Predicting Depression Symptoms Using $-omics$ Data

Joy Doong
Stanford University
joydoong@stanford.edu
ID: joydoong

Jessica Jones
Stanford University
jjones6@stanford.edu
ID: jjones6

Michael Wornow
Stanford University
mwornow@stanford.edu
ID: mwornow

Topic: Healthcare

1 Introduction

Depression, ranging from minor to major in severity, affects millions of individuals in the United States. It is the most common mental disorder in the US, with roughly 17% of the population experiencing at least one major depressive episode during their lives [10]. The prevalence of the disorder has steadily increased over the past decade [14] and has dramatically spiked during COVID-19 [6].

Despite depression's prevalence and correlation with elevated risks for diabetes, cancer, cardiovascular disease, hypertension, strokes, general functional impairment, and suicide, the biological basis and metabolic markers behind it are not well understood [10]. Current diagnoses rely on subjective psychiatric evaluation and criteria outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [13]. The DSM is published by the American Psychiatric Association and is used by experts to diagnose individuals for mental disorders [13].

However, the highly subjective and often vague directions provided by the DSM can cause wildly divergent clinical evaluations of the same patient [13]. More recent concerns include a high rate of incorrect diagnoses (i.e. false positives) [13], as well as criticism that pharmaceutical companies with conflicting financial interests have undue influence in shaping the DSM's diagnostic guidelines [5].

We hope to remedy some of these issues by presenting an alternative model to the DSM that is more strongly grounded in objectively measurable patient-level data. Thus, this project aims to develop a deep learning model that can predict whether a patient is experiencing some level of clinical depression, based solely on the metabolomics, lipidomics, cytokine, and/or clinical profile of that individual. To accomplish this, we first conducted exploratory data analysis on our high-dimensional dataset, imputed missing values into the dataset, and stratified patients into training/development/testing datasets. Next, we established several baselines to compare our deep learning model against by training several traditional machine learning models. Finally, we developed a deep learning model in Keras to predict the depression state of a patient based on his/her biomarkers.

2 Dataset

We are utilizing a dataset drawn from a currently unpublished study conducted in Michael Snyder's lab by Nikki Solanki in 2019 which sought to evaluate the effectiveness of Inquiry-Based Stress Reduction (IBSR) on alleviating the symptoms of depression. IBSR is a non-pharmacological 9-day retreat program that treats depression through cognitive belief restructuring.

A total of 63 patients (28 depressed and 35 healthy individuals) were initially enrolled in the study, but only 47 of these patients were selected for the *deep profiling* conducted by the study's authors. This deep-profiling involved the collection of psychological, physiological, and *-omics* data from each individual at five primary time points: before the retreat, during the retreat (during which data was collected daily for psychological surveys and thrice for -omics data), one month after the retreat, three months after the retreat, and six months after the retreat. These deeply profiled individuals formed the basis of our dataset.

At each of these time points, each patient also took the Beck Depression Inventory-II (BDI-II) survey. The BDI-II survey lists symptoms of depression and asks respondents to rank the severity of each symptom on an integer scale from zero to three [11]. The overall score is the sum of the respondent's scores for each symptom, and this overall score corresponds to the DSM criteria for depressive disorders [7].

Not every individual was profiled at every time point, however, and some tests had to be discarded after quality control measures were taken. Thus, there were actually only a total of $N = 201$ total distinct metabolite measurements taken from patients at distinct time points.

We split the patients into a training, development, and test set such that there were 5 patients' measurements in each of the development and test sets, and the remainder of the patients were left in the training set. We separated the data by patient, rather than simply shuffling the entire dataset, in order to prevent the model from "cheating" during training by observing a data point belonging to a patient who was also included in the test set. This division more accurately reflects the way the model would be used in the real world, where we receive a measurement from a patient we've never seen before and need to predict whether they are depressed.

3 Approach

3.1 Classification or Regression?

At first, we tried to frame this project as a regression problem, and trained our model to predict the exact BDI-II score that a patient would have based on their biomarkers. However, the sparsity and noise of our very small dataset made it difficult to successfully train such a model.

Additionally, given the desired use case of our model – predicting whether a given patient is depressed or not based on their *-omics* data – this prediction did not offer much diagnostic benefit nor shed much additional insight on a patient's status given how noisy these scores typically are.

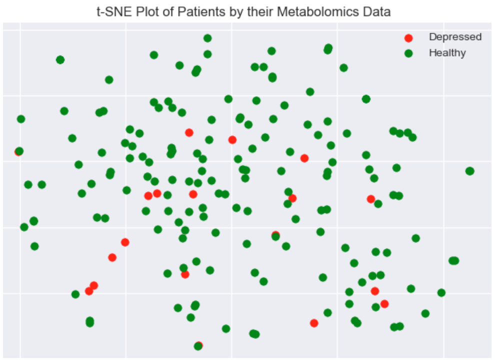
Thus, we decided to re-frame our problem as one of binary classification by binning the BDI-II scores based on expert domain knowledge. The dependent variable Y in our model was a binary random variable such that $Y^{(i)} = 1$ if patient i is depressed, otherwise $Y^{(i)} = 0$. A score of 0-9 on the BDI-II typically indicates that a patient is not depressed, while 10-18 corresponds to mild depression, 19-29 to moderate depression, and 30+ to severe depression [3]. Thus, we considered a patient i to be depressed, and thus have $Y^{(i)} = 1$, if his/her BDI-II score was greater than 9, and $Y^{(i)} = 0$ otherwise.

3.2 Feature Engineering: Imputing Missing Values and Dimensionality Reduction

Our dataset contained a small number of unique datapoints ($N = 205$ in total across all splits) and was of high dimensionality.

Within a single sample of a patient, many biomarkers can be measured, each individual measurement of a specific biomarker will be noisy, and many biomarkers will simply have no reported value due to measurement error. Thus, we faced the challenge of effectively processing this data to minimize noise, and imputing enough of the missing values to ensure that the model could discern some signal for each of the biomarkers.

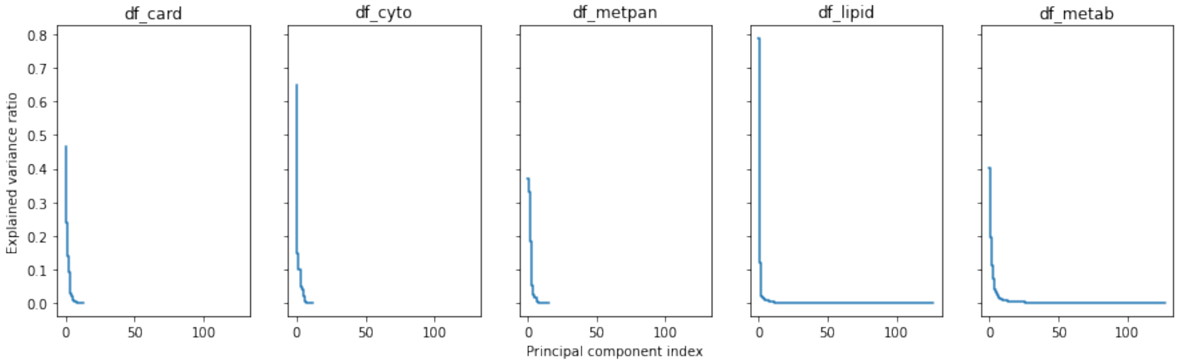
Simple statistical techniques were insufficient to separate this high dimensional data where noise was often larger than signal. The t-SNE plot below projects the vector of metabolomics measurements for each patient onto a 2D plot. As the fairly randomly interspersed depressed patients' measurements (red points) among the non-depressed patients (green points) illustrate, the raw data does not appear to partition itself into obvious clusters, at least when viewed through this projection.



We initially used the metabolomic measurements for a particular patient at a particular timepoint as the feature set X, rather than the other *-omics* data that we had access to. We focused on metabolomics because metabolomics data has been widely praised in the literature as an emerging diagnostic tool [15], and thus we were curious about the predictive power of metabolomics data specifically in practice as it related to mental health. However, we also had access to a much wider range of *-omics* data at our disposal. Thus, we decided to use metabolomics data as our baseline for performance, and tried to come up with a way to take advantage of the other *-omics* data modalities to get a higher performing model.

After imputing missing values (detailed later in this report), we then applied PCA to our imputed dataset to address the high dimensionality of our features. We then tuned our deep learning model with different subsets of the PCA features, varying the amount of total variance that was captured by the features fed into the end model. [9]. We observed that explained variance ratio decreases quickly as the rank of component increases, confirming that PCA is helpful in reducing the number of features for our dataset.

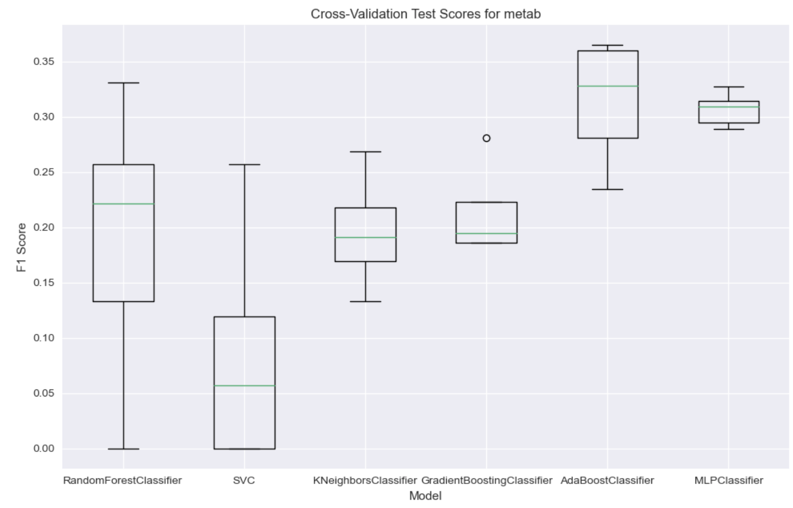
In the chart below, each plot corresponds to a different data modality (e.g. "df_metab" is metabolomics data, "df_metpan" is a separate metabolic panel, "df_lipid" is lipidomics, "df_cyto" is cytokines, and "df_card" are cardiovascular risk factors).



3.3 Baselines

To establish baselines for model performance, we trained several non-deep-learning models, including the following: random forests, gradient boosting, SVCs, and other classical machine learning techniques. We ran grid search over the relevant hyperparameters for each model, and performed 5-fold cross-validation for each set of hyperparameters.

The box plot of the F1 Scores on each of the held out folds of the cross-validation procedure for the best-fitted model as selected by grid search is shown below:



The best F1 Score for each classification model was as follows, with AdaBoost achieving the best F1 Score but SVC and Gradient Boosting tying for best diagnosis accuracy. Given the large bias towards negative values in our dataset, it seemed that F1 Score is a more useful metric for evaluating the utility of a model than accuracy:

Model	Best F1 Score	Best Accuracy
Support Vector	0.26	0.89
K-Nearest Neighbors	0.27	0.75
Gradient Boosting	0.28	0.89
Random Forest	0.33	0.83
Multi-layer Perceptron (4 layers)	0.33	0.81
AdaBoost	0.37	0.67

3.4 Deep Learning Model

Several recent studies have applied straightforward feed-forward deep neural networks to evaluate *-omics* data for disease diagnosis [1; 12]. Since our problem is similar, the code provided by Alakwaa, Chaudhary, and Garmire in their breast cancer metabolomics study served as our starting point [1]. Following their methods, we performed a hyperparameter search for a feedforward neural network with Keras Tuner, evaluating potential models based on binary cross-entropy. We skipped over the quantile normalization that they conducted on their dataset since our Y values were already clearly labeled and interpretable.

4 Results

The breast cancer metabolomics study authors optimized their hyperparameters using random search. They used RMSProp as their learning algorithm and varied the algorithm's learning rate, momentum, and rho values over the random search. They limited their hyperparameter search to relatively small neural network architectures: 4 hidden layers maximum with 10 to 100 units per layers. We ported their code from R and performed a similar search. We tried both random search and Bayesian optimization, as provided in Keras Tuner, to find the best hyperparameters as measured by accuracy on the training set.

We found a relatively small neural network (a single dropout layer and three hidden layers consisting of 40, 20, and 20 units, respectively) could achieve perfect or near-perfect accuracy on the training set. However, accuracy declined sigificantly on the validation set: initially, the model would predict $Y = 0$ for all examples in the validation set. Since our data is imbalanced, with relatively few samples from depressed individuals, the model could achieve high accuracy simply by overfitting to the depressed individuals in the training set.

To address the overfitting problem, we tried several approaches. We used the Synthetic Minority Oversampling Technique (SMOTE) as implemented in the imbalanced-learn library to augment the training data with synthetic examples of depressed individuals [4; 8]. We achieved our best results by computing the five most significant features from each testing panel (metabolomics, lipids, cytokines, cardiovascular risk, metabolic) using PCA. Therefore each example consisted of 25 features. Initially, the training set contained 25 positive examples (samples from depressed individuals) and 103 negative examples. With SMOTE, we created an additional 26 positive examples to improve the class balance. After training on this dataset, the model achieved 86.49% accuracy and 0.44 F1 score on the development set used for tuning. Unfortunately, its performance on the held-out test set was far worse: 69.44% accuracy and 0.0 F1 score. The model did predict $Y = 1$ for some examples in the test set, but none of its predictions were correct.

We also tried to reduce overfitting through regularization. We did a second hyperparameter search for dropout and L2 regularization values, optimizing for accuracy on the validation set. This proved less effective than data augmentation. For example, we started with the model trained on the augmented dataset described above. Holding the basic architecture constant, we searched for the optimal L2 regularization value for each hidden layer and the optimal dropout rate for additional dropout layers. (The search space included zero, so the algorithm could potentially have returned the model unchanged.) The regularized model's accuracy decreased to 78.38% on the development set and 61.11% on the test set.

We anticipated our dataset's high dimensionality and relatively small size might hinder our model's performance. We mitigated these challenges somewhat with data augmentation and dimensionality reduction techniques. To predict depression based on blood test data reliably, a deep neural network will likely require training data from thousands of individuals, a significant percentage of whom are suffering from depression.

5 Conclusion

Detecting and diagnosing mental health disorders remains a complicated task: even the best human experts can only offer educated guesses based on largely qualitative measurements, and thus diagnoses for the same patient and set of symptoms can vary.

In this project we attempted to leverage the vast amount of quantitative data that is being generated by unseen biological processes in patients in order to train a more sophisticated model that could relate patterns in this data to a more accurate diagnosis of depression. The limited size of our dataset, likely exacerbated by its high dimensionality, however, prevented us from developing an effective diagnostic tool.

Additional experiments in the clinic to generate larger datasets of depressed/non-depressed patients and their metabolomics data could improve model performance and generalization. An ensemble approach of multiple models with different biases may also prove useful. For example, Asakura, Date, and Kikuchi developed an ensemble deep neural network (EDNN) to relate phenotypes in fish to metabolomics data [2], and their EDNN produced a lower RMSE than using a single deep neural network for all species in their study.

6 Access to code

https://github.com/Miking98/cs230-project

References

[1] F. M. Alakwaa, K. Chaudhary, and L. X. Garmire. Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. *Journal of Proteome Research*, 17(1): 337–347, 2018. doi: 10.1021/acs.jproteome.7b00595. URL `https://doi.org/10.1021/acs.jproteome.7b00595`. PMID: 29110491.

[2] T. Asakura, Y. Date, and J. Kikuchi. Application of ensemble deep neural network to metabolomics studies. *Analytica Chimica Acta*, 1037:230 – 236, 2018. ISSN 0003-2670.

doi: https://doi.org/10.1016/j.aca.2018.02.045. URL http://www.sciencedirect.com/science/article/pii/S0003267018302605. Analytical Metabolomics.

[3] J. N. Butcher, J. Taylor, and G. Cynthia Fekken. 4.14 - objective personality assessment with adults. In A. S. Bellack and M. Hersen, editors, *Comprehensive Clinical Psychology*, pages 403 – 429. Pergamon, Oxford, 1998. ISBN 978-0-08-042707-2. doi: https://doi.org/10.1016/B0080-4270(73)00018-3. URL http://www.sciencedirect.com/science/article/pii/B0080427073000183.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, Jun 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL http://dx.doi.org/10.1613/jair.953.

[5] L. Cosgrove and S. Krimsky. A comparison of dsm-iv and dsm-5 panel members' financial associations with industry: a pernicious problem persists. *PLoS Med*, 9(3):e1001190, 2012.

[6] C. K. Ettman, S. M. Abdalla, G. H. Cohen, L. Sampson, P. M. Vivier, and S. Galea. Prevalence of Depression Symptoms in US Adults Before and During the COVID-19 Pandemic. *JAMA Network Open*, 3(9):e2019686–e2019686, 09 2020. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2020.19686. URL https://doi.org/10.1001/jamanetworkopen.2020.19686.

[7] G. Jackson-Koku. Beck Depression Inventory. *Occupational Medicine*, 66(2):174–175, 02 2016. ISSN 0962-7480. doi: 10.1093/occmed/kqv087. URL https://doi.org/10.1093/occmed/kqv087.

[8] G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18 (17):1–5, 2017. URL http://jmlr.org/papers/v18/16-365.

[9] U. W. Liebal, A. N. T. Phan, M. Sudhakar, K. Raman, and L. Blank. Machine learning applications for mass spectrometry-based metabolomics. *Metabolites*, 10, 2020.

[10] K. A. McLaughlin. The public health impact of major depression: a call for interdisciplinary prevention efforts. *Prevention Science*, 12(4):361–371, 2011.

[11] K. L. Smarr and A. L. Keefer. Measures of depression and depressive symptoms: Beck depression inventory-ii (bdi-ii), center for epidemiologic studies depression scale (ces-d), geriatric depression scale (gds), hospital anxiety and depression scale (hads), and patient health questionnaire-9 (phq-9). *Arthritis Care & Research*, 63(S11):S454–S466, 2011. doi: 10.1002/acr.20556. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/acr.20556.

[12] D. Stamate, M. Kim, P. Proitsi, S. Westwood, A. Baird, A. Nevado-Holgado, A. Hye, I. Bos, S. J. Vos, R. Vandenberghe, C. E. Teunissen, M. T. Kate, P. Scheltens, S. Gabel, K. Meersmans, O. Blin, J. Richardson, E. De Roeck, S. Engelborghs, K. Sleegers, R. Bordet, L. Ramit, P. Kettunen, M. Tsolaki, F. Verhey, D. Alcolea, A. Lléo, G. Peyratout, M. Tainta, P. Johannsen, Y. Freund-Levi, L. Frölich, V. Dobricic, G. B. Frisoni, J. L. Molinuevo, A. Wallin, J. Popp, P. Martinez-Lage, L. Bertram, K. Blennow, H. Zetterberg, J. Streffer, P. J. Visser, S. Lovestone, and C. Legido-Quigley. A metabolite-based machine learning approach to diagnose alzheimer-type dementia in blood: Results from the european medical information framework for alzheimer disease biomarker discovery cohort. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5:933 – 938, 2019. ISSN 2352-8737. doi: https://doi.org/10.1016/j.trci.2019.11.001. URL http://www.sciencedirect.com/science/article/pii/S2352873719300873.

[13] J. C. Wakefield. Diagnostic issues and controversies in dsm-5: return of the false positives problem. *Annual review of clinical psychology*, 12:105–132, 2016.

[14] A. H. Weinberger, M. Gbedemah, A. M. Martinez, D. Nash, S. Galea, and R. D. Goodwin. Trends in depression prevalence in the usa from 2005 to 2015: widening disparities in vulnerable groups. *Psychological Medicine*, 48(8):1308–1315, 2018. doi: 10.1017/S0033291717002781.

[15] L. Yang, Y. Wang, H. Cai, S. Wang, Y. Shen, and C. Ke. Application of metabolomics in the diagnosis of breast cancer: a systematic review. *Journal of Cancer*, 11(9):2540, 2020.