Using Machine Learning and DNA Methylation to Diagnose PTSD

By:

William Guo

March 28, 2020

Honors Integrated Science II

Jennifer Nguyen - Period 2

Abstract

The objective was to determine if DNA methylation which affects gene expression can be used as a biomarker to diagnose post-traumatic stress disorder (PTSD), which is a recurring psychological response to traumatic events involving life threatening situations, serious injury, or death. Soldiers, doctors, and emergency workers are among those with a high risk of developing PTSD, which, in recent studies, has been suggested to have a causal association with DNA methylation, a chemical process which adds a methyl group (CH3) to cytosines in a DNA sequence. Numerous machine learning algorithms were used to build multiple models through taking significant methylated sites out of a list of 429,948 from 163 subjects containing both patient and control groups, in order to select the model with the highest accuracy of predicting PTSD. A final list of 143 sites, identified through using probability (p) values as a baseline to determine if the site itself was actually significant or not, was used as the dataset for multiple successful machine learning models. The model with the highest accuracy utilized the algorithm SVC (Support Vector Classifier), with a testing accuracy of 92.3%. This means that we can expect this model to correctly predict PTSD 92.3% of the time. In all, as supported by the findings and research, it can be reasonably concluded that the hypothesis of the feasibility of using DNA Methylation as a biomarker to diagnose PTSD, was indeed supported

Table of Contents

Introduction

When asked to consider mental conditions prevalent in modern society, many people's initial thoughts may include popular examples like clinical depression or anxiety disorder. However, one mental condition that is often overlooked is post-traumatic stress disorder, better known as PTSD. Triggered by some traumatic event, this disorder is characterized by three main types of symptoms, 1) re-experiencing the trauma through intrusive distressing recollections of the event, flashbacks, and nightmares, 2) emotional numbness and avoidance of places, people, and activities that are reminders of the trauma, and 3) increased restlessness such as difficulty sleeping and concentrating, feeling jumpy, and being easily irritated and angered (Anxiety and Depression Association of America). Approximately 3.5 percent of U.S adults are affected with PTSD, with an estimated one in every 11 people who will be diagnosed with this disorder in their lifetime (Parekh, Torres 2017). Additionally, in a study of nearly 9,000 military members by King's College London, it found that PTSD in the military increased from 4 percent in 2004-5 to 6 percent in 2014-16 (Ives, 2018). Despite the rise and clear prevalence of PTSD cases in the U.S, there is still need for a scientific way to identify if someone has PTSD. The current go-to method of diagnosing this disorder is through the comparison of possible symptoms. However, recent studies including one conducted by researchers from the University of Michigan may just hold the key to the potential creation of a prediction-based model, which is the focus of this project. By using data from the Detroit Neighborhood Health Study, a five-year project funded by the National Institute of Health, the U-M researchers identified several genes that appeared more active in people with PTSD. The findings went one step further by identifying a specific biochemical reaction that may indicate a causal association with PTSD, DNA methylation

(Arbor, 2010). DNA methylation is an important epigenetic process by which a methyl group (CH3) is added to DNA, typically affecting gene expression. As such, the reason why certain genes may be more associated with PTSD patients is likely due to CpG sites in the promoter region being silenced due to DNA methylation, thus repressing the expression and preventing that specific part from being expressed. Therefore, this experiment aims to test the hypothesis of whether DNA methylation can be used as a biomarker, a measurable substance that is indicative of some phenomenon, to diagnose PTSD. This will be accomplished through the use of a machine learning model to distinguish PTSD patients and controls based on their respective DNA methylation levels. Machine Learning is defined as a research field at the intersection of statistics, artificial intelligence, and computer science and is also known as predictive analytics or statistical learning. (Guido, Müller 2016). The variables in this experiment are the sites, which refer to the areas near or in the promoter region of a gene, with data from 163 subjects (81 PTSD, 82 controls) including their DNA methylation levels. A list of 143 very significant sites out of 429,948 in conjugation with different machine learning algorithms will be used to build multiple models in hopes of creating one with high accuracy in predicting PTSD. The applications of this model, developing a reliable and scientific method to assist in the diagnosis of this disorder, are significant. For instance, those with a high risk to develop PTSD, such as law enforcement, military personnel, and emergency workers can take an annual test using this model to check if they have PTSD. This will allow those who do to receive appropriate treatment sooner. As a result, this will help reduce instances of possible suffering, self-harm, or even suicide due to this disorder being left undiagnosed. This model can also serve as a method of monitoring a patient's recovery process and evaluating the effectiveness of their treatment.

Ultimately, this model would give PTSD patients the opportunity to receive an earlier diagnosis and streamline the treatment process to help them live a happy and normal life.

The hypothesis in this experiment is that DNA Methylation, which affects gene expression, can be used as a biomarker to diagnose post-traumatic stress disorder. Once research is complete, the question of whether DNA methylation can truly be used to diagnose PTSD will be addressed, as well as if an accurate prediction-based machine learning model can be built.

To reiterate, the variables in this experiment are sites near or in the promoter region of a gene, with data from 163 subjects (81 PTSD, 82 controls) regarding DNA methylation levels. A list of 143 very significant sites out of 429,948 in conjugation with different machine learning algorithms will be used to build multiple models in hopes of creating one with high accuracy of predicting PTSD.

Materials and Methods

Materials:

Databases such as PubMed and GEO (Gene Expression Omnibus) host a large amount of biological publications and datasets freely available to the public (https://www.ncbi.nlm.nih.gov/pubmed/). The keywords of post-traumatic stress disorder (PTSD) and DNA methylation were used to search the GEO DataSets for data containing methylation values, PTSD patients and controls. Data science analysis tools and programming languages and platforms used include Scikit-Learn, R, Python, Jupyter Notebooks, and Linux. The Database for Annotation, Visualization and Integrated Discovery (DAVID) was utilized to investigate biological functions and pathways involved by PTSD-associated genes.

Physical computer hardware:

1 Personal Computer: 1.8 GHz processor, 8 GB RAM, Windows 10

Digital dataset and computer software:

GSE89218 Dataset from Pubmed	DAVID Informatics	Jupyter Notebook
RStudio (Computer Program)	Linux (Operating System)	Microsoft Excel

Methods:

DNA methylation levels in the blood samples of PTSD patients and controls measured using methylation array technology are downloaded from the GEO database. Preprocessing and quality control steps are applied such as checking and removing outlier samples. Samples are randomly split into discovery and validation cohorts. A list of methylation sites which have differential methylation levels between patients and controls is first identified using probability (p) values of the Student t test in the discovery cohort. Data from these sites in the validation group are then extracted and analyzed again by t test. A sublist of significant sites is selected ($p < 0.05$). In order to generate a machine learning model for PTSD diagnosis, the validation cohort is split into training and test sets. Five popular machine learning algorithms including KNeighbors, Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine are trained and tested. A PTSD diagnostic model with the best accuracy of predicting PTSD is selected in hopes of assisting physicians in medical decision making.

1. Download and install all the necessary tools, including …

- Anaconda 3 (for Jupyter Notebook): Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing.

Figure 1. Anaconda 3

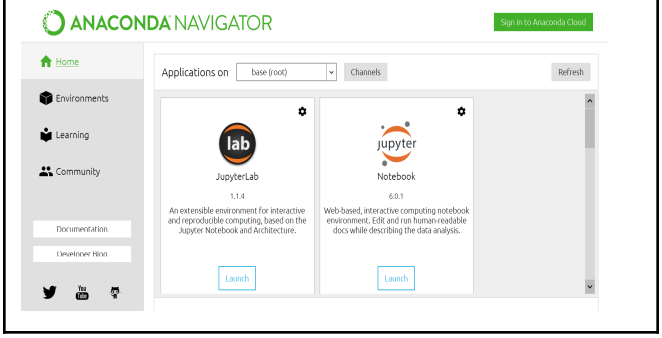
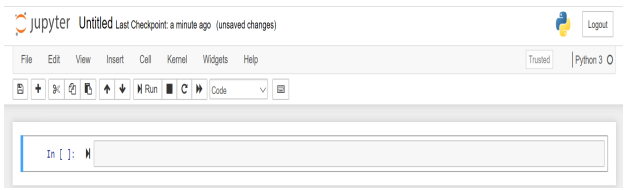
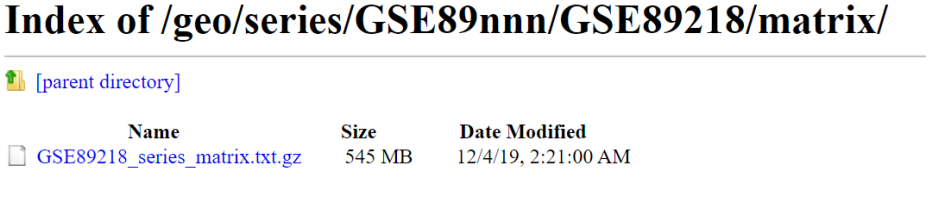


Figure 2. Python 3 in Jupyter Notebook



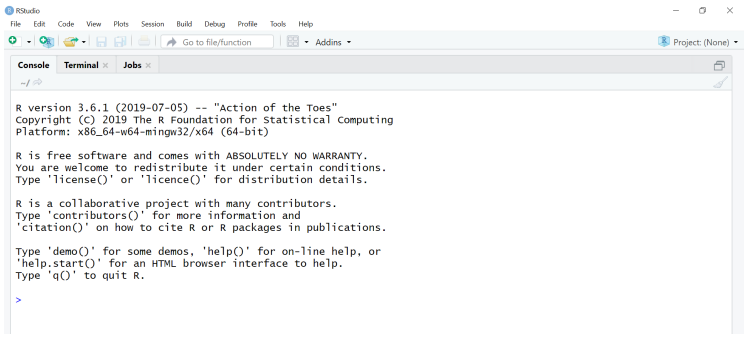
- Download GSE89218 Dataset from Pubmed
    - Go to Pubmed: PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. Its URL is https://www.ncbi.nlm.nih.gov/pubmed/.
    - Select GEO Datasets and type in "PTSD [space] DNA Methylation"
    - Click the first study "Characterization of Whole Genome DNA Methylation Profile Associated with Post-Traumatic Stress Disorder in OIF/OEF Veterans [bisulfite-converted DNA]"
    - Scroll down and click "Series Matrix File(s)"
    - Once on the new tab, right click "GSE89218_series_matrix.txt.gz" and click "Copy link address". The link address is used to download this dataset into Linux for preprocessing.

Figure 3. "GSE89218_series_matrix.txt.gz" file



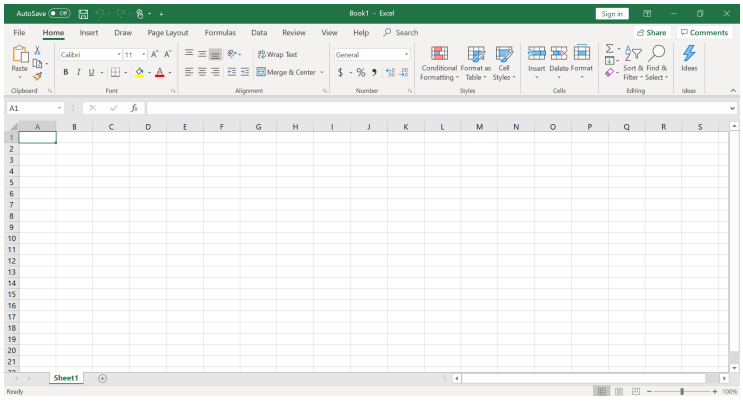
- R Studio: RStudio is an integrated development environment for R, a programming language for statistical computing and graphics.

Figure 4. RStudio



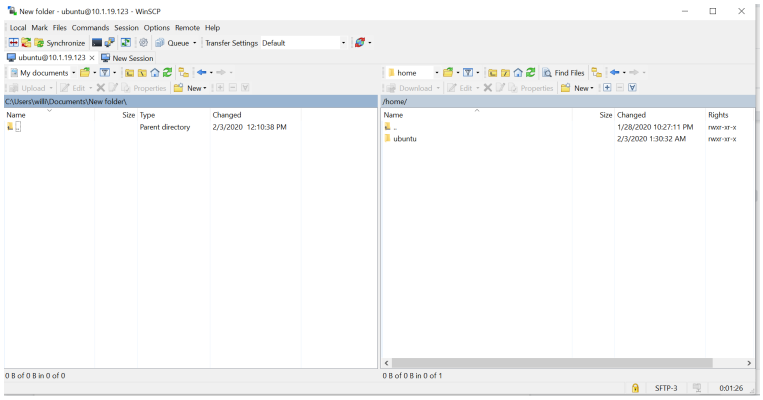
- Excel: Microsoft Excel is a spreadsheet developed by Microsoft.

Figure 5. Excel Spreadsheet



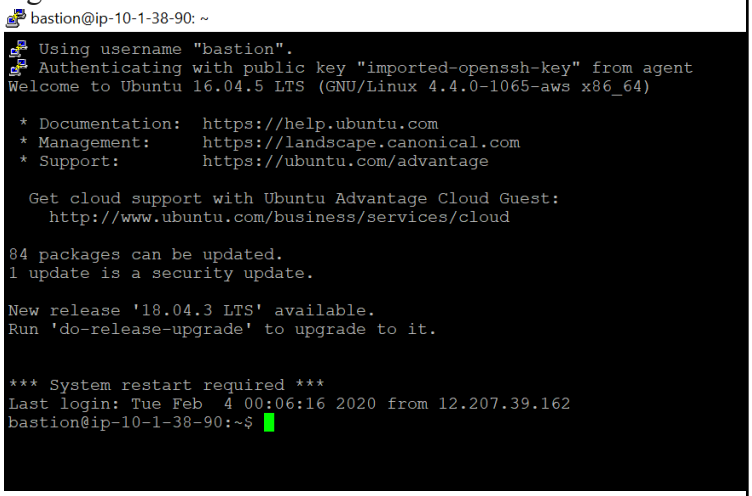
- WinSCP: WinSCP is a free and open-source secure copy program for Microsoft Windows. Its main function is secure file transfer between a local and a remote computer.

Figure 6. WinSCP



- Linux: Linux is a family of open source Unix-like operating systems.

Figure 7. Linux



2. Preprocess the GSE89218 Dataset from Pubmed:

On Linux prompt, type wget, paste the earlier link address and press Enter. After running the command, decompress the "GSE89218_series_matrix.txt.gz" file using 'gunzip', and then use the 'vi' command to view the file and locate the PTSD status and methylation values of each subject. Save the PTSD status and methylation values into separate files for further analysis.

There are 163 samples (81 PTSD, 82 controls) and 429,948 methylation sites in the data.

3. Identify significant PTSD-associated methylation sites.

WinSCP is used to transfer the PTSD status file and the methylation value file from Linux to the Windows laptop. Use the Excel program to open the methylation file where a column represents a sample and a row represents a methylation site. Data of the first 60 samples is saved into a new file and is used as a discovery group. Data of the last 103 samples is saved and used as a validation group. In the discovery methylation file, the average methylation value of each site is calculated for the PTSD and control groups, respectively. The difference between the two groups is also calculated. In addition, the t test function of Excel is used to obtain a p value for the difference in methylation between the two groups at each site. To select significant sites, the cutoff of absolute value of methylation difference is >= 0.015 and the p value is < 0.023. Names of the selected significant sites are saved. To obtain methylation values of these sites in the validation group, the match function is used in the RStudio program. The t test is performed on the data from the validation group and a sublist of significant sites were selected ($p < 0.05$). An Illumina 450k site information file is used together with the match function to identify gene names associated with the selected sites.

4. Functional enrichment analysis of PTSD-associated genes.

Go to the website of the DAVID program, click the Functional Annotation icon on the left panel, paste the gene names, select OFFICIAL_GENE_SYMBOL, check the Gene List box, and click Submit List. On the popup window, select Homo sapiens (Human) and click the Select Species icon. On the Annotation Summary Results panel, click to expand Functional_Categories, and save the tables of UP_KEYWORDS and UP_SEQ_FEATURE as text files. Click to expand

Gene_Ontology, and save the table of GOTERM_BP_DIRECT, GOTERM_CC_DIRECT, and GO_MF_DIRECT. Click to expand Pathways and save the table of KEGG_PATHWAY.

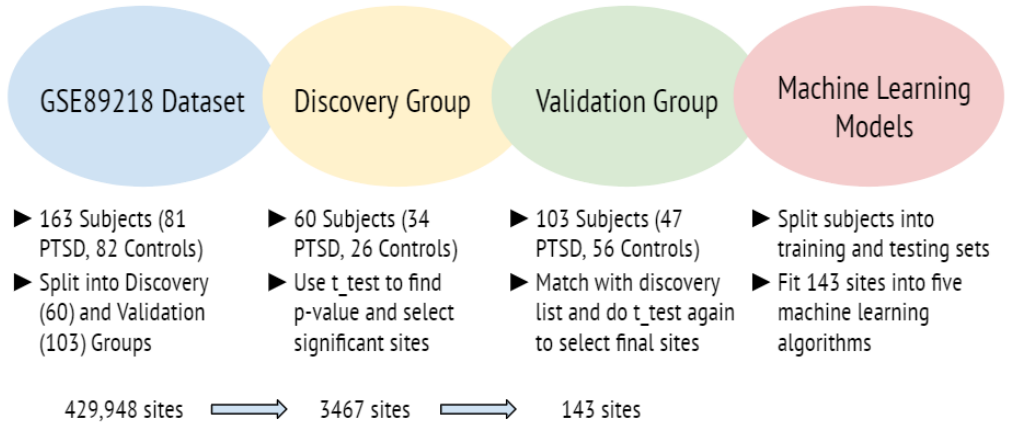
5. PCA analysis and plot heatmap of selected methylation sites.

Principal component analysis (PCA) is performed to obtain a quick assessment of the selected sites in their ability to separate the PTSD and control groups in the validation cohort. The ggplot2 and ggfortify R packages and the prcomp and autoplot functions are used to conduct the PCA analysis. The heatmap of methylation sites and hierarchical clustering of the samples in the validation group are performed using the made4 R package and the heatmap.2 function.

6. Build and evaluate machine learning models for PTSD diagnosis.

The Python programming language and the scikit-learn machine learning package are used in building five PTSD prediction models. The scikit-learn classes used are KNeighborsClassifier, LogisticRegression, SVC (support vector machine), RandomForestClassifier, and GradientBoostClassifier. The programming environment and platform is Anaconda3's JupyterNotebook. To start using JupyterNotebook, click the windows icon on the left bottom of Windows, select Anaconda3 on the menu and click Anaconda Navigator, and click Launch under JupiterNotebook 5.7.4 on the popup window. A new Jupiter window is open automatically on the computer's default web browser. Simply navigate to a desirable work folder, on the right panel click the New icon and select Python 3. A new web tab with a typing space appears and is ready for machine learning programming. Actual codes are listed in Appendix A.

Figure 8. Workflow of the Experiment



Data and Discussion/Analysis

Data:

In order to comprise a list of significantly differentiated DNA methylation sites, the GSE89218 dataset was randomly split into a discovery (34 PTSD, 26 controls) group and a validation group (47 PTSD, 56 controls). A preliminary list of 3,647 significant sites was first identified in the discovery group on the basis of each individual site's 't_test p-value' and 'Diff_Pos-Neg' as shown below in Table 1.

Table 1. A list of 3647 significant sites found in the discovery group

ID_REF	Avg_Pos	Avg_Neg	Diff_Pos-Neg	t_test p-value
cg27224751	0.585788593	0.41108378	0.174704812	0.006937972
cg05388281	0.268165974	0.104847613	0.163318361	0.0000728178
cg12897067	0.232749738	0.077332161	0.155417577	0.011294111
cg16611967	0.26251442	0.110959361	0.151555059	0.005475743
cg07158503	0.686180418	0.542759307	0.143421111	0.00960976

* Note that this is only a very small example of selected sites which are deemed significant

Key:

ID_REF - This column refers to the methylated site, typically near the promoter region which controls the expression of the gene. (ex. cg01655777, cg17132030)

Avg_Pos - The overall average chance as a percentage of the specific site being methylated in those who have developed PTSD.

Avg_Neg - The overall average chance as a percentage of the specific site being methylated in the control group (people who don't have PTSD).

Diff_Pos-Neg - The difference in methylation between Avg_Pos and Avg_Neg. The higher the absolute value, the more the two averages for a specific site are different. In order for the Diff_Pos_Neg value to be useful, the absolute value cutoff is set to be above 0.015.

t_test p-value - A t test is a type of statistic test with the purpose of determining if there is a significant difference between the averages of two specific groups, in this case, the Avg_Pos and Avg_Neg. The output, the t test's p-value, helps indicate if there is strong evidence against the null hypothesis of the experiment. Typically, a p-value < 0.05 means that the difference between the averages of the two groups is significant and thus can be used to reject the null hypothesis.

DNA methylation values of these sites were again extracted from the validation group using the match function of RStudio and subjected to t test. A list of 150 sites was further selected with p-value < 0.05. A final list contained 143 significant sites after excluding seven sites, which were no longer present in the latest version of Illumina's 850k DNA methylation array (Table 2).

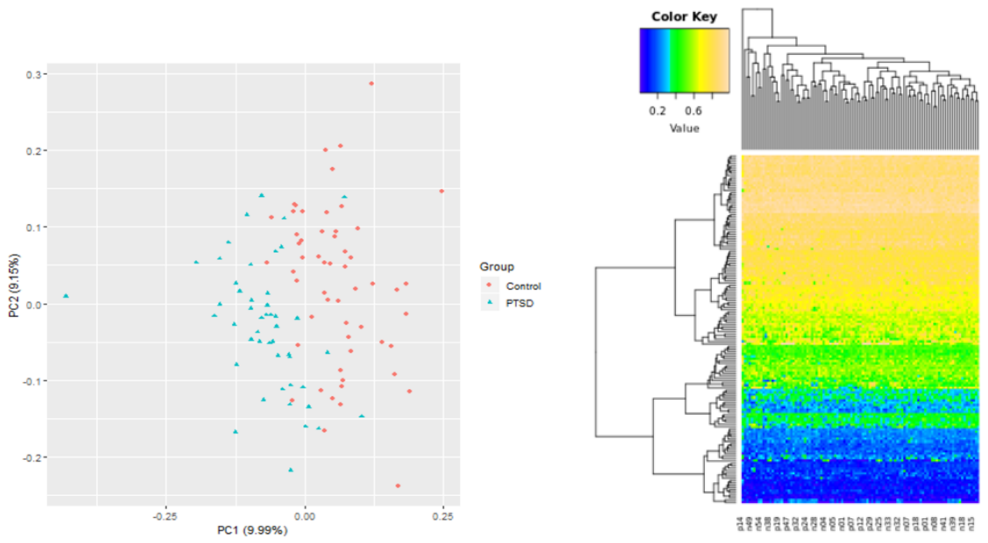
Table 2. A list of 143 significant sites in the validation set

ID_REF	GSM2361079	GSM2361080	GSM2361081	GSM2361082
cg01655667	0.474192043	0.303994547	0.649783226	0.433027857
cg17132030	0.034704582	0.133625897	0.158427957	0.407896318
cg01208318	0.229308564	0.267993701	0.284701453	0.311373831
cg13387235	0.877388555	0.908505541	0.908466235	0.899829319
cg19976404	0.794989117	0.766997955	0.785471562	0.810778825

* Note that this is only a very small example of the final selected sites which are deemed significant

To investigate how well these selected sites can separate the PTSD and control groups, two visual representations were built with principal component analysis (PCA) and a heatmap of methylation sites alongside hierarchical clustering of the samples (Figure 8).

Figure 9. PCA analysis and clustering showing the separation of PTSD and control groups in the 143 selected sites



Through Illumia's Probe Information File, 86 genes in total were identified to be associated with certain sites in the final list (Table 3).

Table 3. A list of genes associated with certain sites in the final list

ACOT7	ARMC7	BRCA1	C7orf20
ACOXL	B4GALT1	C14orf102	CACNG5
ACTC1	B4GALT4	C2orf27B	CALM1
AP2S1	BMP2	C2orf3	CAPZB
ARHGEF18	BOD1	C5orf43	CUEDC1

* Note that this is only a portion of the total 86 genes associated with the list of 143 significant sites

Through the DAVID program, 37 functions and features of these genes were identified, along with its p_value and the number of genes associated with each feature (Table 4).

Table 4. The functions and features of the 86 genes identified

Term	Gene Count	%	P Value
Glycosyltransferase	6	7.69230769230769	0.00181186426302877
Phosphoprotein	44	56.4102564102564	0.0025998904063424
Alternative splicing	51	65.3846153846153	0.00796376986644318
Lipid metabolism	6	7.69230769230769	0.0219947353540561

* Note that this is only a portion of the total 37 functions and features of the 86 genes.

Figure 10. Functional enrichment analysis of the 86 genes through DAVID, showing the UP_KEYWORDS and UP_SEQ_FEATURE.

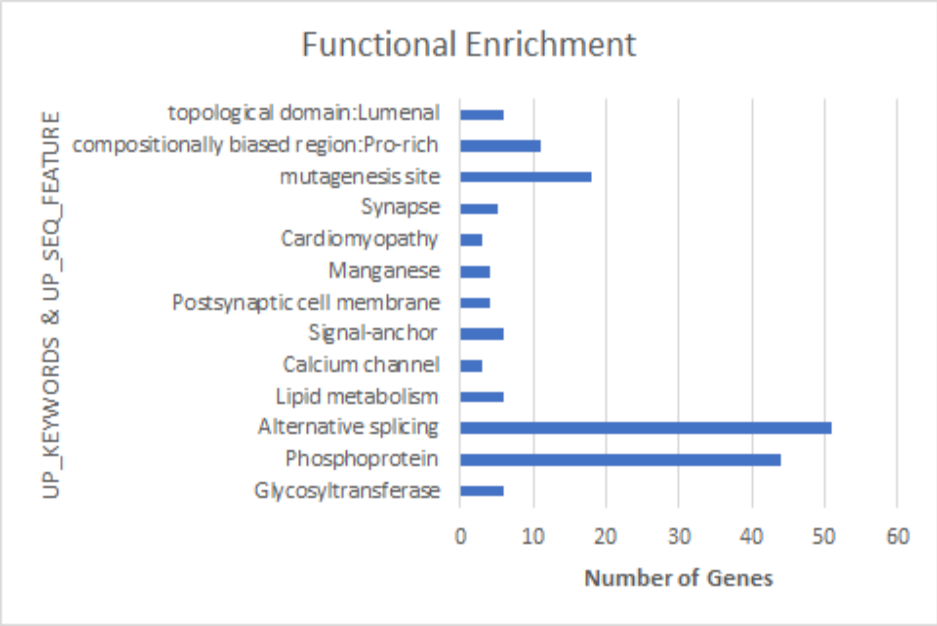
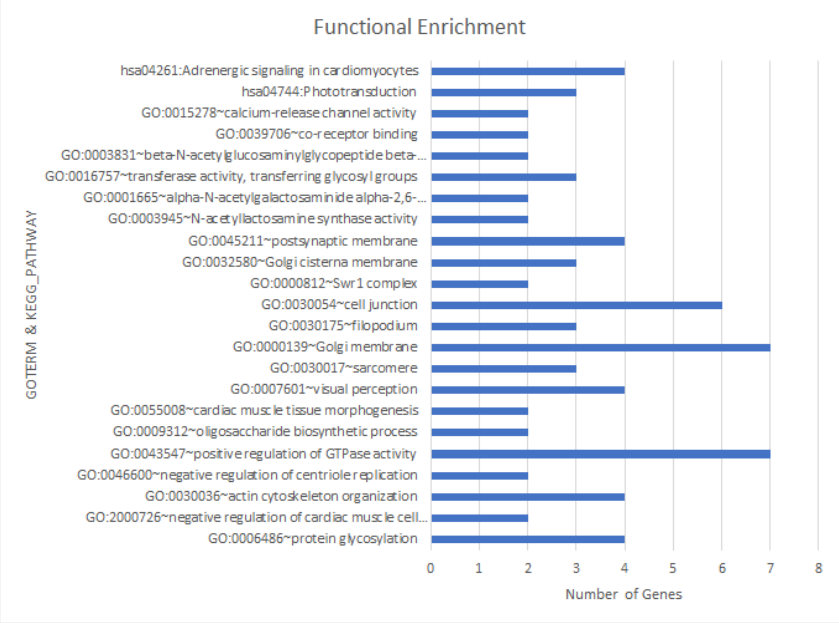


Figure 11. Functional enrichment analysis of the 86 genes through DAVID, showing the enriched terms of gene ontology (GOTERM) and pathways of the KEGG database (KEGG_PATHWAY).



In addition, the validation group was split into a training set and a testing set. Five PTSD prediction models were built by using multiple supervised machine learning algorithms. An algorithm takes a set of input data (features) and known responses to the data (outcomes) and trains a model, which can be used to make predictions of outcome for new data. In this case, the data given to each model is DNA methylation values of the list of 143 significant sites mentioned above. Each method is unique, performing different tasks in order to predict whether or not a person has PTSD. Figure 5 demonstrates the results of my models with different algorithms.

Table 5. Accuracies of five PTSD prediction models on the training and testing data

Model (Algorithms):	Training Set Accuracy:	Testing Set Accuracy:
KNeighborsClassifier	**79.2%**	**80.8%**
LogisticRegression	**89.6%**	**92.3%**
SVC (Support Vector Classifier)	**89.6%**	**92.3%**
RandomForestClassifier	**85.7%**	**92.3%**
GradientBoostingClassifier	**74%**	**61.5%**

Key:

Model (Algorithms): Machine learning algorithms are programs (math and logic) that adjust themselves to perform better as they are exposed to more data.

Training Set Accuracy: The training set accuracy is the accuracy of the individual model based on the examples or pre-determined datasets it was constructed on.

Testing Set Accuracy: The testing set accuracy is the accuracy of the model on data it hasn't seen before, therefore simulating a real scenario of predicting PTSD in this case.

Discussion/Analysis:

The DNA methylation results obtained from my study supports both the claim made by my background research that DNA methylation is connected with PTSD as well as my hypothesis. In the study done by U-M researchers, it was reported that numerous genes, the sequences of nucleotides in DNA or RNA that encodes the synthesis of the genes' product, were identified to be more active in those with PTSD, and that DNA methylation may be involved. Figure 2 and Figure 3 supports this notion, as 143 sites with differential methylation levels between PTSD and controls were identified, as well as the 86 genes that these sites are a part of. This discovery implies that there is a connection between DNA methylation and the development of PTSD, which agrees on what the U-M researchers reported. In addition, the functional enrichment analysis showed the numerous functions and features of the 86 genes that are associated with the 143 selected sites. For example, synapse genes were significantly enriched. A synapse is a junction between two nerve cells, across which impulses pass by diffusion of a neurotransmitter, which might imply a connection with the development of PTSD. Phosphoprotein is a protein posttranslationally modified by attaching either simply a phosphate group, or a complex molecule through a phosphate group. Phosphoproteins play an important role in regulation of cells, and they were enriched in the 86 genes as well.

In addition, my research further supports my hypothesis through the creation of multiple successful prediction-based machine learning models. Table 5 displays the results of my models with SVC (Support Vector Classifier) and Logistic Regression tied for the best results, each having a testing accuracy of 92.3%. This indicates that instead of a 50/50 chance of predicting PTSD, these models can accurately predict PTSD 92.3% of the time, which is a significant

improvement. Again, this implies that there is a correlation between these 143 methylation sites and development and prediction of PTSD. These results support my hypothesis that the use of DNA methylation as a biomarker to diagnose PTSD may be feasible.

Conclusions

Post-traumatic stress disorder is a mental condition that affects around 8 millions of people in the U.S alone in a given year, with a current diagnosis process that is inefficient. To address this issue, I conducted an experiment to explore whether DNA methylation can be used as a biomarker to diagnose PTSD. I discovered 143 significant sites that have a high chance of being associated with PTSD. Through this list of methylation sites, I built five machine learning models that could fairly accurately predict PTSD, with Logistic Regression and SVC having the highest testing accuracy of 92.3%. The ability of these machine learning models to predict PTSD with high accuracy supports the idea that DNA Methylation plays a big role in the development and existence of PTSD. The list of 143 significant CpG sites and the genes they are associated with demonstrate a correlation between these specific methylated areas and the disorder itself and as such can be used as a predictor for PTSD diagnosis in individuals. In all, as supported by my findings and research, it can be reasonably concluded that my hypothesis of the feasibility of using DNA Methylation as a biomarker to diagnose PTSD, was indeed supported. Despite this, more research with large numbers of subjects should be done not only to learn and discover more about the root causes of PTSD in the term of DNA methylation but in other diseases and conditions as well.

Literature Cited/Reference List

Arbor, Ann. “U-M Study Shed Light on the Biological Roots of Post-Traumatic Stress Disorder.” *University of Michigan News*, 7 May 2010, news.umich.edu/u-m-study-sheds-light-on-the-biological-roots-of-post-traumatic-stress-disorder/

Ives, Laurel. “'Higher Levels of PTSD among Veterans', Says Study.” *BBC News*, BBC, 8 Oct. 2018, www.bbc.com/news/health-45761546.

Müller Andreas Christoph, and Sarah Guido. *Introduction to Machine Learning with Python: a Guide for Data Scientists*. O'reilly Et Associates Inc, 2016.

Parekh Ranna, and Felix Torres. “What Is Posttraumatic Stress Disorder?” *What Is PTSD?*, www.psychiatry.org/patients-families/ptsd/what-is-ptsd.

“Symptoms of PTSD.” *Anxiety and Depression Association of America, ADAA,* adaa.org/understanding-anxiety/posttraumatic-stress-disorder-ptsd/symptoms.

Appendix A

Code for Machine Learning Models

Figure 1: Essential Libraries and Tools

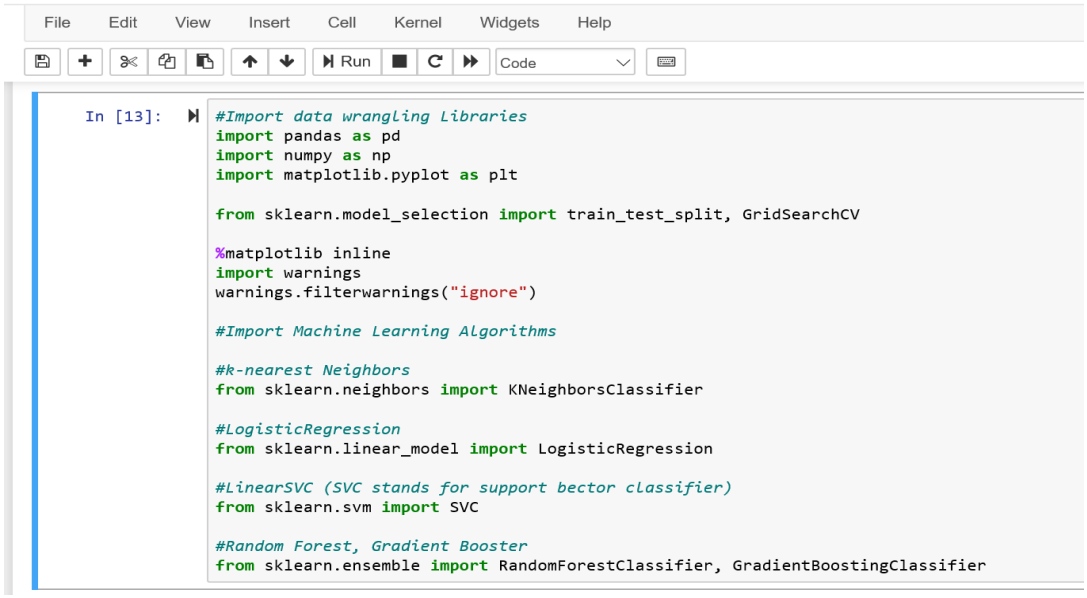


Figure 2: Reading in Data

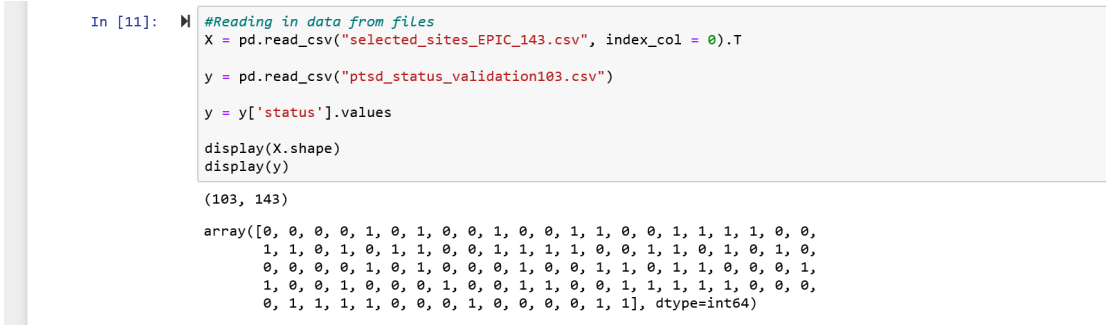


Figure 3: Splitting our Data into Training and Testing Sets

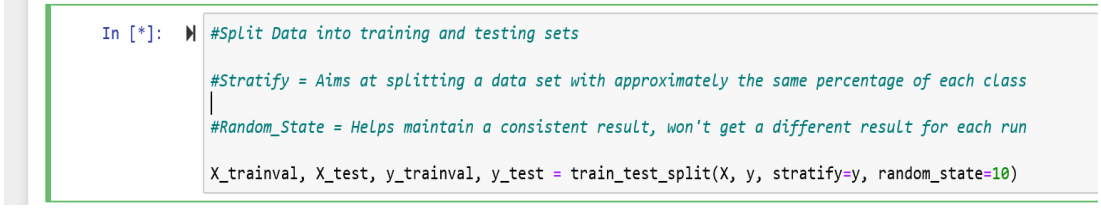


Figure 4: KNeighborsClassifier - Best Parameters and Training Set Accuracy

```
#Finding the best paramaters for KNeighborsClassifier
param_grid = {'n_neighbors' : [1,2,3,4,5]}
grid_search=GridSearchCV(KNeighborsClassifier(), param_grid, cv=5)
grid_search.fit(X_trainval, y_trainval)

#Model Best Paramaters
print(grid_search.best_params_)

#Extra space
print()

#Model Accuracy - Training Set
print("Training set accuracy: {:.3f}".format(grid_search.best_score_))
```

```
{'n_neighbors': 3}

Training set accuracy: 0.792
```

Figure 5: KNeighborsClassifier - Testing Set Predictions and Accuracy

```
#KNeighborsClassifier
knc = KNeighborsClassifier(n_neighbors=3)
knc.fit(X_trainval, y_trainval)

#Test predictions
print("Test set predictions: {}".format(knc.predict(X_test)))

#Extra space
print()

#Model Accuracy - Test Set
print("Test set accuracy: {:.3f}".format(knc.score(X_test, y_test)))
```

```
Test set predictions: [0 1 0 1 1 1 1 0 1 1 1 1 0 0 1 0 1 1 1 0 1 1 0 1 1 0]

Test set accuracy: 0.808
```

Figure 6: LogisticRegression - Best Parameters and Training Set Accuracy

```
#Finding the best paramaters for Logistic Regression
param_grid = {'C': [0.01,0.1,1,2,5,8,10,20]
              }
grid_search=GridSearchCV(LogisticRegression(random_state=0), param_grid, cv=5)
grid_search.fit(X_trainval, y_trainval)

#Model Best Paramaters
print(grid_search.best_params_)

#Extra space
print()

#Model Accuracy - Training Set
print("Training set accuracy: {:.3f}".format(grid_search.best_score_))
```

```
{'C': 5}

Training set accuracy: 0.896
```

Figure 7: LogisticRegression - Testing Set Predictions and Accuracy

```
#Logistic Regression
logreg = LogisticRegression(C=5)
logreg.fit(X_trainval, y_trainval)

#Test predictions
print("Test set predictions: {}".format(logreg.predict(X_test)))

#Extra space
print()

#Model Accuracy - Testing Set
print("Testing set accuracy: {:.3f}".format(logreg.score(X_test, y_test)))
```

```
Test set predictions: [0 1 0 0 1 1 1 0 1 1 0 1 0 0 0 0 1 1 1 0 1 0 0 0 1 0]

Testing set accuracy: 0.923
```

Figure 8: SVC - Best Parameters and Training Set Accuracy

```
#Finding the best paramaters for SVC
param_grid = [{'kernel':['linear'],
               'C':[0.001,0.01,0.1]},
              {'kernel':['rbf'],
               'C':[0.001,0.01,0.1,1,1.1,1.5,2,3,3.5,4,5,7,10],
               'gamma':[0.001,0.01,0.1,1,1.5,2,3,3.4,5,7,10]}]
grid_search = GridSearchCV(SVC(random_state=0), param_grid, cv=5)
grid_search.fit(X_trainval, y_trainval)

#Model Best Paramaters
print(grid_search.best_params_)

#Extra space
print()

#Model Accuracy - Training Set
print("Training set accuracy: {:.3f}".format(grid_search.best_score_))
```

```
{'C': 1, 'gamma': 1, 'kernel': 'rbf'}

Training set accuracy: 0.896
```

Figure 9: SVC - Testing Set Predictions and Accuracy

```
#SVC
svc = SVC(C=1, gamma=1, kernel='rbf', random_state=0)
svc.fit(X_trainval, y_trainval)

#Test Predictions
print("Test set predictions: {}".format(svc.predict(X_test)))

#Extra space
print()

#Model Accuracy - Testing Set
print("Testing set accuracy: {:.3f}".format(svc.score(X_test, y_test)))
```

```
Test set predictions: [0 1 0 0 1 1 1 0 1 1 0 1 0 0 0 0 1 1 1 0 1 0 0 0 1 0]

Testing set accuracy: 0.923
```

Figure 10: Random Forest - Best Parameters and Training Set Accuracy

```
#Find the best paramaters for RandomForestClassifier
param_grid = RandomForestClassifier()
param_grid = {
    "n_estimators":[5,10,50,60,70,80,85,90,100,250],
    "max_depth":[1,2,3,4,5,16,32,None]
}

grid_search = GridSearchCV(RandomForestClassifier(random_state=0), param_grid, cv=5)
grid_search.fit(X_trainval, y_trainval)
print(grid_search.best_params_)

#Extra space
print()

#Model Accuracy - Training Set
print("Training set accuracy: {:.3f}".format(grid_search.best_score_))
```

```
{'max_depth': 4, 'n_estimators': 80}

Training set accuracy: 0.857
```

Figure 11: Random Forest - Testing Set Predictions and Accuracy

```
#RandomForstClassifier
forest = RandomForestClassifier(n_estimators=80, max_depth=4, random_state=0)
forest.fit(X_trainval, y_trainval)

#Test Predictions
print("Test set predictions: {}".format(forest.predict(X_test)))

#Extra space
print()

#Model Accuracy - Testing Set
print("Testing set accuracy: {:.3f}".format(forest.score(X_test, y_test)))
```

```
Test set predictions: [0 0 0 0 1 1 1 0 1 1 0 1 0 1 0 0 1 1 1 0 1 0 0 0 1 0]

Testing set accuracy: 0.923
```

Figure 12: GradientBoostingClassifier - Best Parameters and Training Set Accuracy

```
#Finding the best paramaters for GradientBoostingClassifier
param_grid={'max_depth':[1,2,3,4,5],
           'learning_rate': [0.005, 0.01, 0.015, 1]
           }
grid_search=GridSearchCV(GradientBoostingClassifier(random_state=0), param_grid, cv=5)
grid_search.fit(X_trainval, y_trainval)

print(grid_search.best_params_)

#Extra space
print()

#Model Accuracy - Training set
print("Training set accuracy: {:.3f}".format(grid_search.best_score_))
```

```
{'learning_rate': 0.01, 'max_depth': 1}

Training set accuracy: 0.740
```

Figure 13: GradientBoostingClassifier - Testing Set Predictions and Accuracy

```
#GradientBoostingClassifier
gbc=GradientBoostingClassifier(random_state=0, learning_rate = 0.01, max_depth = 1)
gbc.fit(X_trainval, y_trainval)

#Test Predictions
print("Test set predictions: {}".format(gbc.predict(X_test)))

#Extra space
print()

#Model Accuracy - Testing set
print("Testing set accuracy: {:.3f}".format(gbc.score(X_test, y_test)))
```

```
Test set predictions: [0 1 1 1 1 0 0 0 0 1 1 0 0 1 0 0 1 1 0 0 0 0 0 0 1 0]

Testing set accuracy: 0.615
```

Appendix B

Helpful Resources

1) "A Practical Guide to Linux Commands, Editors and Shell Programming" by Mark G. Sobell

2) "Introduction to Machine Learning with Python: A Guide for Data Scientists" by Andreas C. Müller & Sarah Guido

3) DAVID (The Database for Annotation, Visulatzation, and Integrated Discovery) Bioinformatics Resources 6.8

4) Pubmed - National Center for Biotechnology Information

Appendix C

Rough Draft/Plan of Notebook

Figure 1:

Topic: Using Machine Learning and DNA Methylation to Diagnose PTSD

Hypothesis: DNA methylation, which affects gene expression, can be used as a biomarker to diagnose post-traumatic stress disorder

What to work on: PTSD is a psychological and physiological response in a fraction of people exposed to a potentially traumatic event, involving life threat, serious injury, or death. Soldiers, doctors, emergency workers are among those with high risk to PTSD. DNA methylation is a chemical process which adds a methyl group to cytosines in a DNA sequence, affecting gene expression. Recent evidence suggests a causal association of DNA methylation patterns and PTSD. I postulate that a machine learning model can be built to diagnose PTSD using DNA methylation levels of blood samples. Machine learning algorithms, computing power and genome-wide methylation profiling data have advanced tremendously and have become widely available. They will be used in the current project of developing PTSD diagnostic model.

Materials:

- Public databases such as PubMed, which host a large amount of accessible datasets on life sciences and biomedical topics, will be used to view and download data related to PTSD and DNA methylation as well as to investigate the biological function and meaning of PTSD-associated genes.

- Readily available data science analysis tools and programming languages (ex. Python, Jupyter Notebook, Linux)

Procedure: DNA methylation levels of blood samples from certified PTSD patients and controls is obtained digitally from the National Center for Biotechnology Information (NCBI). To generate a machine learning model for PTSD diagnosis, the samples will be randomly split into training sets, where it is 'trained' and used to build the model off of, and testing sets, where new data is used to access the performance and accuracy of the model. Necessary preprocessing and quality control steps will be taken such as the checking and removal of outliers. The model will be based off of machine learning algorithms such as supervised learning and unsupervised learning. A diagnostic model with the best accuracy of diagnosing PTSD will be selected in hopes of not only assisting physicians in their medical decisions, but to act as a stepping stone for those who developed PTSD to get the help that they need.

Steps:

1) Discover datasets of DNA methylation levels of both PTSD patients and controls from resources such as PubMed or GEO (Gene Expression Omnibus)
2) Utilizing Linux, do preprocessing on the datasets
3) Split into testing and training sets
4) Then, create a model which will read the training set with pre-made code (the goal is to find a specific pattern/trend which exists at least most of the time - in those diagnosed with PTSD, in order to find which sites are methylated as a result of developing PTSD)
5) Then, find certain genes (data) or methylation sites which are identified using probability (p) values

6) Using the information gained by the model, we can find who has (or most likely to have) PTSD by checking if their (specific) genes are methylated and align with the pattern of those who are confirmed to have PTSD
7) The application of this model are quick significant, as we can now have a (mostly) reliable and scientific method in diagnosing PTSD. Through this, we can have those who are more likely to develop PTSD (such as law enforcement and military personnel) to take an annual test by using the model in order to figure out if they have PTSD, which can then allow them to be given the appropriate help they need. This will, in addition, help reduce the chances of possible self-harm or suicide by the hands of this disorder. Not only that, this model can act as a method in finding out how PTSD treatment is going for those who have PTSD.

Goals:

- Learn how to code in Linux and properly separate the code
- Learn how to split the code and develop a pattern by finding genes which are methylated in PTSD patients

General Research:

What is PTSD?

- Post-traumatic stress disorder is a mental health condition that's triggered by a terrifying event - either experiencing it or witnessing it.

What are its symptoms?

- Symptoms may include flashbacks, nightmares and severe anxiety, as well as uncontrollable thoughts about the event.

How is PTSD applicable in our world today?

- Anyone can develop PTSD at any age. This includes war veterans, children, and people who have been through a physical or sexual assault, abuse, accident, disaster, or other servious events.
- According to the National Center for PTSD, about 7 or 8 out of every 100 people will experience PTSD at some point in their lives

How does PTSD develop (scientifically)?

- In a study done by University of Michigan researchers, it identified what appears to be a crucial step in the chain of biological events leading to post-traumatic stress disorder.
- Their findings support the idea that exposure to a traumatic event can trigger genetic changes that alter the body's immune system, leading to PTSD.
- The new U-M findings identified a specific biochemical reaction that may be involved. That biochemical reaction is a process called DNA methylation, in which methyl groups (CH3 groups) are added to some of the molecular letters that spell out the genetic code. DNA methylation can alter gene activity, typically reducing it.

How is PTSD diagnosed?

- Currently, PTSD is only diagnosed through viewing if the patient has symptoms associated with post-traumatic stress disorder, there is no scientific way to do so.

Why is this important to your overall goal for this project?

- This is essential to my project idea as my overall goal is to create a model that can be used to diagnose PTSD through scientific means, specifically by creating a pattern of certain methylated sites (genes) that are (mostly) exclusive to those who have

developed PTSD. Thus, this pattern can be used as a comparison with a testing set to differentiate who has or doesn't have this disorder.

Appendix D

Copyright Page

1) Introduction to Machine Learning with Python by Andreas C. Müller and Sarah Guido

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95272

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (https://oreilly.com/safari). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreily.com*

Editor: Dawn Schanafelt

Indexer: Judy McConville

Production Editor: Kristen Brown

Interior Designer: David Futato

Copyeditor: Rachel Head

Cover Designer: Karen Montgomery

Proofreader: Jasmine Kwityn

Illustrator: Rebecca Demarest

October 2016: First Edition

Revision History for the First Edition

2016-09-22: First Release

2017-01-13: Second Release

2017-06-09: Third Release

See http://oreilly.com/catalog/errata.csp?isbn=9781449369415 for release details

2) Sobell, Mark G.

A Practical Guide to Linux Commands, Editors, and Shell Programming / Mark G. Sobell

p. Cm.

Includes bibliographical references and index.

ISBN 0-13-147823-0 (alk.paper)

1. Linux. 2. Operating systems (Computers) I. Title.

Appendix E

Original Data Log

Figure 1: A Sample of the Raw Dataset in Excel

* Note that this only a small representation of the full dataset, as uploading the full raw data is impossible since there consists a total of 429,948 individual rows depicting seperate sites

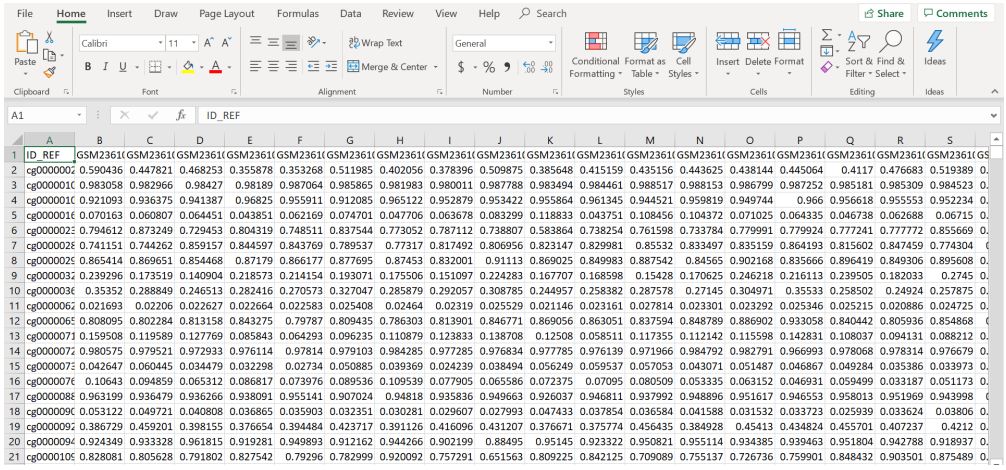


Figure 2: A Sample of the Raw Dataset in Linux

* Note that this only a small representation of the full dataset, as uploading the full raw data is impossible

Appendix F

Photographs

Figure 1. Linux

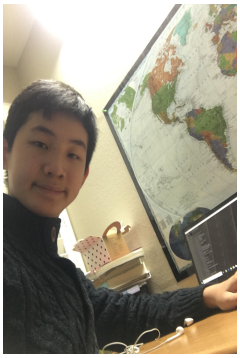


Figure 2. Excel Spreadsheet

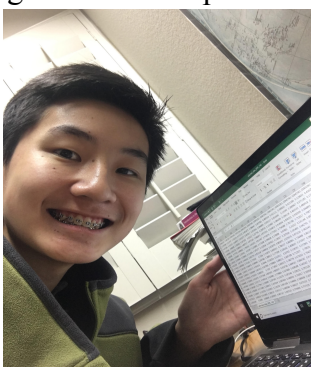
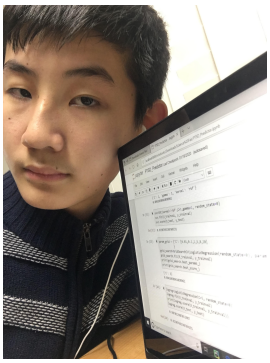


Figure 3. Jupyter Notebook

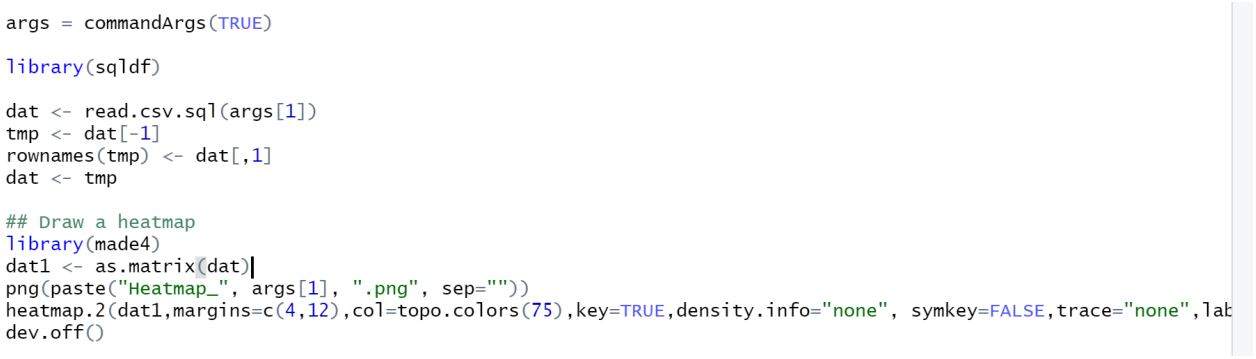


Appendix G

Code For PCA and Heatmap

Figure 1. PCA

Figure 2. Heatmap



Appendix H

Acknowledgments

From William Guo,

I would like to conclude this paper by expressing my utmost gratitude to my Dad who helped supervise this project on Using Machine Learning and DNA Methylation to Diagnose PTSD. I learned so much about so many different topics along this journey, including how to code in new programming languages as well as the science behind PTSD and DNA Methylation, and it couldn't have been done without his guidance and support. Whenever I was stuck or needed advice, my Dad was always there to provide me with assistance to the best of his abilities. In short, this project wouldn't have existed without him and I'm thankful for his everlasting and endless support.