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 [**?**]. 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 [**?**].

Left unchecked, depression increases risk for morbidity, suicide, decreased cognitive and social functioning, self-neglect, and early death [**?**]. 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 [**?**].

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 [**?**].

Motivation and Objective

The Personal Health Questionaire 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 [**?**]. 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 [**?**]. 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 [**?**]. 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. [**?**].

In the past decade, there have been research successfully correlating emotion with voice production and speech acoustics [**?**]. 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 [**?**].

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 emo-

tion recognition [**?**], GP models have been proven to consistently outperform SVM models on the task of music emotion recognition [**?**]. 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 Personal Health Questionaire 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

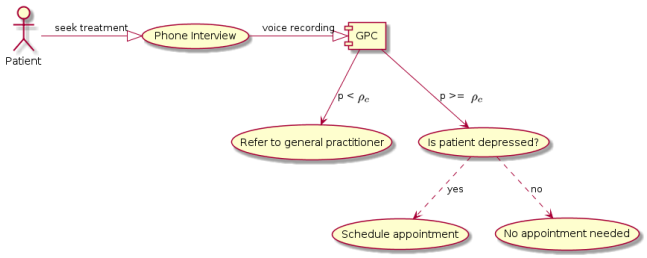

Figure 1: Automation Flow

Figure **??** 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 the 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 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 **??**) 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 will be referred to a general practitioner for further observations (Figure **??**). We note that he/she would need to be diagosed 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.

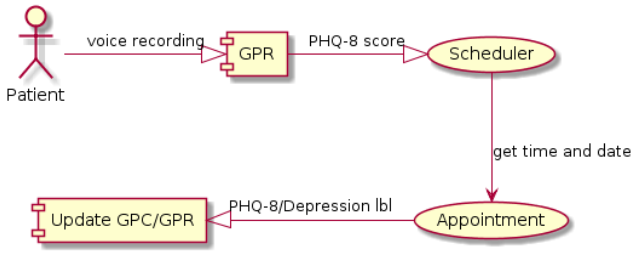

Figure 2: Flow for Scheduling an Appointment

Figure **??** represents the case that he/she needs an appointment. Similarly, the patient's voice is use 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 use 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 use 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) [**?**], 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 theorical 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 [**?**].
2. Speech signals extracted from people suffering from depression should share some similarities and thus admissible for prediction with the Gaussian Process models [**?**].

In our formulation GPs must be able to train on new data. Since the GP models recieve 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 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 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 [**?**]. 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 **??**. We observe that with a $\rho_c = 70\%$, it would predict the depression label with 100% accuraccy. Henceforth, in this paper we will define $\rho_c = 70\%$.

Scheduling : For the cases when the person needs to be scheduled for an appointment with the psycharist, 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 [**?**; **?**], where priority of one's treatments are decided by the severity of their ailments [**?**]. We have seen that this method dispenses limited resources with efficiency [**?**] 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 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 of appointments, piroritizing people with higher predicted PHQ-8 scores.

$p_c(\%)$	**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 accuraccy

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 [**?**; **?**; **?**] 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. 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 [**?**]. 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) [**?**]. 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

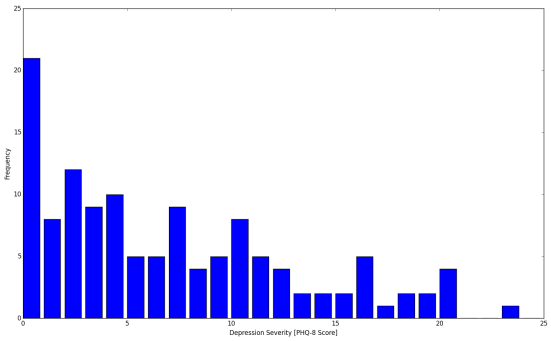

Figure 3: Histogram of the PHQ-8 scores

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) [**?**]. 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 **??**. The data provided are split into 2 sets: training and development. An overview of the data is given in Table **??**.

	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 **??**.

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 $= 6.7418$ [**?**]. 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 **??**) is a commonly used in the machine learning community to measure the differences between the values predicted by a model and the ground truth [**?**].

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

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 [**?**].

Since we have more features than data points, it tends to lead to overfitting [**?**]. 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 [**?**]: CIFE [**?**], Relief [**?**], CFS [**?**]. 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 [**?**]. We used the implementation of GP ARD from GPy, a Gaussian Processes framework in Python [**?**]. 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 **??**.

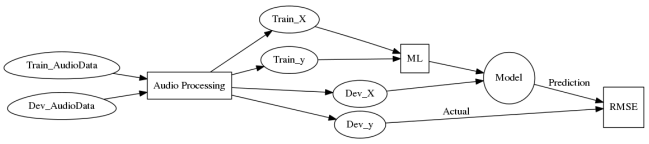

Figure 4: Experimental process

Results

We first ran the experiment across the dataset using all 526 features, without feature selection. As we would expect [**?**], 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 **??**. 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.

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 **??**. 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 **??** and in Table **??**. The line shown across the bar chart represents the baseline RMSE provided.

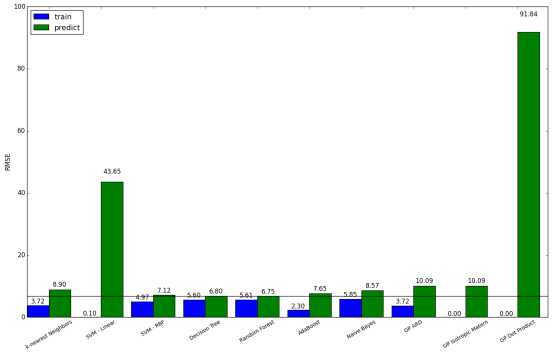

Figure 5: Results across all features

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

Table 4: 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.

Here are the 95% confidence interval for the predicted PHQ8 score for each subject of the test set.

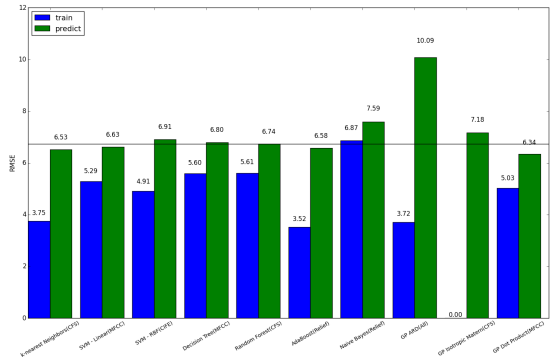

Figure 6: Best Results across all feature subsets

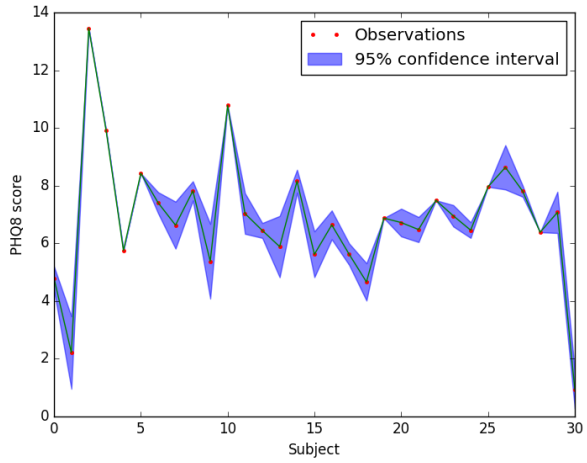

Figure 7: Confidence interval for each subject of the test set

Methodology to investigate depression through voice

1. Using the same application to record sound – Example: Call Recorder
2. Ask the questions to the person during recording. The answer gathered is not important, as long as we could collect the voice. The voice recorded is preferably over 1 minute.
   (a) Ask for the person's personal contact, such as email, or phone number, or address.
   (b) Asking the person's availability to come to the clinic, the date and time, or period of time
   (c) Asking the person's individual health screening history: Is this the first time you find a psychologist to talk with? Do you have any medical records in the past?
   (d) If the previous questions doesn't let the person's voice recording long enough, the agent might ask the person, is he/she is willing to help out in a questionnaire – (Can even collaborate with some questionnaires)
3. It is not necessary to ask all the questions above as long as the patient's voice recording can pass about a minute for each session.
4. Avoid saying something like, "Cheer up!", "Lighten up!" to provide help, or giving suggestions like "Take a hot bath. That's what I always do when I upset!" to the patients. Each phone call should obey the principle of "Ask for information and NOT providing unprofessional guidance". [**?**]
5. End with, thank you Sir/Ms for your time. We will contact you within 3 working days.
6. Cut the voice recordings of the patient out from the whole phone call recordings manually and pass it to the GP model for processing.

Works done on Depression Classification

1. I extended the result of Depression Regression to Depression Classification.
2. I tried using the predicted score as the feature but obtained poor result:

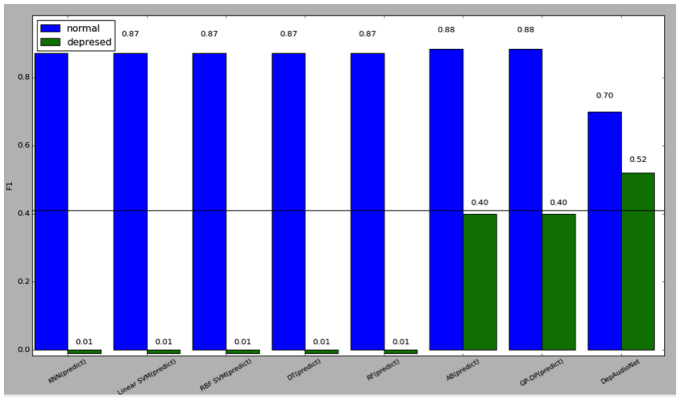

Figure 8: First results of the classification

3. So I changed back to speech feature and obtained a very good result by using GP Dot Product!
4. The GP-dot product obtains a depression F1 score of 0.67 whereas the normal F1 score of 0.92 which beats DeapAudioNet 0.52 for depression and 0.70 for normal. It also beats the baseline which has 0.41 for depression and 0.58 for normal
5. Other than the graph, the following table is generated:

	F1	**Precision**	**Recall**
Gaussian Process (Dot-Product)	0.67(0.92)	0.57(0.96)	0.8(0.88)
DepAudioNet	0.52(0.7)	0.35(1.0)	1.0(0.54)
BaseLine	0.41(0.58)	0.27(0.94)	0.89(0.42)

Table 6: F1, Precision, Recall and Accuracy for each model

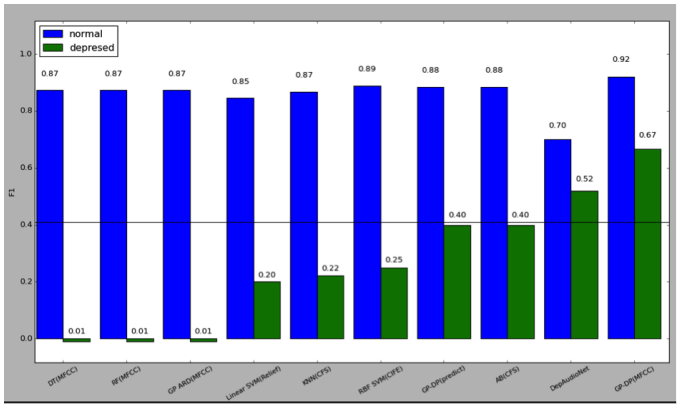

Figure 9: Results of the classification with the GP Dot Product

Probability Range	Accuracy	
Normal		Depression
[0[10[20[30[40[50[60[70[80[90100%	No Sample	No Sample
NAN	90% (9 out of 10)	90% (9 out of 10)

Table 7: Correlationship between Probabilities and Accuracy of classification

According to the table, we could certainly said that the interval [30The resulting predicted result when getting a NAN is always not depressed (0), which might not be the case. One possible explanation the model keeps choosing the non-depressed when it is NAN, so it depends on the luck, when the model gets NAN, we probably needs to explore more. How to balance between Exploration and Exploitation?

Performance Metrics

We are using F1 score to be the performance metrics of the depression classification task, so in the information retrieval world, there is something call precision and something call recall. Precision is among the result I retrieved, how accurate it is. (correct retrieved result / total retrieved) High Precision would show that the model retrieves accurate result. Recall is among the result I retrieved, how relevant it is. (correct retrieved result / total relevant) High Recall would show that the model retrieves result which is highly relevant to the query. So for our case, in order to calculate the precision and recall of depression classification, we first look at the following table:

	True - Well predicted	False - Wrongly predicted
Depressed? Yes (Positive)	True Positive (tp)	False Positive (fp)
Depressed? No (Negative)	True Negative (tn)	False Negative (fn)

Table 8: Precision and Recall for Depression Classification

Imagine we are retrieving a list of "Is his depressed?" list, the precision would be tp / (tp + fp) and so this is the formula for precision(P) is:

$$\mathrm{P} = \frac{tp}{tp + fp} \tag{2}$$

Then what are the relevant list? The one who is correctly diagnosed as depressed and the one who is falsely diagnosed as not depressed are all relevant to depression and thus the formula for recall (R) would be:

$$\mathrm{R} = \frac{tp}{tp + fn} \tag{3}$$

Having a high precision and low recall or low precision and high recall are not good, like if you have no relevant document, then recall is infinite. If nothing is retrieved, precision is also infinite, so we need something to strike for a balance, and F1 is the balance between precision and recall. It is defined as the harmonic mean of precision and recall, and thus the formula for F1 would be:

$$\mathrm{F1} = \frac{1}{\frac{1}{P} + \frac{1}{R}} = \frac{PR}{P + R} \tag{4}$$

Moreover, we also added accuracy, which is every result that is predicted correctly / total result:

$$\text{Accuracy} = \frac{tp + tn}{tp + fp + tn + fn} \tag{5}$$

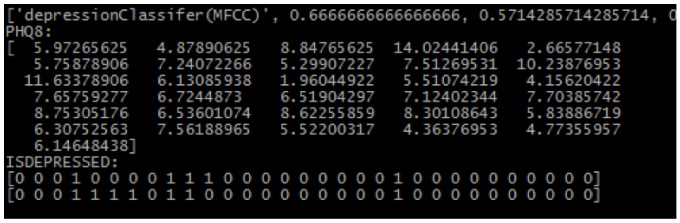

Figure 10: Results of the accuracy

Methodology to investigate depression through voice

1. Using the same application to record sound – Example: Call Recorder
2. Ask the questions to the person during recording. The answer gathered is not important, as long as we could collect the voice. The voice recorded is preferably over 1 minute.
   (a) Ask for the person's personal contact, such as email, or phone number, or address.
   (b) Asking the person's availability to come to the clinic, the date and time, or period of time
   (c) Asking the person's individual health screening history: Is this the first time you find a psychologist to talk with? Do you have any medical records in the past?

(d) If the previous questions doesn't let the person's voice recording long enough, the agent might ask the person, is he/she is willing to help out in a questionnaire – (Can even collaborate with some questionnaires)

3. It is not necessary to ask all the questions above as long as the patient's voice recording can pass about a minute for each session.
4. Avoid saying something like, "Cheer up!", "Lighten up!" to provide help, or giving suggestions like "Take a hot bath. That's what I always do when I upset!" to the patients. Each phone call should obey the principle of "Ask for information and NOT providing unprofessional guidance". [**?**]
5. End with, thank you Sir/Ms for your time. We will contact you within 3 working days.
6. Cut the voice recordings of the patient out from the whole phone call recordings manually and pass it to the GP model for processing.

Works done on Depression Classification

1. I extended the result of Depression Regression to Depression Classification.
2. I tried using the predicted score as the feature but obtained poor result:

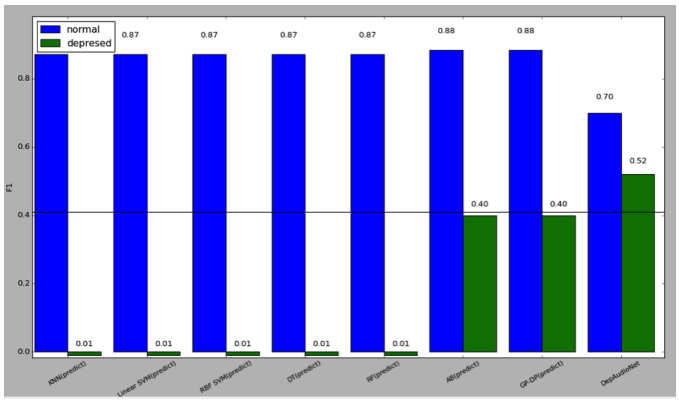

Figure 11: First results of the classification

3. So I changed back to speech feature and obtained a very good result by using GP Dot Product!
4. The GP-dot product obtains a depression F1 score of 0.67 whereas the normal F1 score of 0.92 which beats DeapAudioNet 0.52 for depression and 0.70 for normal. It also beats the baseline which has 0.41 for depression and 0.58 for normal
5. Other than the graph, the following table is generated:

According to the table, we could certainly said that the interval [30The resulting predicted result when getting a NAN is always not depressed (0), which might not be the case. One possible explanation the model keeps choosing the non-depressed when it is NAN, so it depends on the luck, when

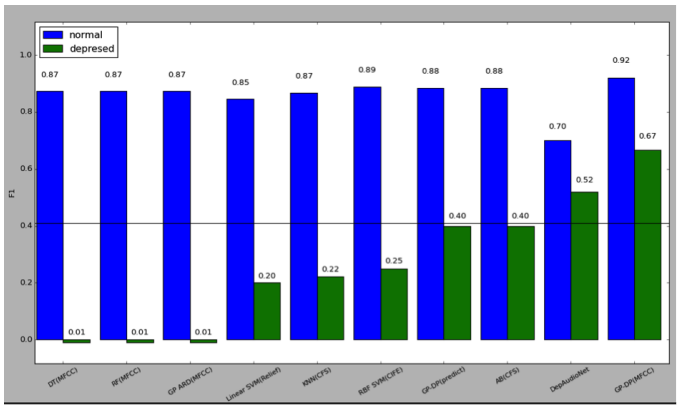

Figure 12: Results of the classification with the GP Dot Product

	F1	Precision	Recall
Gaussian Process (Dot-Product)	0.67(0.92)	0.57(0.96)	0.8(0.88)
DepAudioNet	0.52(0.7)	0.35(1.0)	1.0(0.54)
BaseLine	0.41(0.58)	0.27(0.94)	0.89(0.42)

Table 9: F1, Precision, Recall and Accuracy for each model

~~Probability Range~~	Accuracy	
Normal		Depres
[0[10[20[30[40[50[60[70[80[90100%	No Sample	No San
NAN	90% (9 out of 10)	90% (9 ou

Table 10: Correlationship between Probabilities and Accuracy of classification

the model gets NAN, we probably needs to explore more. How to balance between Exploration and Exploitation?

Performance Metrics

We are using F1 score to be the performance metrics of the depression classification task, so in the information retrieval world, there is something call precision and something call recall. Precision is among the result I retrieved, how accurate it is. (correct retrieved result / total retrieved) High Precision would show that the model retrieves accurate result. Recall is among the result I retrieved, how relevant it is. (correct retrieved result / total relevant) High Recall would show that the model retrieves result which is highly relevant to the query. So for our case, in order to calculate the precision and recall of depression classification, we first look at the following table:

	True - Well predicted	False - Wrongly
Depressed? Yes (Positive)	True Positive (tp)	False Positi
Depressed? No (Negative)	True Negative (tn)	False Negati

Table 11: Precision and Recall for Depression Classification

Imagine we are retrieving a list of "Is his depressed?" list, the precision would be tp / (tp + fp) and so this is the formula for precision(P) is:

$$\mathrm{P} = \frac{tp}{tp + fp} \tag{6}$$

Then what are the relevant list? The one who is correctly diagnosed as depressed and the one who is falsely diagnosed as not depressed are all relevant to depression and thus the formula for recall (R) would be:

$$\mathrm{R} = \frac{tp}{tp + fn} \tag{7}$$

Having a high precision and low recall or low precision and high recall are not good, like if you have no relevant document, then recall is infinite. If nothing is retrieved, precision is also infinite, so we need something to strike for a balance, and F1 is the balance between precision and recall. It is defined as the harmonic mean of precision and recall, and thus the formula for F1 would be:

$$\mathrm{F1} = \frac{1}{\frac{1}{P} + \frac{1}{R}} = \frac{PR}{P + R} \tag{8}$$

Moreover, we also added accuracy, which is every result that is predicted correctly / total result:

$$\text{Accuracy} = \frac{tp + tn}{tp + fp + tn + fn} \tag{9}$$

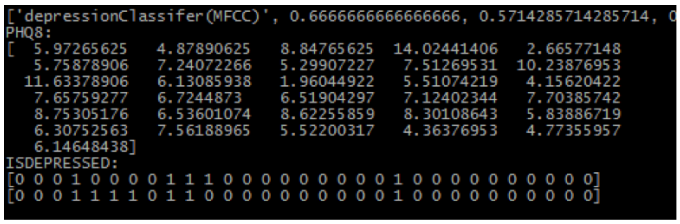

Figure 13: Results of the accuracy

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 [**?**] 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.