Using Machine Learning and DNA Methylation to Diagnose PTSD

William Guo – Northwood High School – Irvine, CA 92620

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 this disorder, which, in recent studies, has been suggested to have a causal association with DNA methylation, a chemical process which adds a methyl group (CH_3) to cytosines in a DNA sequence.

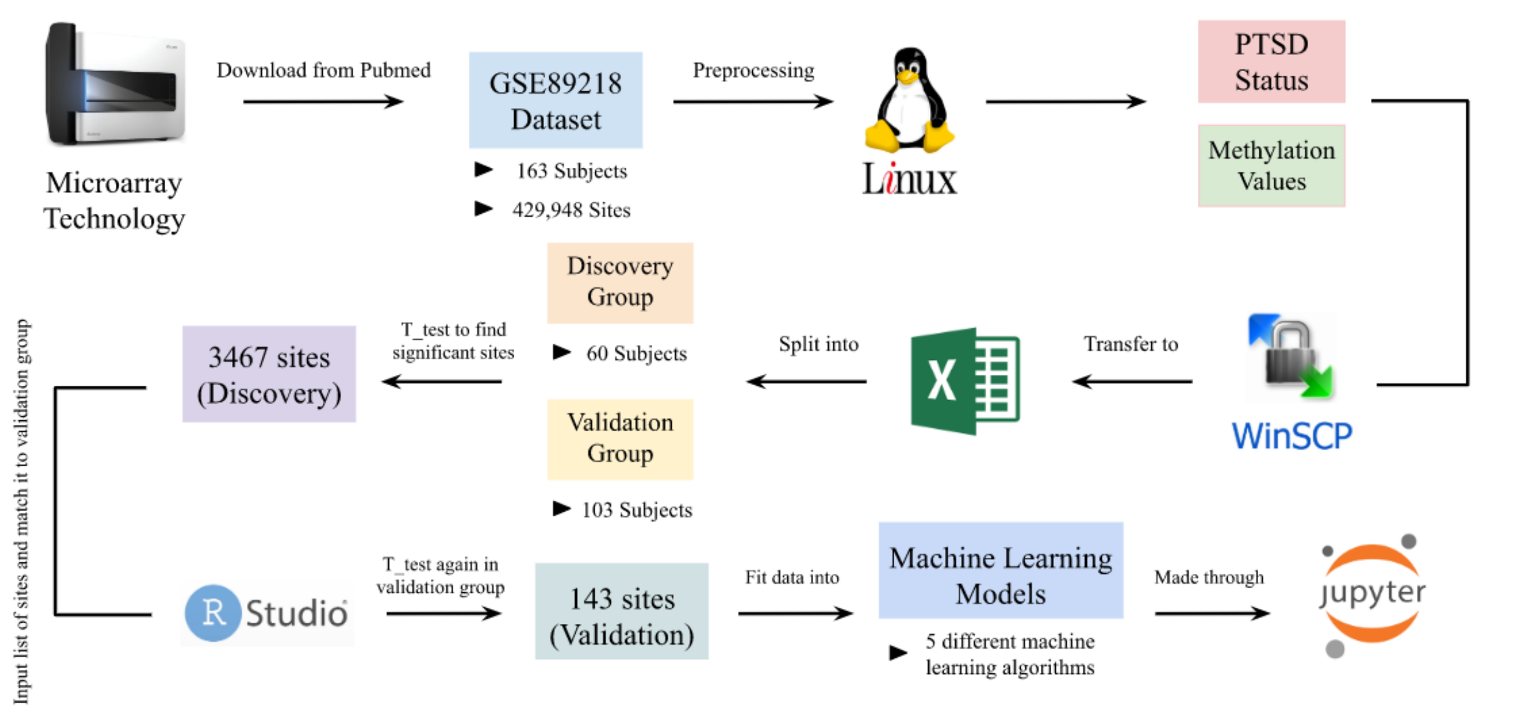


Figure 1: Simplified roadmap of the experiment

Numerous machine learning algorithms were used to build multiple models through taking significant methylated sites out of a list of 429,947 from 163 subjects containing both patient (81) and control (82) 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 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%, 91.7% sensitivity, and 92.9% specificity.

This means that the model can be expected 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 using DNA methylation as a biomarker to diagnose PTSD was indeed supported. A PTSD diagnostic model using DNA methylation and machine learning was successfully built in the hopes of assisting physician in their medical decision making.

BACKGROUND

- Approximately 3.5% of U.S adults are affected with PTSD. An estimated one in 11 people will be diagnosed with this disorder in their lifetime (Parekh, Torres 2017).
- In the United States, despite the rise and clear prevalence of reported PTSD cases, there is still a need for a molecular diagnostic method to identify if someone has PTSD. The main method of diagnosis currently is through comparing symptoms.
- Researchers from the University of Michigan published a study where they identified numerous genes that appeared more active in people with PTSD. The findings also identified a specific biochemical reaction that may have a causal association: DNA methylation (Arbor, 2010).

HYPOTHESIS

DNA methylation, which affects gene expression, can be used as a biomarker to diagnose post-traumatic stress disorder (PTSD).

METHODS

DNA methylation levels taken from blood samples of PTSD patients (81) and controls (82) measured using methylation array technology are downloaded from the GEO database in PubMed. Samples are randomly split into discovery and validation groups.

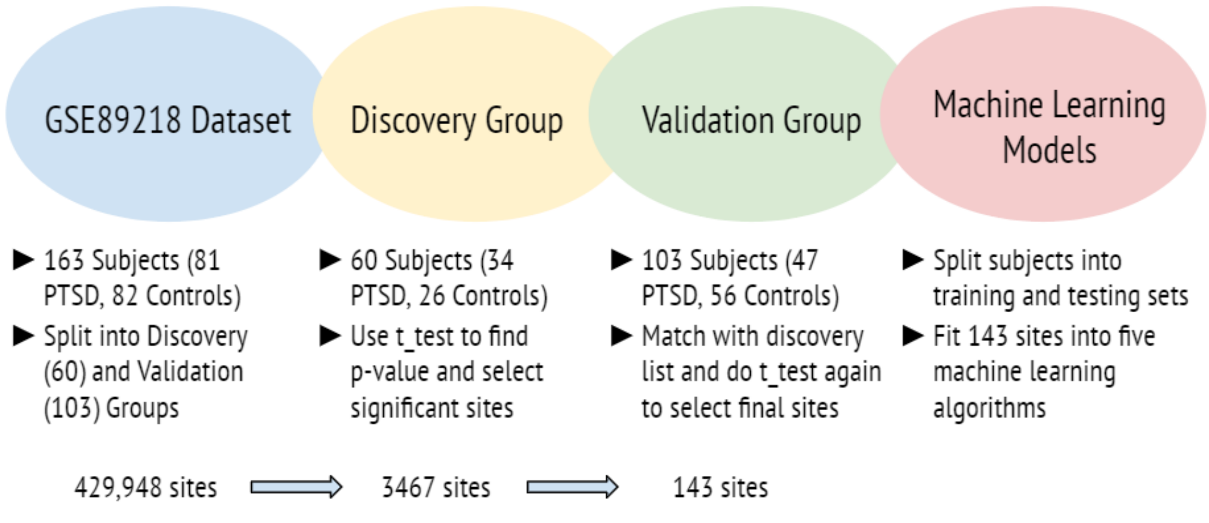
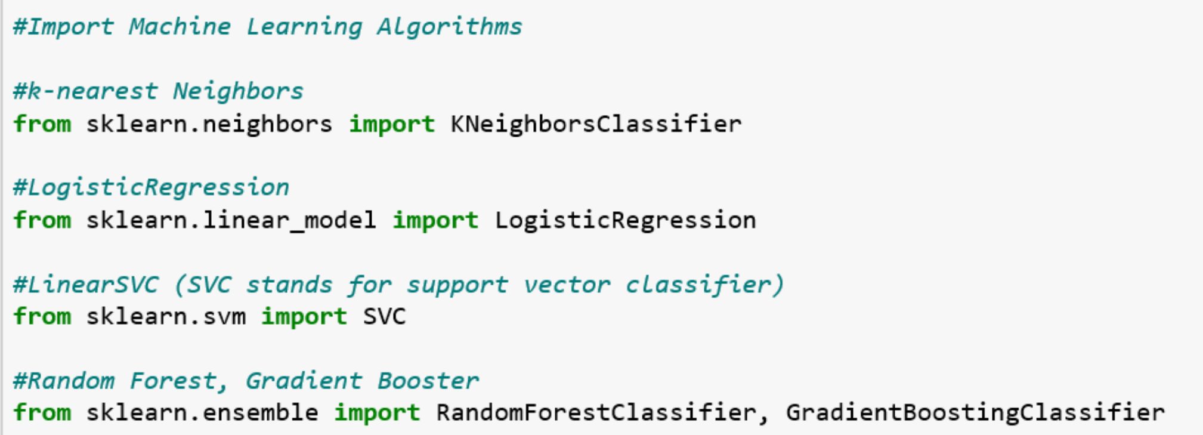

GSE89218 Dataset
Discovery Group
Validation Group
Machine Learning Models
163 Subjects (81 PTSD, 82 Controls)
Split into Discovery (60) and Validation (103) Groups
60 Subjects (34 PTSD, 26 Controls)
Use t_test to find p-value and select significant sites
103 Subjects (47 PTSD, 56 Controls)
Match with discovery list and do t_test again to select final sites
Split subjects into training and testing sets
Fit 143 sites into five machine learning algorithms
429,948 sites
3467 sites
143 sites

Figure 2: The workflow of the experiment

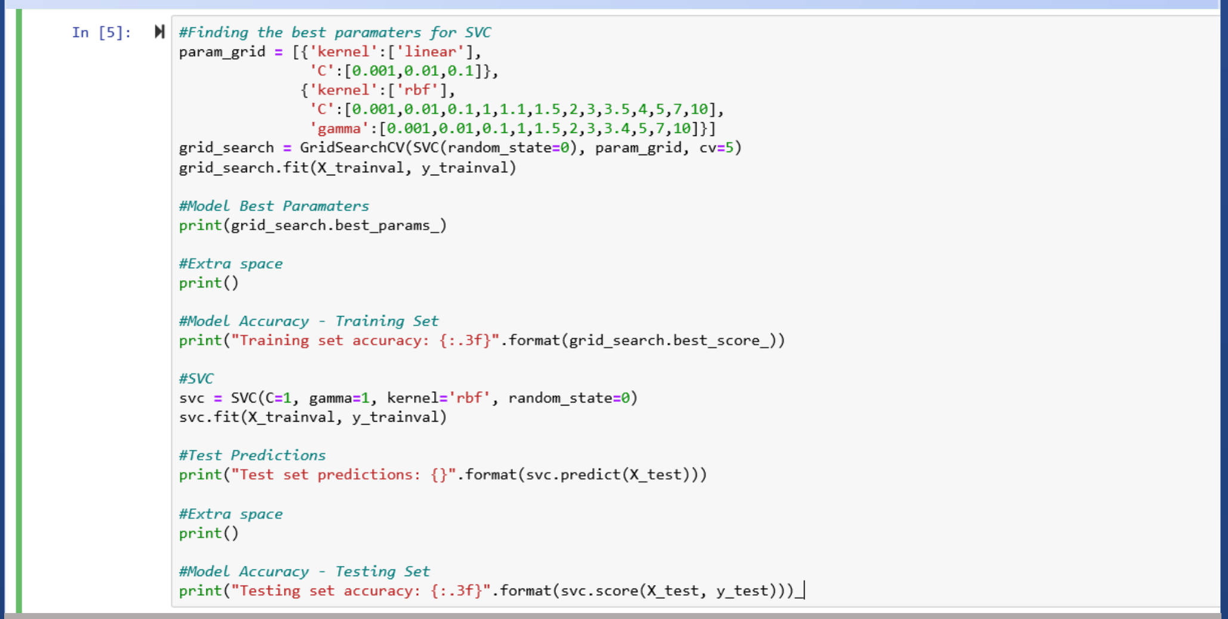
A list of methylation sites with varying methylation levels between patients and controls is identified using probability (p) values in the Student's t-test in the discovery group. Data of these sites in the validation group are then extracted and analyzed again by t-test. A sub list of significant sites is selected ($p < 0.05$). The validation cohort is split into training and testing sets for the machine learning models. Five machine learning algorithms are trained and tested.



```
#Import Machine Learning Algorithms

#k-nearest Neighbors
from sklearn.neighbors import KNeighborsClassifier

#LogisticRegression
from sklearn.linear_model import LogisticRegression

#LinearSVC (SVC stands for support vector classifier)
from sklearn.svm import SVC

#Random Forest, Gradient Booster
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
```

Figure 3: Importing ML algorithms into Jupyter Notebook

A PTSD diagnostic model with the best accuracy of predicting PTSD is selected in hopes of assisting physicians in their medical decision making.



```
In [5]:  #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_))

         #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)))
```

Figure 4: Main code for PTSD Predictor Model
* Note that this is code for only one machine learning model (SVC)

MATERIALS

Physical Computer Hardware:

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

Digital Dataset and Computer Software:

GSE89218 Dataset	DAVID Informatics	RStudio
Jupyter Notebook	Microsoft Excel	Linux

RESULTS

1. A list of 143 sites out of 429,948 were determined to be significantly associated with PTSD.

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

Table 1: A list of 143 significant sites in the validation set
* Note that this is a very small sample of the final selected sites

2. 86 genes were identified to be associated with certain sites in the list. In addition, 37 functions and features of these genes were found through DAVID.

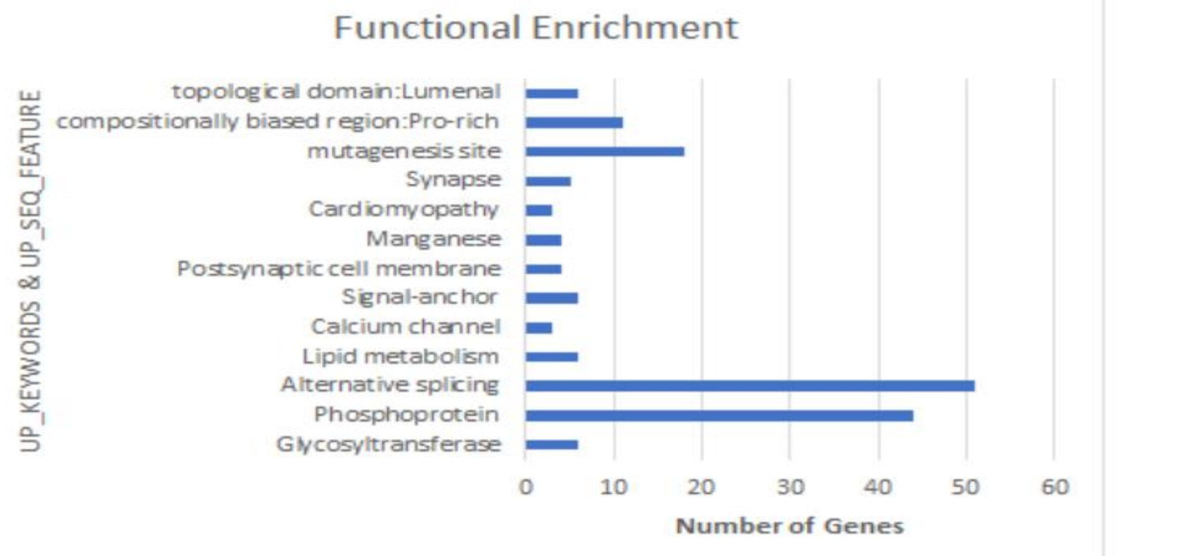


Figure 5: Functional Enrichment of 86 genes

3. The 143 sites were shown to separate PTSD and control groups by PCA analysis and clustering.

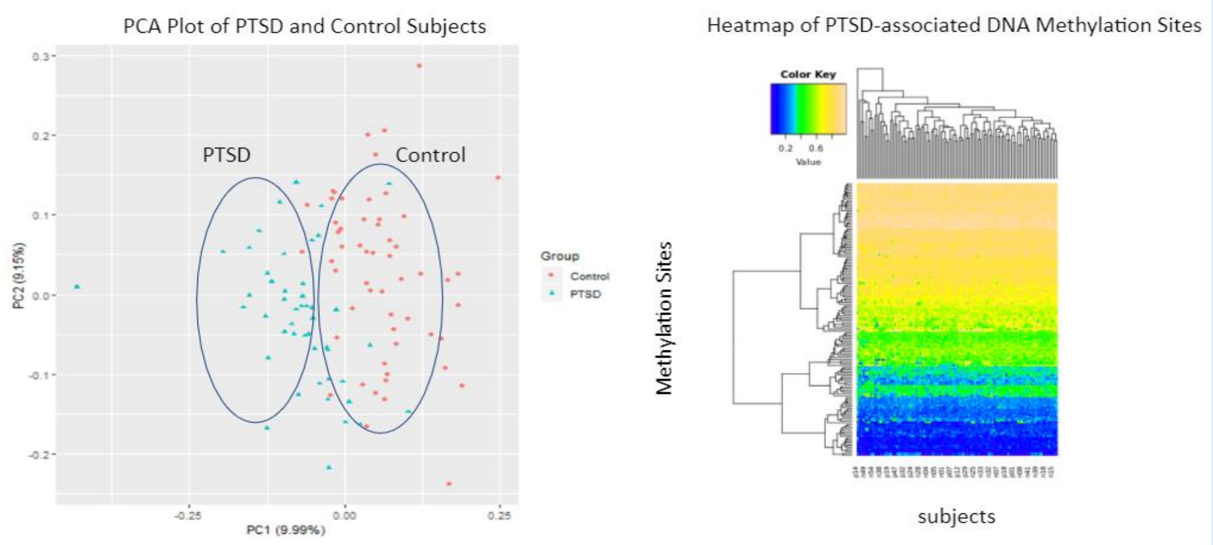


Figure 6: Separation of PTSD and control subjects in the validation group using the list of 143 significant sites

RESULTS

4. Five machine learning models using various algorithms were each built with the same methylation values in the list of 143 sites.

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	**74.0%**	**61.5%**

Table 2: Accuracies of five PTSD prediction models on the training and testing data

5. The SVC PTSD prediction model achieved 91.7% sensitivity, 92.9% specificity, and 92.3% accuracy in the testing set.

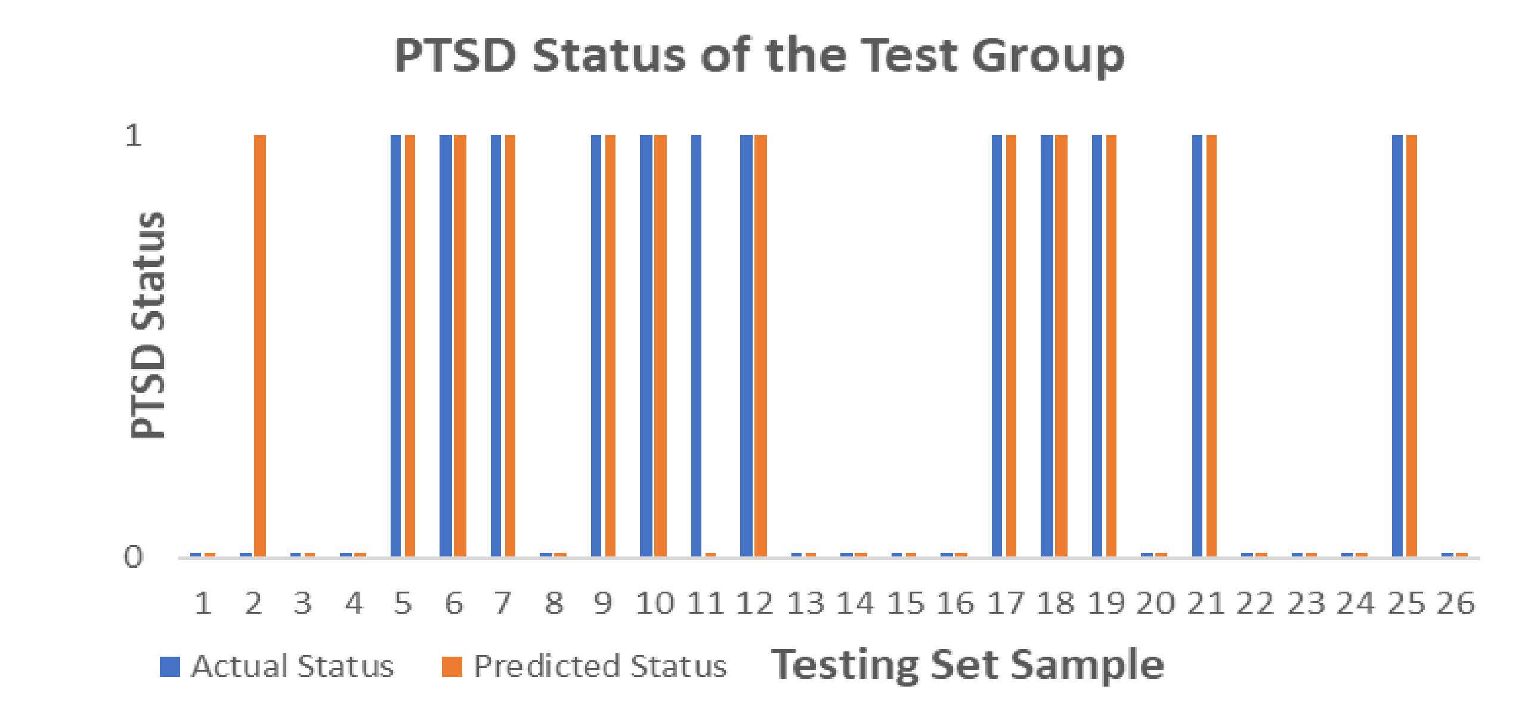


Figure 7: The actual vs. predicted PTSD status of the test group by the SVC model

CONCLUSIONS

143 significant methylation sites were discovered that have a high chance of being associated with PTSD. This list of significant CpG sites and the genes they are associated with demonstrat a correlation between these specific areas and the disorder itself, and as such can be used as a biomarker for diagnosis in individuals.

Through this list of sites, multiple machine learning models were built that could accurately predict the disorder. The prediction ability of these machine learning models with a high accuracy supports the idea that DNA methylation plays a big role in the development and existence of PTSD.

In all, as supported by the research and findings, it can be reasonably concluded that the hypothesis of using DNA methylation as a biomarker to diagnose PTSD was indeed supported. A PTSD diagnostic model using DNA methylation and machine learning was successfully built in hopes of assisting physicians in their medical decisions.

FUTURE STUDIES

How is the biochemical process of DNA methylation triggered?

What other biomarkers can be used to diagnose PTSD?