CS4246 AI Planning and Decision Making - Project 1
Depression Prediction

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 barrier and decrease the accuracy of a diagnostic test. In this study we aim to prove that prediction of depression severity on the PHQ-8 scale with voice acoustics is possible through the use of Gaussian Process, potentially lowering entry barriers to diagnostic tests and aiding early detection.

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 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 [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 some success, proving that it is an effective indicator of depression severity [Hashim et al., 2016].

The aim of the present study is to determine if acoustic measures of voice, characterizing specific spectral and timing proerties, predict predetermined clinical ratings of depression severity on the PHQ-8 scale in a sample of patients, using an existing dataset of their voice recordings.

Future Applications

Findings from this experiment will reinforce results from similar experiments and make a stronger case for the possibility of lowering the entry barriers of depression diagnostic tests and at the same time, increasing the accuracy of depression diagnosis by abstracting the Hawthorne Effect. Diagnostic tests can be designed in a friendly way that does not require the patient to consciously answer questionaires such as the PHQ-8 or even to have complete awareness of the diagnostic process. By taking into account the Hawthorne Effect and the status quo of the social stigma and working around them, diagnoses can be made easier, faster and more accurately and with that, education and treatment can begin early before the mental condition of the patient deteriorates.

Modelling and Approach

As Gaussian Process Regression Model is a new state of the art machine learning model, we are trying to apply it for depression prediction with the hope that the accuracy of depression prediction will be improved.

The following assumptions have been made for the experiment so that the GP Regression Model is suitable for depression prediction:

1. For any depressed individual, it is possible to identify depression via his/her speech. [Cummins et al., 2015]
2. Depression Prediction is an event-based recognition which provides a single depression estimates over a certain amount of time. [Valstar et al., 2016a]
3. The speech signals extracted from different depressed people should share some similarities and thus suitable for prediction with the Gaussian Process Model. For example, a diminished prosody and monotonous and "lifeless" sounding speech is indicative of depression.[Cummins et al., 2015]

Qualitative Advantages

GP is exceptionally useful in this application as it enables us to perform prediction on various training data including outliers or those with irregular sampling rates. As our assumption of we are able to determine of depressed person according to the speech and all the signals extracted share some similarities, we could make use of the training data in GP to perform prediction for a depressed person based on the speech of the person.

With GP, we can obtain a probability distribution such as the predictive mean and variance. These properties have a significant role to play in estimating the depression levels. For instance, the mean could be use as the baseline for an average working-human mood. The variance here would give us a picture to how good our existing data is. So, if some area of the probability distribution gives very high variance, it means that we don't have enough data for that area. Therefore, using the properties of GP we could determine how well our database is and we can find out what is the average depression level for everyone in the world nowadays.

In addition, GP produces truly probabilistic outputs with an explicit degree of prediction uncertainty. The prediction uncertainty can be further analyse to produce useful results. In addition, there exist algorithms for GP hyperparameter learning—something the other Machine Learning methods like SVM framework lacks. It is very useful as hyperparameter estimation could provide improvement to the Machine Learning algorithm especially in sound/speech recognition. [Hashimoto et al., 2015]

Previously, SVM gave a very high accuracy of prediction in speech emotion recognition, scoring about 94 % of accuracy in classification of emotions. [Chavhan, Dhore, and Yesaware, 2010]. However, in a more recent research, GP actually beats SVM in emotion recognition. In the paper "Music Genre and Emotion Recognition Using Gaussian Processes" [MARKOV and MATSUI, 2013], both music genre classification and music emotion estimation tasks, the GP performed consistently better than the SVM.

Therefore, as we are trying to prediction on depression, we would like to use GP to process our data to produce a more accurate prediction result.

Requirements

Aside from the audio recordings, our GP model makes use of the following:

1. **Audio Features**
   What are the differences between sounds? To differentiate between sounds, we need to learn about audio features, which will be described below:
   (a) **Energy**
   The Energy feature of a sound refers to the loudness of the sound at various timeframes, hence it is obvious that energy of a sound is directly proportional to the amplitude of the soundwave. This shows that the higher the energy, the louder the sound is going to be.
   (b) **Mel Frequency Cepstral Coefficients (MFCC)**
   Mel Frequency Cepstral (MFC) is a representation of short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. The greatest benefit of using MFCC is that the scale approximates the human's auditory system response more closely, hence it allows for a better representation of sound. In order to obtain MFCC, Fourier transform is performed on the sound signals.
   (c) **Magnitude Spectrum**
   Magnitude spectrum can be produced by converting the input signal of an audio into frames. Fast Fourier Transform (FFT) is performed on each frames and this will form the Magnitude Spectrum.

(d) **Zero-Crossing Rate**
Zero crossing rate is the rate of sign-changes along a signal. This feature is extremely useful in speech recognition and music information retrieval.

2. **Kernels**
The similarity measure of all features is usually called kernel. As Gaussian Process uses Kernels to predict the value for an unseen point from training data, understanding different kernels that we are going to use is extremely useful.
The kernels that would we are intested in is shown below (The kernels are assumed to be defined on two samples $x = (x_1 x_2 x_3 ... x_n)$ and $x' = (x_1{}' x_2{}' x_3{}' ... x_n{}')$, represented as feature vectors in some input space):

(a) **Radial Basis Function (RBF)**
The RBF kernel is defined as

$$K(x, x') = \exp^{-\frac{||x-x'||^2}{2\sigma^2}} \quad (1)$$

$||x - x'||^2$ represents Squared Euclidean Distance And $\sigma > 0$ can either be a scalar (isotropic variant of the kernel) or a vector with the same number of dimensions as the inputs X (anisotropic variant of the kernel).
It is also known as the squared exponential kernel. This kernel is infinitely differentiable, which implies that GPs with this kernel as covariance function have mean square derivatives of all orders, and are thus very smooth.

(b) **Matern**
The class of Matern kernels is a generalization of the RBF and the absolute exponential kernel parameterized by an additional parameter nu. The smaller the value of nu, the less smooth the approximated function is. For nu=∞, the kernel becomes equivalent to the RBF kernel and for nu=0.5 to the absolute exponential kernel. Important intermediate values are nu=1.5 (once differentiable functions) and nu=2.5 (twice differentiable functions).

(c) **Automatic Relevance Detection (ARD) Kernels**
If we chose the anisotropic variant of the RBF and Matern kernel, then these kernels would be also called as ARD kernels. So, what it means for the model is, (let's assume the numbe of dimensions of the input is N) for each σ_i, where i = 1,2,3,...,N, might be different, depending the importance of the corresponding input features.

ARD is advantageous for the experiment such that: [Cawley and Talbot, 2014]

i. Generalisation performance is potentially improved.
ii. The data would be better explained.
iii. Increase the efficiency of feature extraction for audio.

(d) **Dot Product**
The Dot Product kernel is non-stationary and can be obtained from linear regression by putting N(0, 1) priors on the coefficients of x_d (d = 1, . . . , D) and a prior of N(0, σ_0^2) on the bias. The Dot Product kernel is invariant to a rotation of the coordinates about the origin, but not translations. It is parameterized by a parameter σ_0^2. For $\sigma_0^2 = 0$, the kernel is called the homogeneous linear kernel, otherwise it is inhomogeneous. The kernel is given by

$$K(x, x') = \sigma_0^2 + x \cdot x' \quad (2)$$

While Matern is a very generalized kernel, both isotropic and ARD variant of the Matern kernel is selected for our model. On the other hand, as dot product kernel is a very efficient and simple kernel, we decide to give it a try too.
As we favor simplicity, we actually think that dot product kernel would give us the best result.

Machine Learning Methods Applied

Machine learning is about using models to learn from existing data for some improvement or predictions. The following is some methods used in machine learning:

1. **k-nearest neighbors algorithm**
The k-Nearest Neighbors algorithm is a non-parametric method for classification and regression. The training examples are vectors in a multidimensional feature space, each with a class label. The data will be classified into the class with most number of occurrence within the defined k radius.

2. **Support Vector Machines (SVM)**
Support Vector Machine (SVM), with the help of the libsvm, is a type of linear classifiers that aims at finding a unique solution in the form of an optimal hyperplane which is defined as the one that maximizes the margins between the two classes.

3. **Random Forest Regressor**
Random Forest is an ensemble of decision trees, whereas a decision tree tries to separate the data into different leaf nodes where the data points in each node have certain similarity. Each decision tree will predict a value and the values will be averaged to be the result.

4. **AdaBoost**
AdaBoost is the short form of Adaptive Boost which is sensitive to noisy data and outliers. AdaBoost first find out the classifier with the least errors, adding the weight to the outliers, then find out another classifier with least errors again. A final classifier will be combined and indicate the class of data.

5. **Naive Bayes**
Naive Bayes is a prediction algorithm that uses Bayes rule. All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. The possibilities is represented by the number of times the particular event occurs in the dataset over the total number of events.

Hypothesis: Doing feature selection before applying to the ARD Kernel Gaussian Process model will improve the result. Without doing feature selections, as the feature dimensions would be huge, we would also have large number of kernel parameters. As a result, it is likely to lead to overfitting and generating poor result. [Cawley and Talbot, 2014] Thus, to increase the effectiveness of the Gaussian Process Model, we would apply feature selection to the data first.

Hypothesis: MFCC is the most important feature. Depression is actually tied to emotions : it can either mean sadness or people who deny feeling sad. [Stratou et al., 2015] On the other hand, as the baseline accuracy of using only MFCC to predict emotions is quite high, [El Ayadi, Kamel, and Karray, 2011], we actually guess that MFCC is the most important feature in depression prediction. In other words, since we have limited datas and large feature dimension, using MFCC feature only for the input vector of the model might actually improve the result.

Evaluation

In order to test our proposed Gaussian Process models, we conducted tests on data obtained from Audio/Visual Emotion Challenge and Workshop(AVEC 2016) [Valstar et al., 2016b]. The goal of AEVC is to weigh-in on the various approaches(visual, audio) 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 to test our hypotheses.

Data

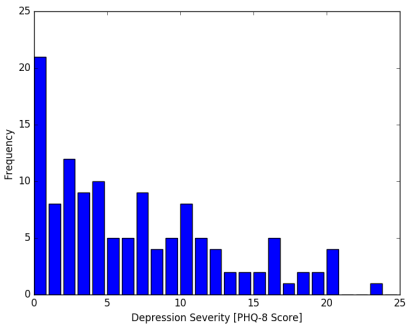

Figure 1: Histogram of the PHQ-8 scores

The depression data used in AVEC 2016 was obtained from the benchmarking database, the 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 1. The data provided are split into 2 sets: training and development. A summary of the data is given in Table 1.

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

Table 1: 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 features (Energy, MFCC, Magnitude Spectrum, Zero-crossing) as presented in section above titled Audio Features. Each audio features is composed of several individual features, the breakdown of the actual number of feature columns are as follows.

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

Table 2: 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 would also use Root Mean Square Deviation Error (RMSE) to measure the error rate on both Training and Development datasets. RMSE(Equation 3) is a commonly used in machine learning communities to measure the differences between the values predicted by a model and the ground truth[Dhanani et al., 2014].

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \tag{3}$$

Feature Selection

Feature selection is the process of selecting a subset of relevant features including variables or predictors to be used in a

model construction for machine learning. The usage of feature selection is to reduce the complexity of a model in machine learning to be interpreted easier. The objective of variable selection 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 leads to overfitting[Smith and Somorjai, 2011]. We would need to perform feature selection on the data before using the machine learning algorithm.

The litany of 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], reliefF[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 Gaussian Models against commonly used machine learning algorithms 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 algorithm 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 2.

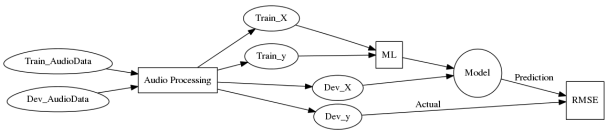

Figure 2: Experimental process

Results

We first run the experiment across the dataset across all 526 features, without feature selection. As we would expect[Cawley and Talbot, 2014], the results are terrible as the number of features is more than the number of data points, potentially causing overfitting. The results of the experiment is illustrated in figure 3. We would expect the GP ARD would be able to theoritcally extract relevant features and improve prediction. However, we have seen experimentally that GP ARD performs poorly, along with other GPs. The results for all the features are shown in figure 3.

We re-run the experiment across the data with feature selection. We run each of the feature subset gathered from the feature selection algorithms against each of the machine learning algorithm. We observe that reliefF, CIFE and CFS selected a large number of MFCC features. This confirms

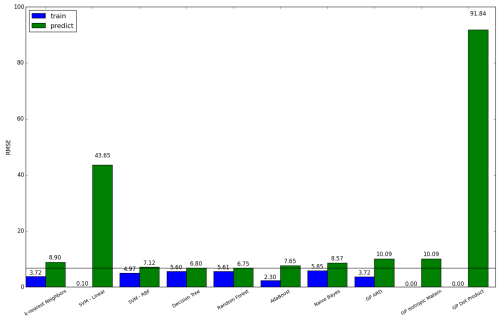

Figure 3: Results across all features

our hypothesis that MFCC gives the best predictive power to predict PHQ-8. Hence, in addition we also ran the experiment on MFCC features, and are represented as MFCC alongside the feature subsets obtained from feature selection. The best results across all feature subsets are shown in figure 4 and in table 3.

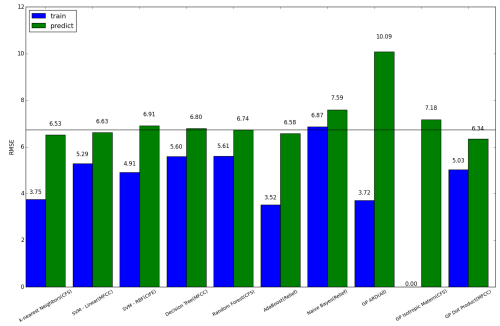

Figure 4: Best Results across all feature subsets

Conclusion

Our work has succesfully modelled the various levels of depression with GP by taking audio recordings as our input.

Contributions

- **Antoine Charles Vincent Garcia**: Scripting the program, setting up machine learning libraries and running tests.
- **Chan Jun Wei**: Project technicalities such as problem formulation and modelling, mathematics and experiment planning.
- **Chen Tze Cheng**: Project technicalities such as problem formulation and modelling, mathematics and experiment planning.
- **Eric Ewe Yow Choong**: Formatting of the report, resolution of LaTeX issues and keeping track of requirements

		RMSE	
Algorithm	Subset	Train	Dev
K-Nearest Neighbors	x	x	
SVM - Linear	x	x	
SVM - RBF	x	x	
Decision Tree	x	x	
Random Forest	x	x	
AdaBoost	x	x	
Naive Bayes	x	x	
GP ARD	x	x	
GP Isotropic Matern	x	x	
GP Dot Product	x	x	

Table 3: RMSE results of the different machine learning algorithms

- **Han Liang Wee, Eric**: Scripting the program, setting up machine learning libraries and running tests.
- **Ho Wei Li**: Background research and writing up of the abstract and introduction section with its subsections.

References

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

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.

El Ayadi, M.; Kamel, M. S.; and Karray, F. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44(3):572–587.

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).

Hashimoto, K.; Zen, H.; Nankaku, Y.; Lee, A.; and Keiichi. 2015. Hyperparameter Estimation for Speech Recognition Based on Variational Bayesian Approach.

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.

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.

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.

Stratou, G.; Scherer, S.; Gratch, J.; and Morency, L. P. 2015. Automatic nonverbal behavior indicators of depression and PTSD: the effect of gender. *Journal on Multimodal User Interfaces* 9(1):17–29.

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.