An Interpretable Depression Machine Learning Classifier on Brain Imaging Data

Tayden Li
Stanford University
Stanford, CA 94305
taydenli@stanford.edu

Abstract

There are growing concerns about the generalizability and interpretability of machine learning classifiers on brain imaging data. Prior studies attempted to classify psychiatric disorders, yet the datasets in these studies have small sample sizes, with most of them having fewer than 100 individuals. In this study, I use machine learning approaches (logistic regression, decision trees, and neural networks) to predict depression using brain imaging variables on a dataset extracted from the UK Biobank with 5,002 individuals. This is currently larger than any brain imaging datasets of any psychiatric disorders in the literature. The best classification result is obtained with LightGBM, with an area under precision-recall curve (AUPRC) of 0.385.

1 Introduction and related work

Major depressive disorder (MDD), commonly known as depression, is a psychiatric disorder with a 20.6% life time prevalence (Hasin et al., 2018). On an individual level, it can significantly decrease a person's quality of life. On a societal level, it costs tens of billions of dollars each year in the U.S. alone (P. S. Wang et al., 2003). Despite its huge individual and societal impact, the biological mechanisms underlying MDD are still largely unknown.

Our poor understanding of MDD leaves clinicians with no choice but to rely on subjective diagnostic measures. Psychiatrists base their diagnosis on the Diagnostic and Statistical Manual of Mental Disorders-5 (DSM-5), yet the diagnostic criteria listed in DSM-5 are often criticized for being arbitrary in nature (Association et al., 2013; Chmielewski et al., 2015). A more accurate, objective diagnosis is needed for psychiatrists to prescribe appropriate medications.

A machine learning approach to predict MDD with biological data may improve our understanding of the biological basis of depression and make diagnosis more objective. Previous medical research has shown that some neuroimaging abnormalities may be related to depression (Zhuo et al., 2019). It is thus promising to use magnetic resonance imaging (MRI) data to predict whether a person has MDD.

Plenty of studies have tried to predict different psychiatric disorders using processed MRI data, where biologically meaningful variables are extracted from raw MRI images. In the review papers by Du et al. (2018), Gao et al. (2018), Patel et al. (2016), and Wolfers et al. (2015), support vector machine (SVM) is the most widely used method because it can be fairly resistant to overfitting when we input high dimensional data with brain attributes. Prior studies also often use methods like logistic regression, random forest, and neural networks (Chauhan & Choi, 2020; Hu et al., 2020; Plitt et al., 2015). Mutliple studies in the aforementioned review papers report to have achieved some accuracy higher than 90%.

However, we should cautiously interpret these results as they might not generalize well to the general clinical population. Since recruiting participants can be challenging, prior studies are conducted on small datasets, with the majority of them having fewer than 100 individuals (Andrews et al., 2018; Chauhan & Choi, 2020; Gao et al., 2018; Price et al., 2014; X. Wang et al., 2017; Wolfers et al., 2015). With small sample sizes, these studies report cross-validation or leave-one-out cross-validation accuracy in place of hold-out test set accuracy. Lanka et al. (2020) found that the nature of the small sample sizes and the use of cross-validation accuracy may have significantly inflated the results of prior studies. The performance of the models in prior studies might be close to random on a hold-out test set.

In this study, I curate a brain imaging dataset with 5,002 individuals and 2,331 processed MRI variables. This is larger than any exisiting neuroimaging dataset of some psychiatric disorders in the literature. I use these brain imaging attributes as input to predict whether an individual has depression with logistic regression, two decision tree methods (random forest and LightGBM), and two neural networks. As the primary goal of this project is to improve our understanding of the biological basis of depression, I use SHapley Additive exPlanations (SHAP), an estimation of shapley values to visualize and help us interpret how different brain imaging features contribute to the best model's prediction.

Preprint. Under review.

2 Dataset

2.1 Dataset Cleanup

The dataset used in this study is extracted from UK Biobank. UK Biobank is a database with extensive biological samples and medical information from over 500,000 participants aged 40–69 years (Sudlow et al., 2015).

In this huge database, I first handpick brain imaging variables out and combine files in various formats. I average over duplicated variables if it makes biological sense to do so or when it is appropriate to do so given how the data is collected. Otherwise, individuals that have other duplicated or missing variables are removed. I also change the coding of some variables, such as depression status, as they were originally coded to fit other unrelated variables. The data is then randomly split into train:validation:test $= 70\% : 15\% : 15\%$ with same ratios of cases and controls in each subset.

2.2 Dataset characteristics

The final curated dataset contains 5,002 individuals, of which 1,421 have depression and 3,581 does not. Each individual has 2,331 features. There are four types of features—participant backgrounds, processed structural MRI, processed diffusion MRI, and processed resting functional MRI attributes.

Participant background features include age, sex, BMI, and a metric of socioeconomic status. Structural MRI is an imaging technique that examines the anatomy of the brain. Example variables include volume of grey matter in amygdala. Diffusion MRI captures the mapping of the diffusion process of molecules, which can then be used to infer the structure of nerve tracts. Lastly, functional connectivity values derived from resting functional MRI accounts for the organization and relationship among spatially separated brain regions.

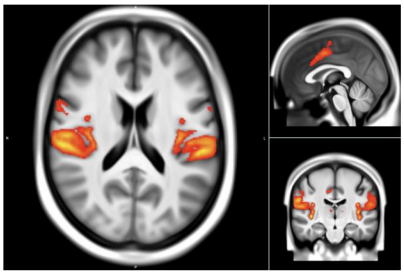


Figure 1: An example fMRI brain map from UK Biobank.

2.3 Dimensionality Reduction

As the dataset used here has a large feature space, I used dimensionality reduction techniques to visualize the dataset and create two lower-dimensional datasets.

2.3.1 t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, nonlinear dimensionality reduction technique. t-SNE algorithm works by first constructing a probability distribution in a high dimensional space such that similar points are assigned higher probability and vice versa for dissimilar points. This distribution is built according to a Gaussian distribution that can be manipulated through the variable "perplexity" in the implementation. It similarly constructs a distribution in a lower dimensional space using Student t-distribution. The algorithm then uses gradient descent to minimize Kullback-Leibler divergence, which measures the similarity of the two probability distributions.

Figure 2a shows the result of t-SNE. Even with a variety of perplexity and numbers of iterations, t-SNE cannot cluster datapoints into different depression outcomes, suggesting that this could be a challenging classification task.

2.3.2 PCA

Principal component analysis (PCA) is another dimensionality reduction method. In PCA, principal components are constructed such that the first component accounts for the largest possible variance. The subsequent principal components will be the ones that are orthogonal to previous ones and account for the largest possible remaining variance. PCA works by first standardizing all variables to a comparable scale. We can then calculate the covariance matrix and its eigenvectors and eigenvalues. After sorting the eigenvectors by their eigenvalues in descending order, the principal components by variance expalained in descending order will point to the direction that the associated eigenvectors do.

We can see in Figure 2b that, under the first two components, datapoints of different labels are clustered together. Notice that there are outliers in the top left and lower right corners even after data standardization.

To balance the tradeoff between the number of components and retained information in our two lower-dimensional dataset, I utilized Figure 2c to create a dataset consisting of the first 500 principal components, which explains 71.8% of the variance of original data, and a dataset with the first 1,000 components, which explains 90.3% of the original variance.

3 Methods and experiments

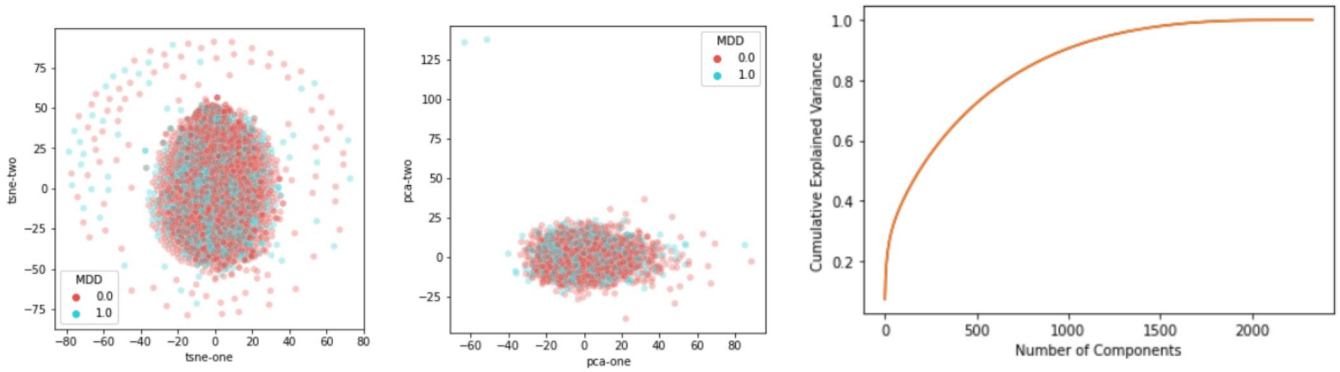
Due to the space constraint, I will focus on presenting the experiments conducted on the original dataset without dimensionality reduction. However, similar hyperparameter tuning experiments are conducted on the two lower-dimensional datasets as well.

3.1 Metrics

Due to the class imbalance in our data (the number of samples for each label is different), accuracy might be misinterpreted. Precision and recall may alter with the threshold that determines if a class is positive or not. Therefore, I use area under precision-recall curve (AUPRC) as my main metric, along with area under receiver operating characteristic curve (AUROC). Recall and precision can still be inferred from the precision-recall curve presented in the results section.

A receiver operating characteristic curve is a plot of true positive rate (or recall) versus false positive rate. The quantities used in AUROC and AUPRC are defined as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP}$$



(a) t-SNE.

(b) PCA: first two principal components.

(c) The relation between cumulative explained variance of the data and number of principal components.

Figure 2: Dimensionality reduction on our dataset.

$$\text{False positive rate} = \frac{FP}{FP + TN}$$

where TP is true positive, FP is false positive, TN is true negative, and FN is false negative.

Note that the baseline for AUROC is 0.5, and the baseline for AUPRC is 0.284. This is because an average random classifier will have a horizontal precision-recall curve with precision being the fraction of positive examples in the data.

3.2 Logistic regression

Method I choose logistic regression as my baseline model because it is one of the simplest model for a classification task. It aims to minimize the following loss function $J(\theta) = \lambda||\theta||_p^p$

$$-\sum_{i=1}^{n}\left(w_1 y^{(i)} \log(h_\theta(x^{(i)})) + w_0(1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))\right)$$

where $x^{(i)}$ is the variable vector of the i-th individual, n is the number of examples, θ is the coefficient vector, $||\theta||_p$ is the L^p norm of theta, λ is the regularization strength, $h_\theta(x^{(i)}) = \frac{1}{1+e^{-\theta^T x^{(i)}}}$, and $y^{(i)} = 1$ if the i-th individual has depression and 0 otherwise. When class weights are used to account for class imbalance,

$$w_k = \frac{\text{total number of individuals}}{2 \times \text{ number of individuals with } y^{(i)} = k}.$$

Otherwise, $w_0 = w_1 = 1$.

Experiments The effects of regularization and class weights on AUPRC are shown in Table 1. Note that the AUPRC values under different experimental conditions often differ by less than 0.001.

Table 1: Validation set AUPRC under different logistic regression settings on the original dataset. For each regularization setting, the highest AUPRC of differnt regularization strength is reported.

Regularization	Class weight	
	Not used	Used
None	0.326	0.327
L1	0.327	0.327
L2	0.326	0.327

3.3 Decision Tree-based Methods

3.3.1 Random forest

Method Random forest ensembles multiple decision trees trained in parallel with bagging (Breiman, 2001). Bagging allows individual trees to be trained on subsets of training data that are randomly sampled with replacement. For a tree, when deciding which feature to split on, we pick the one that decreases Gini impurity the most. Gini impurity is given by

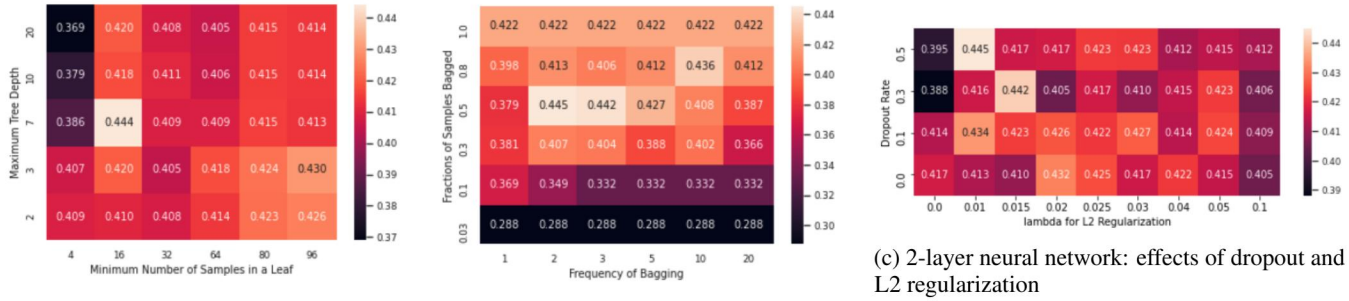
$$I_G(p) = \sum_{i=1}^{K} p_i(1 - p_i)$$

where p_i is the probability of an item with label i being randomly chosen from a set and K is the number of labels. Bagging, along with that each tree is trained on a random subset of features, ensures that each tree is unique. The aggregation of unique trees reduces the variance of the algorithm, making random forest more robust to noise in the training data than a decision tree is.

Experiments To prevent overfitting, I experiment with different tree attributes, including maximum tree depth, maximum number of leaves, and minimum number of samples in a leaf. Maximum proportion of samples and features used are also tested. Only part of the experiments are shown in Figure 3a due to the limit of space. Our best performing model on the original dataset has 100 trees in the forest, a maximum tree depth of 7, at least 16 samples in each leaf, and at most 150 leaves in each tree. At most 50% of features and 70% of instances are used to train each tree. Class weights are used to account for class imbalance.

3.3.2 LightGBM

Method Light Gradient Boosted Machine, or LightGBM, is a fast-to-train, state-of-the-art algorithm in numerous classification tasks on tabular data (Ke et al., 2017). LightGBM is a gradient boosting decision tree (GBDT) algorithm, which combines weak learners, or decision trees that perform at least slightly better than random, into a strong model iteratively. Mathematically, our goal is to find a strong model $\hat{F}$ that minimizes some loss function $L(y, \hat{F}(x))$. Let F_m be



(a) Random forest: effects of tree structure

(b) LightGBM: effects of bagging

(c) 2-layer neural network: effects of dropout and L2 regularization

Figure 3: Parameter tuning for decision trees and neural networks. Validation set AUPRC on the original dataset is presented.

the combined model we have at the m-th iteration. We first train a weak learner $h_{m+1}(x)$ that is closest to the gradient

$$\sum_{i=1}^{n} \nabla_{F_m} L(y^{(i)}, F_m(x^{(i)})).$$

We then get F for our next step

$$F_{m+1}(x) = F_m(x) + \gamma_{m+1} h_{m+1}(x) \quad \text{where}$$

$$\gamma_{m+1} = \text{argmin}_\gamma \sum_{i=1}^{n} L(y^{(i)}, F_m(x^{(i)}) + \gamma h_{m+1}(x)).$$

In addition to GBDT, LightGBM incorporates special techniques to speed up the training. Two key techniques are gradient-based one-sided sampling (GOSS) and exclusive feature bundling (EFB). Instead of using the full training set to train each tree, GOSS focuses on using training instances with large gradients. EFB, on the other hand, reduces the number of features by merging mutually exclusive features.

Experiments Similar to random forest, I experiment with tree structures, proportion of features used in each tree, and bagging parameters (shown in Figure 3b) to prevent overfitting. On the original dataset, the best model has 200 boosting rounds and a 0.03 learning rate. For every 2 iterations, 50% of samples are bagged. Each weak learner is trained with 50% of features. Each tree has a maximum depth of 3, at least 128 samples in each leaf, and at most 30 leaves. Class weights are used to address class imbalance.

3.4 Neural networks

A 2-layer and a 4-layer neural network are implemented here. Both networks are trained with a batch size of 32 and a learning rate of 0.001 with Adam optimizer. Weighted binary cross entropy loss is used to account for class imbalance. Data is first normalized prior to feeding into the networks.

In the 2-layer network, there is a dense layer of varying units with ReLU, followed by a prediction layer with sigmoid activation. The effects of number of input layer units are captured in Table 2. The 4-layer network begins with a 128-unit dense layer with ReLU, followed by two 256-unit dense layers with ReLU and a prediction layer with sigmoid activation.

I experiment two methods of regularization—L2 regularization on weights and dropout layers after each dense layer (Figure 3c). On the original dataset, the best 2-layer network uses a 0.01 L2 regularization strength and a 0.5 dropout rate, whereas the best 4-layer one has a 0.025 L2 regularization coefficient and a 0.1 dropout rate.

Table 2: 2-layer network: validation set AUPRC with different input layer units on the original dataset.

Input layer units	128	256	512	1024
AUPRC	0.455	0.436	0.433	0.430

3.5 SHAP

For the best model, I use model-agnostic SHapley Additive exPlanations (SHAP) to explain how features contribute to a model's prediction (Lundberg & Lee, 2017). SHAP computes Shapley values from coalitional game theory by fairly distributing prediction probability among all features. An explanation model g satisfies the following equation

$$g(z) = \phi_0 + \sum_{j=1}^{M} \phi_j z_j.$$

where M is the number of features, $\phi_j \in \mathbb{R}$ is the Shapley value for feature j and $z \in \{0,1\}^M$ is the coalition vector that has 1 as an entry iff a feature is present. Shapley value represents the expected marginal contribution of each feature after all possible combinations of them have been considered.

4 Results and Discussion

The results of various models are presented in Figure 4a-c.

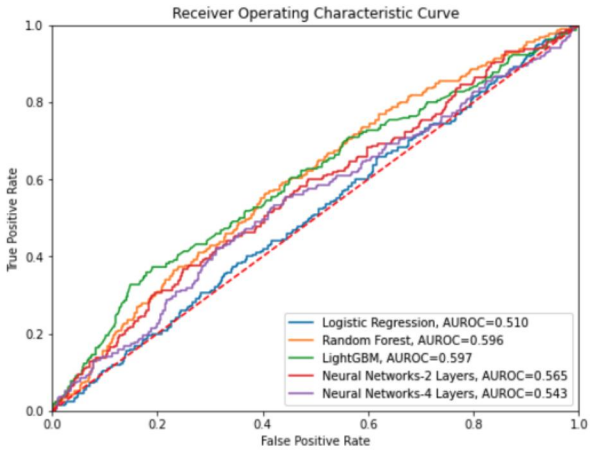
Logistic Regression performs poorly on the original dataset, which is expected as it is a relatively simple model. However, on the other two dataset, it improved significantly and achieved an AUPRC of 0.368, outperforming decision trees.

Decision Trees perform well on the original dataset. LightGBM has the highest performance (AUPRC $= 0.385$ and AUROC $= 0.597$). However, decision trees suffer a drop of performance when trained on the other two datasets, where LightGBM specifically is a lot more prone to overfitting.

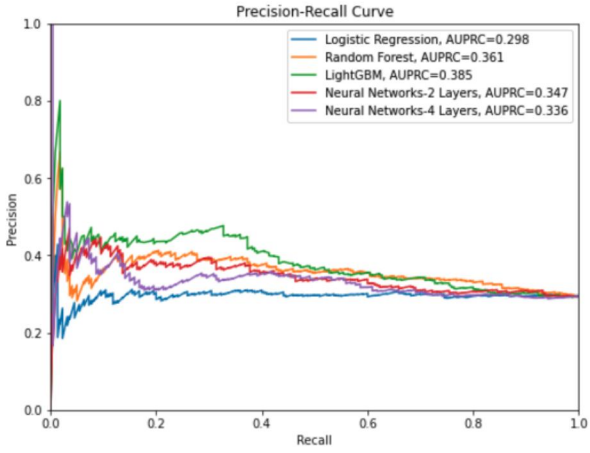
Neural Networks, specifically the two-layer one, have the highest performance on the datasets with 500 and 1000 principal components. However, these models perform worse than decision trees on the original dataset. The 4-layer neural network is, in general, more prone to overfitting and performs worse than the two-layer one on test sets.

Dataset	LR		RF		LGB		NN-2		NN-4	
	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC
Original-train	.606	.365	.985	.969	.656	.425	.884	.783	.926	.853
Original-test	.510	.297	.596	.361	.597	.385	.565	.347	.543	.336
PCA500-train	.784	.559	.828	.667	.968	.932	.900	.776	.992	.979
PCA500-test	.592	.368	.545	.332	.506	.325	.593	.385	.566	.353
PCA1000-train	.874	.748	.871	.725	.808	.623	1.	1.	.996	.992
PCA1000-test	.553	.346	.567	.349	.527	.322	.560	.361	.556	.352

(a) Results of various algorithms on the original dataset and two others with reduced dimensions. ROC and PRC denotes AUROC and AUPRC respectively; LR denotes logistic regression, RF denotes random forest, LGB denotes LightGBM, NN-2 denotes 2-layer neural network, and NN-4 denotes 4-layer neural network. Note that 0.5 is the baseline for AUROC and 0.284, the fraction of positive instances, is the baseline for AUPRC.



(b) ROC curve on the original dataset.



(c) Precision-recall curve on the original dataset.

(d) The impact of top features on the output of LightGBM. Features are ranked in descending order of their influence on output. SHAP value is proportional to how much a feature can change the prediction probability. FNC variables represent functional connectivity value of specific brain regions.

Figure 4: Summary of results on depression prediction.

Figure 4d shows that although there are quite a few brain imaging variables that are important in LightGBM's prediction, sex and BMI are the most influential features. This is not expected yet consistent with that depression is more prevalent in women (blue in the figure) and individuals with higher BMI (red in the figure) (Speed et al., 2019).

The performance of our models are lower than expected. Some hypotheses behind the low performance are as follows:

- Even though this dataset contains more individuals than the ones in the literature, the dataset size is still small. This can make algorithms here underpowered.
- There might be other algorithms, such as other neural networks, that can perform better than the ones we used here.
- It is also likely that the features in this dataset alone are not sufficient to model depression. Depression involves numerous biological, sociological, and environmental factors, which is the reason why scientists understand them poorly. By having almost all features here being brain imaging data, we might have lost some critical information. Moreover, the depression variable we use here, as mentioned in the introduction section, are based on subjective criteria that scientists still debate about. The lack of consensus in diagnosis might have caused our depression label to be not entirely accurate (Du et al., 2018).

5 Conclusion

On our brain imaging dataset of 5,002 individuals and 2,331 features, we achieve a somewhat discouraging but better-than-random result. The implication of this study is discussed, with a goal of improving future work of depression diagnosis. A larger dataset with more individuals and more biological and environmental variables included may critically benefit this line of work a lot. Even with limited data, I would like to conduct the following experiments in the future:

- Exploring alternative feature selection methods (other than PCA), such as forward/backward feature selection.
- Implementing other algorithms, such as support vector machine, as well as ensembling multiple machine learning methods to combine the advantages of different algorithms.

Contributions

Tayden was responsible for retrieving and cleaning the dataset, proposing and running all experiments, as well as the project write-up.

References

Andrews, D. S., Marquand, A., Ecker, C., & McAlonan, G. (2018). Using pattern classification to identify brain imaging markers in autism spectrum disorder. In *Biomarkers in psychiatry* (pp. 413–436). Springer.

Association, A. P., et al. (2013). *Diagnostic and statistical manual of mental disorders (dsm-5®)*. American Psychiatric Pub.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Chauhan, N., & Choi, B.-J. (2020). Dnn based classification of adhd fmri data using functional connectivity coefficient. *International Journal of Fuzzy Logic and Intelligent Systems*, *20*(4), 255–260.

Chmielewski, M., Clark, L. A., Bagby, R. M., & Watson, D. (2015). Method matters: Understanding diagnostic reliability in dsm-iv and dsm-5. *Journal of abnormal psychology*, *124*(3), 764.

Code libraries used: scikit-learn, tensorflow, keras, shap, pandas, numpy, matplotlib, seaborn, lightgbm. (n.d.).

Du, Y., Fu, Z., & Calhoun, V. D. (2018). Classification and prediction of brain disorders using functional connectivity: promising but challenging. *Frontiers in neuroscience*, *12*, 525.

Gao, S., Calhoun, V. D., & Sui, J. (2018). Machine learning in major depression: From classification to treatment outcome prediction. *CNS neuroscience & therapeutics*, *24*(11), 1037–1052.

Hasin, D. S., Sarvet, A. L., Meyers, J. L., Saha, T. D., Ruan, W. J., Stohl, M., & Grant, B. F. (2018). Epidemiology of adult dsm-5 major depressive disorder and its specifiers in the united states. *JAMA psychiatry*, *75*(4), 336–346.

Hu, J., Cao, L., Li, T., Liao, B., Dong, S., & Li, P. (2020). Interpretable learning approaches in resting-state functional connectivity analysis: The case of autism spectrum disorder. *Computational and Mathematical Methods in Medicine*, *2020*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*, 3146–3154.

Lanka, P., Rangaprakash, D., Dretsch, M. N., Katz, J. S., Denney, T. S., & Deshpande, G. (2020). Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets. *Brain imaging and behavior*, *14*(6), 2378–2416.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Patel, M. J., Khalaf, A., & Aizenstein, H. J. (2016). Studying depression using imaging and machine learning methods. *NeuroImage: Clinical*, *10*, 115–123.

Plitt, M., Barnes, K. A., & Martin, A. (2015). Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. *NeuroImage: Clinical*, *7*, 359–366.

Price, T., Wee, C.-Y., Gao, W., & Shen, D. (2014). Multiple-network classification of childhood autism using functional connectivity dynamics. In *International conference on medical image computing and computer-assisted intervention* (pp. 177–184).

Speed, M. S., Jefsen, O. H., Børglum, A. D., Speed, D., & Østergaard, S. D. (2019). Investigating the association between body fat and depression via mendelian randomization. *Translational psychiatry*, *9*(1), 1–9.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., … others (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos med*, *12*(3), e1001779.

Wang, P. S., Simon, G., & Kessler, R. C. (2003). The economic burden of depression and the cost-effectiveness of treatment. *International journal of methods in psychiatric research*, *12*(1), 22–33.

Wang, X., Ren, Y., & Zhang, W. (2017). Depression disorder classification of fmri data using sparse low-rank functional brain network and graph-based features. *Computational and mathematical methods in medicine*, *2017*.

Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., & Marquand, A. F. (2015). From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience & Biobehavioral Reviews*, *57*, 328–349.

Zhuo, C., Li, G., Lin, X., Jiang, D., Xu, Y., Tian, H., … Song, X. (2019). The rise and fall of mri studies in major depressive disorder. *Translational Psychiatry*, *9*, 335. doi: https://doi.org/10.1038/s41398-019-0680-6