inquired or brought up. Even if some GPs show interest in patients' psychosocial issues, they likely lack the requisite interviewing skills to properly elicit the relevant history while others lack the sensitivity to affective and nonverbal patient cues [Badger et al., 1994]. In addition, GPs have a plethora of tasks during each consultation such as assessment and management of ongoing general medical issues, prevention and health maintenance, as well as paperwork. This severely limits the time budget that remains for assessing mental health issues such as depression and further raises the barrier of accurate diagnosis and proper treatment [Telford et al., 2002].

Motivation

The most recent study conducted by the Institute of Mental Health has shown that Major Depressive Disorder (MDD) is the most common mental disorder in Singapore. An estimated 1 in 17 people having suffered from MDD at some point in their lives which is almost twice that of the next most common disorder, alcohol abuse [Chong et al., 2012]. In 2015, despite falling overall suicide rates, and a shrinking proportion of youths in the ageing population of Singapore, the number of suicide deaths in youths between

10-19 of age was twice as many as the year before and it is also the highest in 15 years [Samaritans of Singapore, 2016]. This is a worrying trend and as suicide is closely correlated with depression, due to the absence of a more recent study, we extrapolate that the prevalence of MDD has increased over the last 5 years.

Diagnosis and treatment of depression in Singapore is difficult. Apart from the highlighted difficulties that GPs face, there are also significant delays in specialized treatment, which can extend up to several months due to the presence of a long referral chain to a psychiatrist, inherent in the public health system. These barriers are compounded by sociocultural factors in asian societies such as stigmatization where a majority of the public, including mental health professionals, think that they should be protected from people with mental health problems or that these people should be locked up. [Chong et al., 2009]. This highlights a need for a better process for diagnosing depression.

Objective

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 a recent study, Gaussian Process Dot Product trained using the MFCC feature set of acoustic analysis was found to be a good model for the prediction of depression incidence and severity with better accuracies than other state-of-the-art machine learning models [Han et al., 2016].

We propose a framework based on the outputs of the Gaussian Process Dot Product on MFCC model. The framework lowers the barriers of diagnosis and treatment of depression by supplanting the diagnosis process in general practice with the predictions of the model in most cases, requiring the GP's attention only in ambiguous cases. In univocal cases, patients are referred directly to specialists for treatment, thereby reducing the referral chain to a single step. Specialized healthcare providers are also given the predicted Patient Health Questionnaire 8 (PHQ-8) score of the patient to better prioritize urgent cases, improving appointment scheduling. On the side of GPs, time and resources saved can be utilized to provide better quality healthcare in all other aspects.

Modelling and Approach

Leveraging on the success of the modelling of depression prediction of Personal Health Questionnaire depression scale (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. In this manner, we cut down on the number of

appointments reducing workload on the staff. Moreover, our process allows prioritizing of patient's appointments by their PHQ-8 score which represents the severity of depression. Hence, we would need to use two gaussian processes (GP): 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. We note that the process is a general framework and that different clinics have different operating processes, clinics should tweak the framework to better fit their needs. Hence, we would omit some detail in our model and leave them to the implementer.

Automation Flow

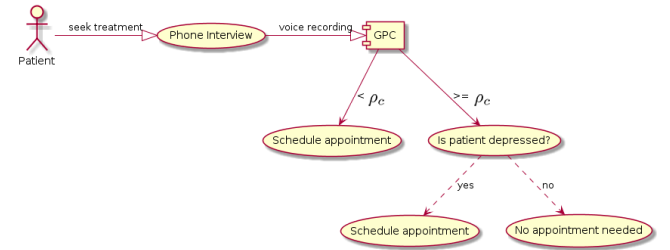


Figure 1: Automation Flow

Figure 1 represents the flow of a person who is seeking for medical attention from a psychiatric clinic. Prior to medical treatment, the patient would be calling up the clinic to arrange an appointment time and date. Before the clinic fixes the appointment, a short phone interview consisting of several questions is conducted. Refer to the sample questions in the section titled 'Proof of Concept'. The objective of the interview is to record the patient's audio signal as he/she answers the questions. The recordings will serve as an input for GPC, which is a classifier that will output both a probability estimate and a label, determining if the patient is suffering from depression or not. The probability estimate represents the confidence on the predicted label. We can then define a specific pre-defined probability estimate ρ_c , such that we have two scenarios as listed below. In the illustration below, we make the argument based on a probability estimate p which is based on some audio recording collected.

1. Confident of prediction ($p \geq \rho_c$)
GPC is confident of its prediction, allowing us to make decision with confidence on the predicted label. Then, we make the decision based on whether the patient is depressed or not depressed as predicted by GPC. A depressed patient will be given an appointment (refer to Figure 2) for treatment, whereas a non-depressed patient would be informed that he/she is not required to go for an appointment.
2. Not confident of prediction ($p < \rho_c$)
GPC is not confident of its prediction, we cannot rely on the predicted label. In this case, patient can proceed with the appointment scheduling (Figure 2). We note

that he/she would need to be diagnosed by the clinician and treatment if appropriate. We would also obtain the ground truth from the diagnosis and use that to train the GPC/GPR on this datapoint.

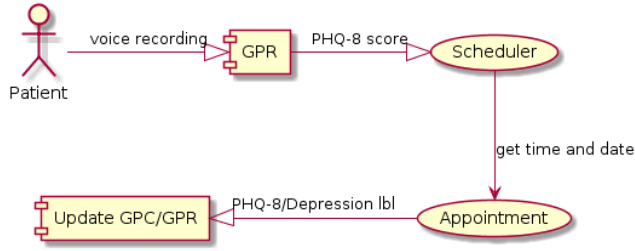


Figure 2: Flow for Scheduling an Appointment

Figure 2 represents the case that he/she needs an appointment. Similarly, the patient’s voice is used as an input for the GPR which predicts the PHQ-8 score. PHQ-8 score determines the levels of depression (from 0-24, 0 means no depression and 24 means very depressed) of the patient, allowing the staff to prioritize appointments based on the PHQ-8 score. We would prioritize patients who are more depressed (predicted by the GPR with a higher PHQ-8 score) will have higher priority and vice versa. Through the appointment(s) with the psychiatrist, the ground truth of the depression and PHQ-8 labels will be obtained and can be used to train both GPC and GPR if required (if this datapoint was predicted with no confidence i.e. $p < \rho_c$). As more data points are observed and subsequently used for training, both GPC’s and GPR’s accuracy would improve incrementally.

Insights to Our Model

Gaussian Processes : In our modelling, we used two GPs for regression and classification tasks respectively. The GP models that we used here are modelled similarly to the modelling done in project 1. GPR predicts the PHQ-8 score, while 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 (provided as part of the dataset) respectively. We applied the same audio signal processing techniques to the audio files as per project 1. We used GP with a dot-product kernel, with the Mel Frequency Cepstral Coefficients (MFCC) feature subset as we have seen the good theoretical and experimental results the feature subset had produced in project 1. More 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 which provides a single depression estimate over a time period [Valstar et al., 2016a].
2. Speech signals extracted from people suffering from depression should share some similarities and thus be admissible for prediction with the Gaussian Process models [Cummins et al., 2015].

In our formulation GPs must be able to train on new data. Since the GP models receive new data incrementally, we cannot use an offline GP as we have described in project 1. Hence, we would need to tweak the GP to fit this problem, needing a GP that learns dynamically, adapting to new data as it becomes available. 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 which will be used in our project.

Phone Interview : We noted that regardless of the questions asked, a depressed person will still exhibit signs of depression in his speech [Kroenke et al., 2008]. Hence, the questions asked are irrelevant. A sample of questions are presented in the section titled ‘Proof of Concept’.

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 on the audio recording. As mentioned in the modelling, we define a particular probability ρ_c . If the GPC predicts with $p \geq \rho_c$, then we can trust the predicted label and continue to decide appropriately based on the label whereas if the GPC predicts with $p < \rho_c$, then we cannot trust the predicted label and refer the patient to an appointment with the psychiatrist to obtain more data. 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. We observe that with a $\rho_c = 70\%$, it would predict the depression label with 100% accuracy. Henceforth, in this paper we will define $\rho_c = 70\%$.

$p_c(\%)$	Depression	No Depression
[0, 50)	-	-
[50, 60)	60%(3/5)	80%(44/55)
[60, 70)	100%(1/1)	84.2%(16/19)
[70, 80)	-	100%(4/4)
[80, 90)	-	100%(1/1)
[90, 100]	-	-
-	-	90%(9/10)

Table 1: Relationship between probability and accuracy

Scheduling : For the cases when the person needs to be scheduled for an appointment with the psychiatrist, we can use the GPR to determine the predicted PHQ-8 scores. With the predicted scores, we can prioritise certain higher risk individuals over the rest. We were inspired by triage algorithms used in emergency services [Shah et al., 2015; Oredsson et al., 2011], where priority of one’s treatments are decided by the severity of their ailments [Wikipedia, 2016]. We have seen that this method dispenses limited resources with efficiency [Rosedale et al., 2011] throughout many emergency services world-wide. Similarly, we want to

apply that idea into psychiatric clinic, which are also facing a shortage of resources. Since the depression clinic is understaffed and struggling to keep up with number of patients, it is wise to prioritize 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 of appointments, prioritizing people with higher predicted PHQ-8 scores.

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 examination if the person in question is depressed or not and administer the appropriate treatment. From the appointment with the psychiatrist, he/she 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. We also prioritize the clinic's resources based on a person's depression severity, allowing better allocation of resources. This ensures that the clinic's resources are directed to people who need them most.

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, a pre-screening process whose accuracy improves incrementally over time, and a better allocation of resources.

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 listed 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

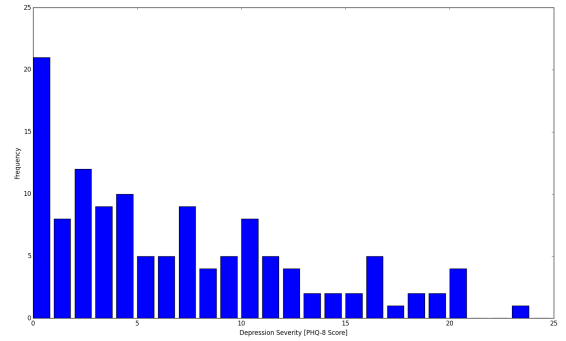


Figure 3: Histogram of the PHQ-8 scores

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 corresponding 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.

	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

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. 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].

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

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.

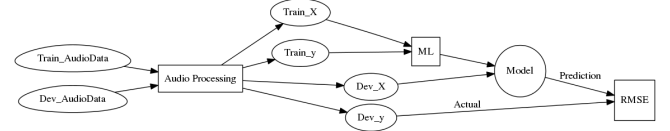


Figure 4: Experimental process

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 theoretically extract relevant features and improve prediction. However, we have observed experimentally that GP ARD performs poorly, along with other GPs.

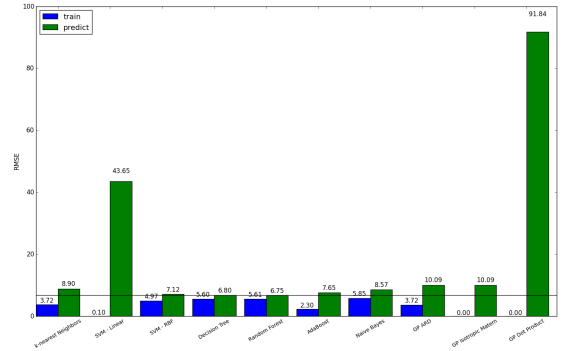


Figure 5: Results across all features

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

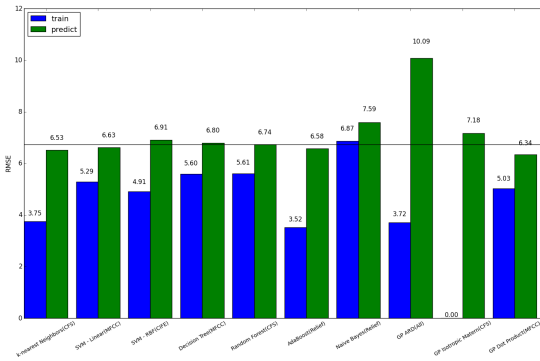


Figure 6: Best Results across all feature subsets

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 successfully 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 uncertainty. 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

- Badger, L.; deGruy, F.; Hartman, J.; Plant, M.; Leeper, J.; Ficken, R.; Maxwell, A.; Rand, E.; Anderson, R.; and Templeton, B. 1994. Psychosocial interest, medical interviews, and the recognition of depression. *Archives of Family Medicine* 3(11):974.
- 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.

- Chong, S.; dS Verma; Vaingankar, J.; Chan, Y.; Wong, L.; and Heng, B. 2009. Perception of the public towards the mentally ill in a developed asian country. *Social Psychiatry and Psychiatric Epidemiology* 42(9):734–739.
- Chong, S. A.; Abidin, E.; Vaingankar, J. A.; Heng, D.; Sherbourne, C.; Yap, M.; Lim, Y. W.; Wong, H. B.; Ghosh-Dastidar, B.; Kwok, K. W.; and Subramaniam, M. 2012. A population-based survey of mental disorders in singapore. *ANNALS Academy of Medicine Singapore* 41:49–66.
- 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.
- Goldman, L. S.; Nielsen, N. H.; and Champion, H. C. 1999. Awareness, diagnosis and treatment of depression. *Journal of General Internal Medicine* 14(9):569–580.
- 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.
- Han, E. L.; Chan, J.; Ho, W. W.; Chen, T.; Ewe, E. Y.; and Garcia, A. C. 2016. Depression prediction with gaussian process. CS4246 AI Planning and Decision Making Project 1.
- 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.
- Kessler, R.; Berglund, P.; Demler, O.; and et al. 2003. The epidemiology of major depressive disorder: Results from the national comorbidity survey replication (ncs-r). *JAMA* 289(23):3095–3105.
- Komer, B.; Bergstra, J.; and Eliasmith, C. 2014. Hyperopt-sklearn: Automatic hyperparameter 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 population. *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.
- 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.
- Oredsson, S.; Jonsson, H.; Rognes, J.; Lind, L.; Göransson, K. E.; Ehrenberg, A.; Asplund, K.; Castrén, M.; and Farrohknia, N. 2011. A systematic review of triage-related interventions to improve patient flow in emergency departments. *Scandinavian journal of trauma, resuscitation and emergency medicine* 19(1):1.
- 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.
- Robnik-Šikonja, M., and Kononenko, I. 2003. Theoretical and empirical analysis of relief and rrelief. *Machine Learning* 53(1):23–69.
- Rosedale, K.; Smith, Z.; Davies, H.; and Wood, D. 2011. The effectiveness of the south african triage score (sats) in a rural emergency department. *SAMJ: South African Medical Journal* 101(8):537–540.
- Rothman, A., and Wagner, E. 2003. Chronic illness management: what is the role of primary care? *Annals of Internal Medicine* 138(3):256–261.
- Samaritans of Singapore. 2016. National suicide statistics.
- Shah, S.; Patel, A.; Rumoro, D. P.; Hohmann, S.; and Fullam, F. 2015. Managing patient expectations at emergency department triage. *Patient Experience Journal* 2(2):31–44.
- Smith, I. C. P., and Somorjai, R. L. 2011. Deriving biomedical diagnostics from nmr spectroscopic data. *Biophysical Reviews* 3(1):47–52.
- Telford, R.; Hutchinson, A.; Jones, R.; Rix, S.; and Howe, A. 2002. Obstacles to effective treatment of depression: a general practice perspective. *Family Practice* 19(1):45–52.
- 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.
- Wikipedia. 2016. Triage — Wikipedia, the free encyclopedia. <http://en.wikipedia.org/w/index.php?title=Triage&oldid=748740826>. [Online; accessed 10-November-2016].
- World Health Organization. 2004. The global burden of disease: 2004 update. Technical report, World Health Organization.