Directed Network Analysis of Genome-Wide Data for Post Traumatic Stress Disorder and Fibromyalgia Diagnosis

Undergraduate Senior Thesis

Directed Network Analysis of Genome-Wide Data for Post Traumatic Stress Disorder and Fibromyalgia Diagnosis

---

Magaly Gutierrez

Adviser: Dean Francis J. Doyle III

Harvard University
John A. Paulson School of Engineering and Applied Sciences
December 5th, 2018

Writing period
06. 2017 – 12. 2018

Adviser
Dean Francis J. Doyle III

Reviewers
Ms. Kelsey Dean
Dr. Burook Misganaw

Readers
Prof. Yaron Singer
Prof. Finale Doshi-Velez

Abstract

Network-based computational analysis of genome-wide data has been used to aid prediction of disease status with success in the past. Machine learning classifiers are usually applied to datasets of gene expression and/or methylation in order to classify a patient as diseased or healthy. Nevertheless, the large number of genes usually contained in these datasets makes it difficult to effectively identify genes with high relevance to the disease condition. The integration of information on interactions between genes, proteins, and other biochemical entities from databases of known biological pathways has been used to improve the analysis of abnormalities in gene expression and methylation with encouraging results [1, 2, 3]. Chuang et al. explored the effects that using combined subsets of measurements in known biological pathways has in identifying biological markers for disease classification [4]. The idea relies on the concept that biologically linked components of a network tend to have similar measurements (expression or methylation) when they are close to one another. In this project, the Chuang et al. algorithm is extended in order to exploit the effect of directionality on networks on identifying disease status [4]. This extended algorithm is applied to datasets on two diseases: Fibromyalgia and Post Traumatic Stress Disorder (PTSD). The results of the algorithm are then used as features in nine different classification algorithms. The results show that it is possible to achieve an accuracy rate similar, and in some cases better, than using individual methylation and expression measurements of genes as features with much less computational power. Furthermore, the use of this technique yields insight into the relationship between these often comorbid disorders.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

Medical conditions in which patients exhibit physical symptoms without clearly understood biological basis are referred to as psychosomatic disorders. The common comorbidity of these conditions and psychological symptoms has led doctors to classify them as mental health disorders often without an explanation or treatment for the physical complaints of patients [5, 6, 7]. As a result, psychosomatic disorders have been the target of lots of scrutiny over the years, from claims that they are not real conditions to being attributed to paranormal activity [5, 6].

Furthermore, the lack of understanding involving these conditions has led to millions of people all over the world to being diagnosed (and often misdiagnosed) with a psychosomatic disorder and then being left with no effective treatment. Shortage of explanations for symptoms prompt patients to continuously look for medical treatment resulting in exorbitant medical bills and often disability [8].

Motivated by this ongoing public health threat and new developments in the use of gene expression analysis to detect differentially expressed genes in diseased patients, I have chosen to focus on analysis of the biological basis for two psychosomatic illnesses: Fibromyalgia and Post Traumatic Stress Disorder (PTSD). In this project, I identified the genes with significant expression or methylation changes in patients suffering from these conditions and tested effectiveness of using these subsets of genes in classification of disease status of patients with different machine learning algorithms.

1.1 Overview of Psychosomatic Disorders

Poor understanding of the pathology of psychosomatic disorders has made it difficult to assert whether different conditions classified as psychosomatic are pathologically similar. Nevertheless, the motivation of this project to study PTSD and fibromyalgia syndrome relies on the reported comorbidity of the two conditions [9]. Some studies have suggested that there is as much as 57% prevalence of PTSD symptoms in fibromyalgia patients [10]. Moreover the onset of fibromyalgia syndrome has been linked to high-stress or traumatic life event making the correlation between PTSD and fibromyalgia more likely [10].

1.1.1 Fibromyalgia

Fibromyalgia is a psychosomatic disorder identified by symptoms of widespread pain, cognitive difficulties, and sleep disturbances [11]. As much as 2% of the US population is reported to suffer from this syndrome [9]. However, as of today, diagnosis of fibromyalgia is purely based on clinical symptoms and on ruling out other conditions with similar signs [12]. Nevertheless, the increase in disease rate among families suggests a genetic basis for the condition. Bondy et al. suggested abnormalities in the T102C polymorphism of the 5-HT2A-Receptor gene to be correlated with the disease [13]. On the other hand, Jones et al. found several upregulated genes in patients with fibromyalgia that could account for

the different symptoms of the condition [14]. There is still, however, not a conclusive gene identified as the sole cause of the disorder.

1.1.2 Post Traumatic Stress Disorder

Post Traumatic Stress Disorder (PTSD), is the name given to an array of symptoms resulting from an extremely stressful event. Such symptoms include reliving the event, nightmares, flashbacks, unwanted recollections of the event, irritability, insomnia, and impaired concentration among others [15]. PTSD is estimated to be the fourth most common psychiatric disorder in the United States but just like with fibromyalgia, there is no biological test for diagnosis [15]. Instead, a questionnaire is clinically administered to the patient believed to be suffering from this disorder and then evaluated by a physician [15]. Nevertheless, biological basis for PTSD disorder have also been suggested. For example, an elevated level of plasma norepinephrine and 3-methoxy-4-hydroxyphenylglycol [16] have been found in patients with PTSD as well as abnormalities in around 448 different genes [17].

1.2 Genome-Wide Analysis for Understanding Disease Status

Genome-wide expression profiles have traditionally been used to identify disease markers by measuring which specific genes are differentially expressed in diseased patients compared to healthy controls [4, 18]. Indicators of disease status have been previously identified by measuring the accuracy with which the gene expression measurement patterns are able to distinguish between patients and controls [19].

For some conditions, however, it hard to classify a sample as diseased or healthy because the small number of samples in a dataset compared to the number of features from genome-wide measurements can suggest incorrect differentially expressed genes. This may be the case in many studied datasets of PTSD and fibromyalgia. Studies in both of these conditions have shown differentially expressed genes that distinguish healthy versus diseased individuals [14, 17]. However because of the large number of suggested genes showing abnormalities in these conditions it is hard to reliably classify a patient as healthy or diseased leading to high rate of false positives. For this reason, it is necessary to employ algorithms to filter out the irrelevant features in the dataset and thus improve the prediction accuracy of the disease status.

1.2.1 The Role of Machine Learning in Disease Status Prediction

In machine learning, there are several approaches to improve prediction accuracy for a dataset with large numbers of irrelevant features [20, 21]. Improving the classification accuracy of a sample given a set of data is usually viewed as a two step problem: feature selection and classification algorithm selection. First, it is imperative to select the most relevant features for the samples we are trying to classify. There has been many proposed algorithms for eliminating irrelevant features from a dataset [20]. This project however, focuses on generating new features from the given dataset in order to 1) reduce the dimensionality of the data and 2) amplify the signal difference between samples belonging to different classes [21].

Reduction of the number of irrelevant features in genome-wide datasets is a common problem. In this case, I chose to generate new features given additional biological information in the form of signaling

networks. I have implemented feature generation algorithm based on previous work by Chuang et al. that combines the value of a node in the network with its neighboring nodes [4]. This approach causes clusters of nodes with large value differences that are then used as features and fed to classification algorithms.

Choosing the correct classification algorithm is also an imperative step for improving the prediction accuracy rate for a dataset. Different classification algorithms weight the different features with which we are trying to classify the samples differently meaning that two different algorithms can yield very different results even on the same set of data [22]. In this project I tested nine different classification algorithms with a 10-fold cross-validation protocol in order to investigate the differences in prediction accuracy rates among them. Furthermore, analysis of the implementation of each of these algorithms helps improve our understanding of the characteristics of the dataset used.

2 Background

2.1 Network-Based Approaches for Disease Status Classification

For diseases with large levels of noise for which it is hard to distinguish between diseased and healthy patients, it has been proposed that finding sets of genes with biologically verified interactions with one another that are differentially expressed could help expand the difference between diseased and healthy individuals allowing for more accurate disease status prediction [4, 23, 24].

Several approaches for integrating biological pathway network information and genome-wide methylation or expression measurements have been proposed. Vaske et al. introduced the Pathway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM) which integrates information on different types of molecular data and uses a probabilistic algorithm to generate pathway activities based the biological pathway networks [25]. On the other hand, the Tie Diffusion through Interactive Events (TieDIE) approach proposed by Paull et al. perturbs nodes on biological networks based on known genetic mutations, then allows the signals to propagate through the pathway over several iterations to assess the potential significance of related genes on disease prediction [26]. COMBINER by Yang et al. also used biological pathways for prediction of disease status but did not account for interactions within the pathway networks and instead generated features based on the similarity of genome data within the pathway [23]. Mitra et al. and Barabasi et al. give through overviews of other methods that have integrated biological network knowledge in the prediction of disease status [24, 27].

Chuang et al. introduced a "Network-Based Classification of Breast Cancer Metastasis" in 2007 which analyzed differentially expressed genes within known pathways. They defined subnetworks as the subset of genes within a given distance in an specific pathway and searched for differentially expressed subnetworks in a set of pathways [4, 19]. Starting from a specific node n_i, Chuang's algorithm seeds a subnetwork at a single node and then adds a node to the subnetwork from the neighbors of the nodes in the current subnetwork. The search ends when no node generates an increase in the score of the subnetwork that is greater than an specified parameter r which represents the maximum possible distance between the seed node and any other node in the subnetwork [4].

In motivation of Chuang et al.'s approach, this project analyzes the effects a modified subnetwork detection algorithm has on the accuracy rate of prediction of disease status. The hope is that this can help us understand how different differentially expressed genes in a condition relate to one another given the directionality of the pathway interactions.

In the algorithm modified from Chuang et al., I define the interaction networks as directed while the original approach used undirected networks. Furthermore, in this approach there is not a minimum score increase value and thus the generation of a subnetwork stops when there is no additional neighbor of the current nodes in the subnetwork that can increase the score of it.

2.2 Overview of Datasets

2.2.1 Fibromyalgia Data

The fibromyalgia gene expression data used in this project was originally published by Jones et al. and can be found in Gene Expression Omnibus (GEO) Database with accession number GSE67311 [14]. The study was conducted on blood samples of Caucasian females of ages 18 and over. The fibromyalgia participants were clinically diagnosed with fibromyalgia for at least six months at the time of the study [14]. More information can be found in the original Jones et al. report [14]. This dataset uses the Affymetrix Human Gene Array platform to quantify expression of 33297 genes.

2.2.2 PTSD Methylation Data

The PTSD Methylation dataset used in this project was first published by Hammamieh et al. and it is published in the GEO Database with accession numbers GSE76401 and GSE85399 [28]. The DNA methylation data was derived from the blood samples of male Iraq and Afghanistan war veterans. The participants ranged from 20 to 60 years of age. The PTSD participants had been clinically diagnosed for at least three months [28]. More information about the experimental protocol and prepossessing steps of the samples can be found in the original publication.

2.2.3 CePa Pathway R Package

The CePa pathways database contains a total of 1004 biologically verified individual pathways with differing numbers of genes, complexes, and compounds from the NCI, BioCarta, KEGG, and Reactome catalogs [29]. The pathways vary in number of nodes and connections. An example of these pathways is the `Fanconi Pathway` shown in Figure 1. In this representation the size of the nodes is relative to their respective t-score. The edges between the nodes represent interactions between genes, proteins and components [29].

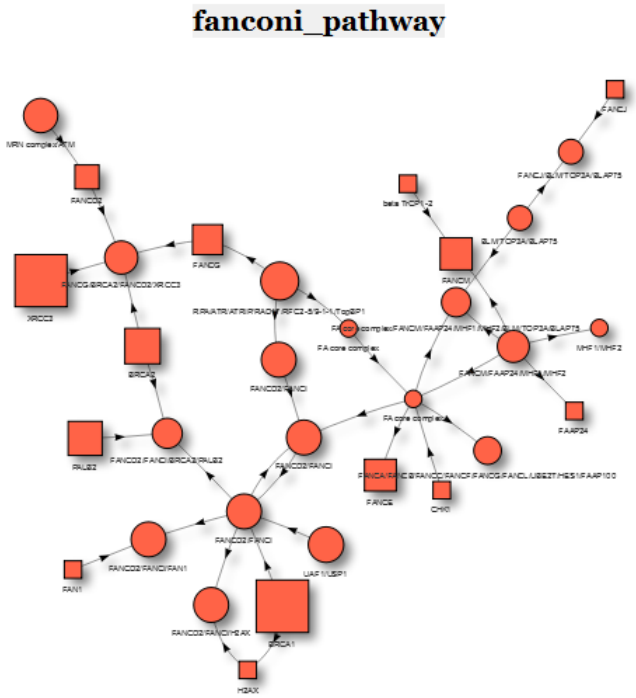

Figure 1: Fanconi Pathway. Key interactions pathway in the repair of damaged DNA. Mutations in this pathway are linked to Fanconi anemia, a type of anemia that generates bone marrow failure, congenital growth abnormalities and cancer predisposition [30]. In this representation circles represent complexes, triangles represent genes and squares represent other components.

3 Approach

Every sample of methylation or genome-wide expression data contains thousands of measurements. Some of these measurements are irrelevant to the condition we are trying to predict. Unrelated measurements add a large degree of noise to the system, making it difficult for classification algorithms to distinguish between diseased and healthy individuals. Thus, before trying to predict the status of a patient from expression or methylation measurements using various classification algorithms, I filtered out irrelevant features from the system.

In this approach, irrelevant measurements are filtered out in two steps. First, features are generated using a modification of the algorithm used in Chuang et al., grouping together subsets of genes in biologically validated pathways that are relevant to disease status [4]. Section 3.1 contains the description of the algorithm used and the definitions of the terms in the results. In section 3.2, the subcore results are passed through nine classification algorithms to measure the prediction rate for each dataset. Then in section 3.3, the values of these subsets of genes is quantified and ranked using a feature elimination algorithm and prediction rate was measured on the reduced set of features.

3.1 Feature Generation

Each node in a given CePa pathway is composed of either one gene, a group of genes, or a complex group of genes and proteins. First, the t-score for each node in a given pathway is calculated using R function `prcomp` [31]. Two different approaches were used for the cases in which a pathway node is a complex node (composed of more than one gene or a combination of genes and proteins). First, a node's t-score was given the t-score of its first principal component after doing a principal component analysis of the node using the `prcomp` function. This function performs singular value decomposition (SVD) which according to R documentation is preferred due to its numerical accuracy [31]. This is referred to as the PCA approach throughout this paper. In the second approach, the mean over all of the components in the node was calculated. This approach is referred to throughout the paper as Mean Value Analysis (MVA) approach.

Algorithm 1 is used to identify statistically significant subcores in a pathway to be used as features in classification algorithms. A subcore is defined as the set of nodes that has a combined statistical significance. The term core is used to refer to a group of connected nodes from different subcores - that is, the group of nodes belonging to statistically significant subcores. Algorithm 1 is started from each individual n_i node in the pathway. The neighbors of node n_i are defined as all of the nodes with an incoming edge from n_i. While there are still neighbors available to expand the subcore, the algorithm checks for the neighbor with the highest t-score. It re-calculates the subcore's t-score by averaging the absolute value of the expression/methylation data of all of the components of the subcore with the

Figure 2: Core Generation Algorithm Summary. 1) Identify the nodes for which the data set does not have available information (grey). 2) Begin with the BRCA1 node (this image just shows a close up of the previous picture). 3) Color BRCA1 as part of the first subcore (yellow), add its neighbors to the neighbor pool in descending t-score order. 4) Add the node with the highest t-score to the subcore if it increases the overall t-score of the subcore. 5) The neighbor's pool now consists of the neighbors of all of the subcore elements that have not yet been tested (orange). 6) Continue attempting to add nodes to the algorithm until the neighbor pool is empty and we have no more neighbors to add. 7) The process will be then repeated starting for each node of the pathway. 8) As it is visible in this picture nodes that were not the highest t-score when initially added can still end in the subcore. 9). All cores in the pathway with a minimum t-score of 1.

new neighbor and makes the decision of adding the neighbor to the subcore if the absolute value of the t-score increases or rejecting the neighbor if the absolute value of the t-score decreases or stays the same (see Figure 2). If information is missing from one of the nodes in the pathway, this node is added to the subcore and its neighbors are added to the neighbor pool.

The algorithm returns the subset of subcores that surpass a specific t-score or that are below a given p-value. The list of all subcores in each pathway is filtered and only the subcores with a t-score greater than 1.978, corresponding to a p-value of 0.05, are kept[1].

[1]Because the number of samples, or degrees of freedom in both of the datasets tested is similar (142 samples for fibromyalgia and 161 samples for PTSD) the same cutoff t-score was used. This t-score corresponds to a p-value of 0.05 with 141 degrees of freedom.

3.2 Training Prediction Models

Nine classification algorithms for the prediction of disease status given the features generated in the previous step were chosen and compared. The package scikit-learn in python contains functions that implements and optimizes each of these approaches and in which one can modify parameters [32, 33]. This package was used for simplicity. Table 1 contains a summary of the prediction models and the parameters used in each of them while the list below contains a brief description of the implementation parameters used for such algorithms:

- **K-Neighbors Classifier**
  In this algorithm a data sample is assigned the classification value of the majority of its `k` neighbors. In this particular implementation by sci-kit learn the training data is kept completely and each sample of the test data is assigned the classification value of the majority of the `k` training samples that are closest to it in distance. The distance between neighbors is the Minkowski distance between all of the `N` feature components used in each specific trial. Using p=2 defines the distance as the Euclidean distance between the features of each sample:

$$distance(D_0, D_1) = (\sum_{i=1}^{N}(|D_{0i} - D_{1i}|^p))^{(1/p)},$$

  where D_{0i} and D_{1i} represent the value of the ith component of D_0 and D_1 respectively. Here, I used k=5 to assign the sample to the majority of its five closest neighbors.

  Although the K-Neighbors classifier is simple, it has been used on genome data with positive results in the past motivating its use in this project [34, 35].

- **Linear Support Vector Machine (SVM)**
  Support Vector Machine (SVM) algorithms define feature spaces in large dimensions so that the Euclidean distance between points in the training data belonging to different classes is maximized [36, 37]. The margin of the SVM is defined as the perpendicular distance between the decision boundary and the closest sample for each of the classes in the selected feature space [37, 38].

  New samples are projected onto the same feature space and classified according to the equation:

$$\hat{y}(x) = w^T\phi(x) + b,$$

  where the sign of the result $\hat{y}(x)$ represents the classification class of the sample x. The parameter b is the bias or intercept, w is the weight vector of the dataset and $\phi(x)$ is a fixed feature space transformation [37, 39]. In the linear kernel example $\phi(x) = x$.

  One of the advantages of SVM algorithms is that local solutions are equivalent to global optimum simplifying the process of classifying a sample [37, 38, 40, 41].

- **Radial Basis Function (RBF) SVM**

The kernel used in this trial is defined as:

$$kernel(x, x') = \exp\left(-\frac{1}{2}d(x, x')^2 * \gamma\right),$$

where d(x,x') is the Euclidean distance between the two samples and the parameter γ is a measure of how close samples need to be to one another to have their outcome classification affected [32, 36].

- **Gaussian Process**
  The Gaussian Process classifier assigns a new sample probabilistically based on the similarity between the new sample and samples in the training data given the specified kernel function [42]. In this trial the Gaussian Process classification algorithm was used with the RBF kernel:

$$kernel(x, x') = \exp\left(-\frac{1}{2}d(x, x')^2\right),$$

  where d(x,x') is the Euclidean distance between the two samples.

- **Decision Tree**
  The Decision Tree algorithm generates models based on the training data in the form of trees. The tree represents a set of rules that end up in the classification of a new sample. The rules are represented as branches in the tree and they culminate in a node that contains the classification result of the given sample [43, 37].

- **Random Forrest**
  The Random Forest algorithm generates several decision trees with bootstrapped samples and classifies a new sample based on the majority of the results of the different Decision Trees it generated [44].

- **AdaBoost**
  The original AdaBoost algorithm combines the results from different weak classifiers (classifiers only required to achieve 50% accuracy) and returns the majority vote of the result of those classifiers [45]. The sci-kit learn algorithm implements a modified version of the original AdaBoost algorithm which uses Stagewise Additive Modeling using a Multi-class Exponential Real loss function (SAMME.R) proposed by Zhu et al. which among other differences requires the weak classifiers to perform better than 50% accuracy [46, 47].

- **Gaussian Naive Bayes**
  The Naive Bayes algorithm assumes independence between each pair of features of the dataset. It classifies a sample according to the equation:

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x \mid y),$$

  where $P(y)$ is the prior probability that the sample is of the class y and $P(x|y)$ refers to the conditional probability of the sample x given it belongs to the class y. In this case the Gaussian Naive Bayes classifier was used. This instance of the Bayesian classifier assumes a Gaussian

distribution on the likelihood of the features of each sample. Thus the probability of a sample x of being of class y is given by the value:

$$P(x \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right),$$

where σ_y^2 is the variance of the class and μ_y is the mean [48].

- **Quadratic Discriminant Analysis (QDA) Classifier**
  The Quadratic Discriminant Analysis (QDA) Classifier uses Bayes' Theorem to classify a sample based on the maximum posterior probability of the sample being in an specific class. The probability of a sample x given that it is of class y=k is given by the equation:

$$P(x|y=k) = \frac{1}{(2\pi)^{N/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right),$$

  where `N` represents the number of features in the dataset, `x` is the sample, μ_k is the mean value for the class `k`, and Σ_k is the covariance matrix of the samples.

  The QDA algorithm was used with no shrinkage in this case meaning that the diagonal matrices were not used as estimates for the covariance matrices or averaged with them. This is important because when the opposite is the case (when the diagonal matrices are used instead of the covariance matrices) this model is equivalent to the Gaussian Naive Bayes classifier as it assumes conditional independence in each class.

  The main limitation about using this algorithm came from the fact that some features in most biological datasets are expected to be correlated. This means that the determinant of the covariance matrices for calculating this classification is close to zero making matrix inversion inaccurate and thus the individual measure of importance for each feature not reliable [49]. If the features are indeed correlated, then the effect of some features will be undermined by the algorithm but more importantly this can affect the overall classification rate of the algorithm as colinear features will be eliminated or not considered for the classification of a sample [49]. The idea behind this choice was to contrast the Gaussian Naive Bayes approach which assumes that features are independent. Using this algorithm will help cement the assumed correlation between features in the dataset and draw insight about the importance of features not clearly correlated to one another.

3.3 Feature Reduction Algorithms

In addition to classifying the samples using all of the subcores found in the dataset, a recursive feature reduction algorithm was applied. The scikit-learn package in python was used with a step size of one (meaning that the 10-fold average accuracy was calculated after every one feature was removed) and a support vector classification model (SVC) with a linear kernel [32, 33].

The purpose of using the scikit-learn feature reduction algorithm is to further decrease the number of features being used to predict disease status [50]. The algorithm calculates the optimal number of

Table 1: Prediction Models. Parameters used in the respective prediction models.

Prediction Model	Parameters
KNeighbors Classifier	algorithm='auto', leaf_size=30, metric='minkowski', n_neighbors=5, p=2, weights='uniform'
Linear SVM	C=1, decision_function_shape='ovr', degree=3, gamma='auto', kernel='linear'
RBF SVM	C=1, decision_function_shape='ovr', degree=3, gamma=2, kernel='rbf', max_iter=-1, probability=False, tol=0.001
Gaussian Process	kernel=1**2 * RBF(length_scale=1), max_iter_predict=100, optimizer='fmin_l_bfgs_b'
Decision Tree	criterion='gini', max_depth=5, min_samples_leaf=1, min_samples_split=2, random_state=None, splitter='best'
Random Forest	bootstrap=True, criterion='gini', max_depth=4, max_features=1, max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, n_estimators=10,
AdaBoost	algorithm='SAMME.R', base_estimator=None, learning_rate=1.0, n_estimators=50, random_state=None
Gaussian Naive Bayes	priors=None
QDA	priors=None, reg_param=0.0, store_covariance=False, store_covariances=None, tol=0.0001

features the data needs for best prediction accuracy.

Algorithm 1 Feature Generation Algorithm

SubcoreList $\leftarrow 0$ ▷ Initialize list of subcores
foreach node **v** in pathway **P do**
  Calculate t-score of node **v**
  SC $\leftarrow$ **v** ▷ Begin a subcore (SC) with node v
  N $\leftarrow$ **v** neighbors ▷ Get all neighbors of node v
  while N is not empty **do**
    u $\leftarrow$ node in **N** with highest t-score.
    Visited $\leftarrow$ **Visited** $\cup$ **u** ▷ Create a vector of visited nodes
    remove **u** from **N**
    if $t - score(v + u) \geq t - score(v)$ **then**
      SC $\leftarrow$ **SC** $\cup$ **u** ▷ Add neighbor to ongoing core
      N $\leftarrow$ **N** $\cup$ **u** neighbors $\notin$ **Visited** ▷ Add neighbors of u that we have not visited to N
    end if
  end while
  SubcoreList $\leftarrow$ **SubcoreList** $\cup$ **SC**
end forreturn SubcoreList

4 Results

4.0.1 Feature Generation

Two datasets were processed according to Algorithm 1 [14, 28]. Table 2 contains an overview of both of the datasets.

Table 2: Summary of Datasets. Overall dimensions of the two datasets tested in this project.

Dataset	Controls	Cases	Sites Measured	GEO Accession
Fibromyalgia gene expression	71	71	33297	GSE67311
PTSD Methylation	80	81	485577	GSE76401, GSE85399

The Fibromyalgia gene expression dataset consists of a matrix of expression data for 33297 sites for 71 healthy subjects and 71 patients with fibromyalgia. Algorithm 1 returned 898 significant subcores (with a p-value of less than 0.05) when complex nodes were processed as an average of their components, referred to as the mean value analysis (MVA) approach. These subcores belonged to 276 different pathways. When complex nodes were processed as their first principal component (PCA approach), Algorithm 1 returned 786 significant subcores belonging to 258 different pathways.

The PTSD dataset consists of a matrix of 485577 measurement sites, 80 control patients and 81 patients with PTSD. Algorithm 1 returned 784 significant subcores belonging to 235 different pathways when complex nodes were processed as averages of their components (MVA). Finally the algorithm showed 136 significant subcores in this dataset belonging to 67 different pathways when complex nodes were processed as their first principal component (PCA).

Table 3 contains a summary of the results obtained from Algorithm 1 for both datasets. It is worth noting that in both datasets the number of subcores found by the algorithm, as well as the number of pathways the subcores belong to is similar for both conditions. The exception to this rule is the PTSD DNA Methylation PCA results which had a surprisingly low number of subcores compared to the other trials. This difference is likely due to the deviation of measurements in the components of the complex nodes. When the first principal component is used as a measurement of a complex node, it is less likely that adding another node to the subcore will increase the t-score of the subcore resulting in smaller subcores and fewer overall significant subcores.

Table 3: Subcore Feature Generation Algorithm Results.

Dataset	All Genes		PCA		MVA	
	Sites	Pathways	Subcores	Pathways	Subcores	Pathways
Fibromyalgia Gene Expression	33297	1004	786	258	898	276
PTSD DNA Methylation	485577	1004	136	67	739	235

4.0.2 Disease Status Prediction

The tables of features produced by Algorithm 1 in the previous section were used as inputs for nine different classification algorithms from the scikit-learn python package [32, 33]. The prediction accuracy rate and processing time for each model and dataset is summarized in the next subsections.

4.0.3 Fibromyalgia Dataset Prediction Results

The prediction accuracy and processing time for each prediction algorithm with the fibromyalgia subcores is summarized in Table 4. The information is also organized in bar plots in Figure 3. The majority of the classification algorithms performed better when using the subcores as features as opposed to using the entire dataset of gene measurements. In three of the algorithms (RBF SVM, Gaussian Process and QDA) the results from all of the genes were better or equal to one of the subcore datasets but not both. In the RBF SVM and Gaussian process approaches, all genes are better than MVA but not better than the PCA prediction scores, in the QDA classification the score of all genes matched the score of the score of the MVA subcore dataset and outperformed the score of PCA prediction. Overall, both PCA and MVA approaches showed increased accuracy averages among all classification algorithms compared to the results of the dataset with all genes.

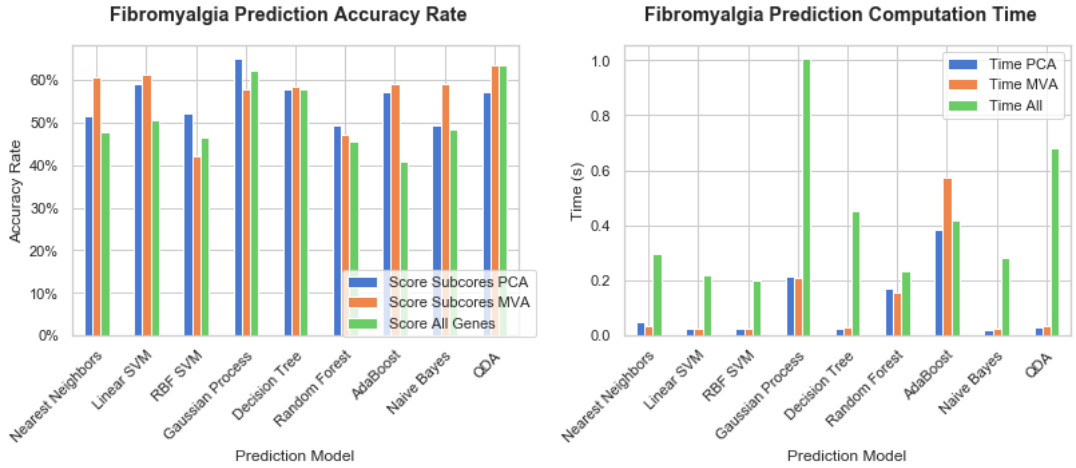

Figure 3: Prediction Results for Fibromyalgia Dataset.
Left: Prediction accuracy rate for fibromyalgia dataset.
Right: Processing time for fibromyalgia dataset per prediction model.

The amount of time it took to train the classification algorithms and classify the samples was in general higher for the dataset with more genes. This result, however, should be taken with a grain of salt because it does not account for the amount of time that it takes for Algorithm 1 to generate the subcores.

A recursive feature reduction algorithm from the scikit-learn package in python was used with a step size of one (meaning that the 10-fold average accuracy was calculated after one feature was removed) and a support vector classification model (SVC) with a linear kernel [32, 33]. The accuracy rate per number of subcores used for the prediction in both approaches is shown in Figure 4. The dataset resulting from the PCA processing had a slightly lower prediction rate than the MVA processed one (70.48% as opposed to 71.87%). The highest prediction accuracy score in PCA however, happened

Table 4: Prediction Results for Fibromyalgia Dataset.

Prediction Model	Subcores PCA		Subcores MVA		All Genes	
	Accuracy	Time(s)	Accuracy	Time(s)	Accuracy	Time(s)
Nearest Neighbors	0.514	0.050	0.607	0.036	0.478	0.297
Linear SVM	0.592	0.026	0.614	0.026	0.507	0.219
RBF SVM	0.521	0.023	0.421	0.023	0.464	0.201
Gaussian Process	0.650	0.215	0.578	0.209	0.621	1.006
Decision Tree	0.578	0.024	0.585	0.029	0.578	0.451
Random Forest	0.492	0.171	0.471	0.155	0.457	0.234
AdaBoost	0.571	0.384	0.592	0.574	0.407	0.420
Naive Bayes	0.492	0.021	0.592	0.023	0.485	0.283
QDA	0.571	0.030	0.635	0.036	0.635	0.681
Average:	0.553	0.104	0.566	0.123	0.514	0.421

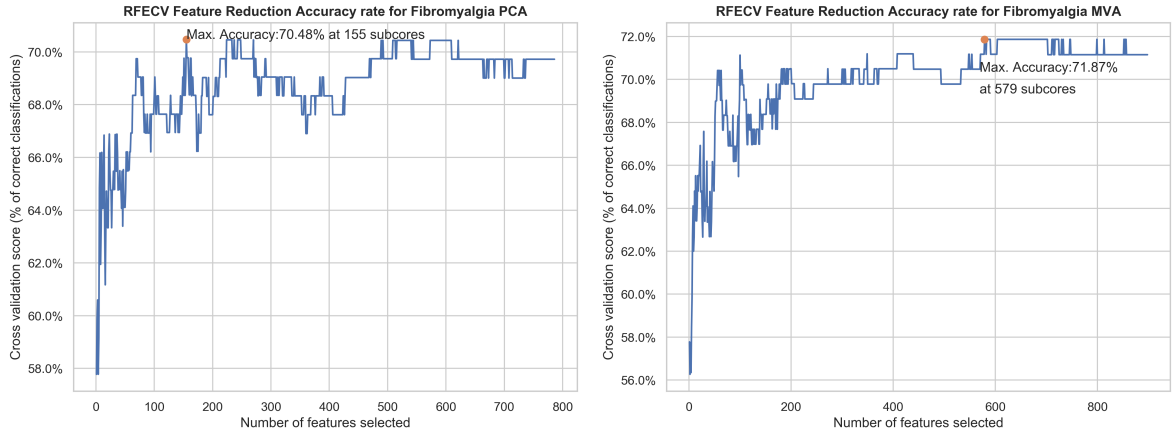

Figure 4: Accuracy Rates for the SVC Model with Linear Kernel: Fibromyalgia.
Left: Fibromyalgia PCA Feature Reduction Plot. The plot shows the different accuracy rates depending on the number of features used in the prediction (x-axis). The maximum accuracy was given by 155 features with 70.48%.
Right: Fibromyalgia MVA Feature Reduction Plot. The plot shows the different accuracy rates depending on the number of features used in the prediction (x-axis). The maximum accuracy was given by 579 features with 71.87%.

with a much smaller number of subcores than the MVA processed dataset (155 as opposed to 579 subcores). Both approaches were able to improve their prediction accuracy rate by more than 10% from the reported accuracy rate with all subcores using this feature elimination method. The PCA approach improved from 59.20% accuracy rate to 70.48% and the MVA approach went from 61.40% to 71.87%. This results seems to support that there is still some irrelevant features among the subcore datasets.

Figure 5 contains a list of the pathways with the highest number of subcores from the resulting highest prediction accuracy rate on the SVC model for both of the PCA and MVA approaches. In the MVA approach, the 579 subcores that yielded the highest prediction accuracy rate (71.15%) belonged to a total of 271 pathways whereas the 155 subcores that yielded the highest prediction rate in PCA approach belonged to 110 different pathways.

A comparison of the pathways with significant subcores in both approaches showed that 108 pathways were common between both approaches though with differing number of subcores in each

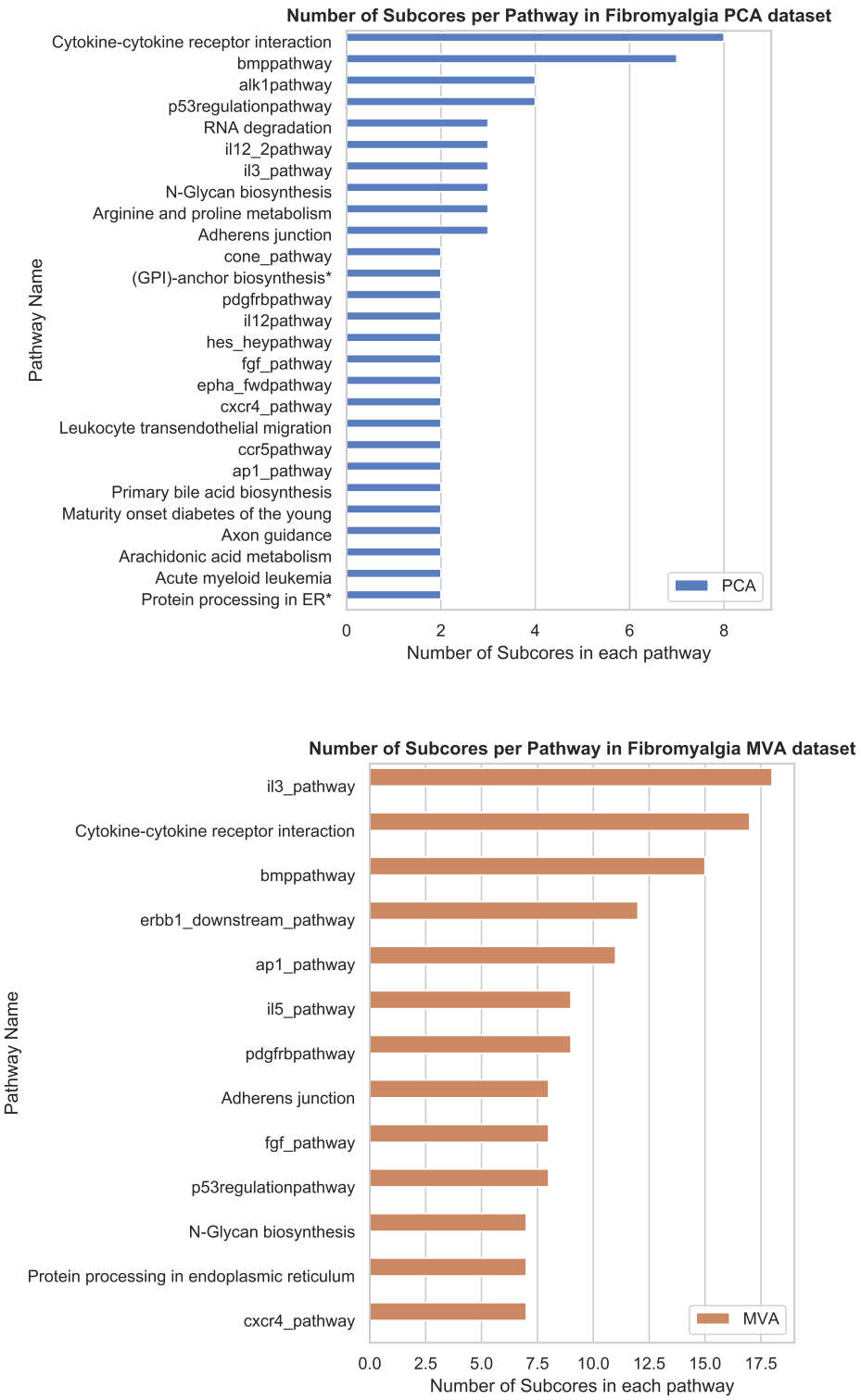

Figure 5: Significant Subcores per Pathway: Fibromyalgia.
Top: Fibromyalgia PCA Pathway Names vs. Number of Subcores. Pathways with two or more subcores in them. Names marked with an asterisk have been shortened from their original CePa name.
Bottom: Fibromyalgia MVA Pathway Names vs. Number of Subcores. Pathways with more than 6 subcores.

pathway. The common pathways with more than three subcores are shown in Figure 6, including the interlukin-3 (il-3) pathway, the Cytokine-cytokine receptor interaction pathway, and the Bone Morphogenic Protein pathway.

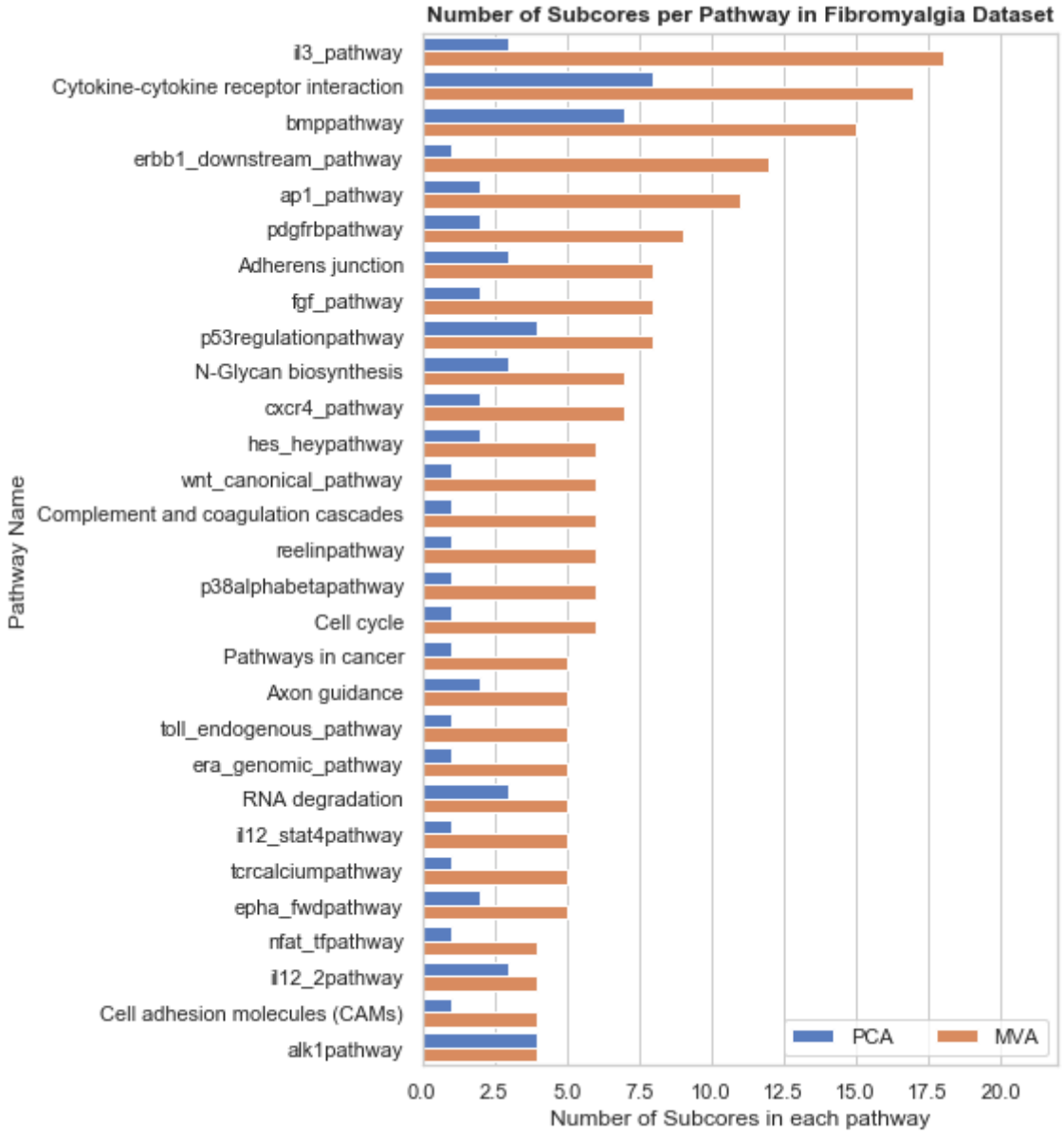

Figure 6: Fibromyalgia Common Pathways between PCS and MVA approaches. Pathways with >2 significant subcores either in the PCA or MVA approach.

4.0.4 PTSD Methylation Prediction Results

The prediction accuracy and processing time for each prediction algorithm with the PTSD Methylation subsets is summarized in Table 5. The information is also organized in bar plots in Figure 7. As with the fibromyalgia dataset, we can see that the classification algorithms yielded similar accuracy rates for the PCA and MVA subcores as with the entirety of the genes in the dataset. The average accuracy rate of prediction for the MVA approach was slightly higher than the average for the dataset with all genes (54.2% to 53.2%) while the average of the PCA approach fell slightly lower (52.7%)

Table 5: Prediction Results for PTSD Methylation Dataset.

Prediction Model	Subcores PCA		Subcores MVA		All Genes	
	Accuracy	Time(s)	Accuracy	Time(s)	Accuracy	Time(s)
Nearest Neighbors	0.487	0.026	0.531	0.026	0.500	7.986
Linear SVM	0.518	0.008	0.556	0.017	0.550	7.001
RBF SVM	0.506	0.009	0.487	0.020	0.562	7.442
Gaussian Process	0.531	0.299	0.575	0.306	0.531	17.024
Decision Tree	0.575	0.010	0.537	0.036	0.506	10.337
Random Forest	0.550	0.131	0.543	0.133	0.512	4.444
AdaBoost	0.493	0.562	0.556	0.769	0.575	10.536
Naive Bayes	0.512	0.011	0.512	0.019	0.493	8.674
QDA	0.575	0.012	0.587	0.025	0.562	10.802
Average:	0.527	0.118	0.542	0.150	0.532	9.360

than both of them. The discrepancy, however, remained small. This is interesting for the PCA strategy in this dataset especially because the number of pathways with significant subcores was very small (67) in comparison to the MVA strategy with the methylation dataset (235) and to the results of both strategies in the fibromyalgia dataset (258 for PCA results and 276 for MVA results).

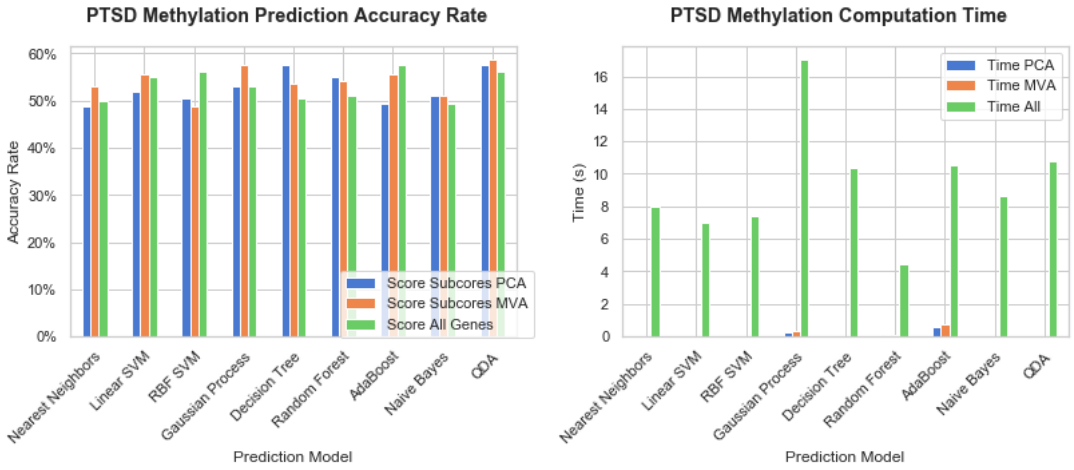

Figure 7: Prediction Statistics for PTSD Methylation Dataset.
Left: Prediction Accuracy Rate for PTSD Methylation dataset.
Right: Processing time for PTSD Methylation dataset per prediction model.

In the PTSD Methylation dataset especially, there are several readings for the gene. Thus, it can be argued that the PCA approach does a better job of singling out the node components with higher signals related to the disease. The MVA approach weights all components in a node equally, reducing the node t-score by including irrelevant components. This could imply that different methylation sites are not equally important in predicting disease condition.

The same feature reduction algorithm from the scikit-learn package in python that was used in the fibromyalgia dataset was used for this dataset with a step size of one (meaning that the 10-fold average accuracy was calculated after one feature was removed) and a support vector classification model (SVC) with a linear kernel [32, 33]. The accuracy rate per number of subcores used for the prediction in both approaches is shown in Figure 8. Once again, the feature reduction algorithm was able to improve the

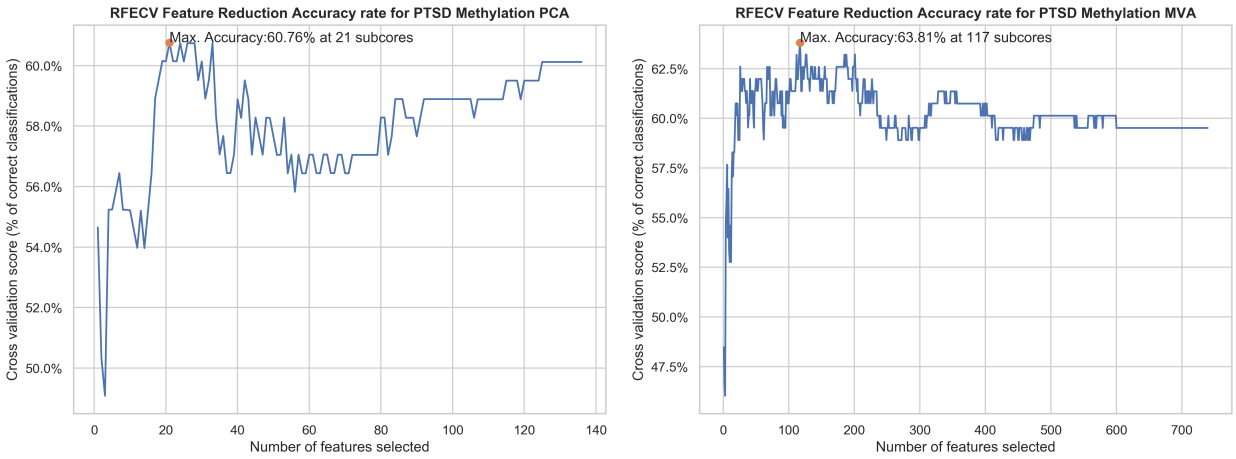

Figure 8: Accuracy rates for the SVC Model with Linear Kernel: PTSD Methylation. Left: PTSD Methylation PCA Feature Reduction Plot. The plot shows the different accuracy rates depending on the number of features used in the prediction (x-axis).The maximum accuracy was given by 21 features with 60.14% accuracy.
Right: PTSD Methylation MVA Feature Reduction Plot. The maximum accuracy was given by 117 features with 63.2% accuracy.

prediction accuracy score. In the PCA approach the score went from 51.8% using all subcores to 60.76% with only 21 subcores. On the other hand, the prediction accuracy score improved from 55.0% with all subcores to 63.2% with only 117 subcores. This suggests that there are some irrelevant subcores in these datasets as well.

In general, it can be noted that these accuracy values were lower than those in the fibromyalgia data. Although it is not initially clear what the reason for this difference is, one of the causes include the the large number of genes measurements contained in the dataset. As mentioned before, it appears that different methylation sites have different relevance to the disease status prediction. It is also worth noting, however, that the prediction rate with the entire dataset for the PTSD methylation dataset was initially lower than the fibromyalgia dataset.

Figure 9 contains a list of the pathways with the highest number of subcores in the results of the PCA and MVA approaches.

The comparison of the pathways with significant subcores in the PCA and MVA approaches was made. A total of 6 pathways were common between both approaches. All of those pathways except for the MAPK Signaling Pathway contained the same number of subcores in both approaches. These pathways are listed in Figure 10, including the glycine, serine, and threonine metabolism pathway, the MAPK signaling pathway, and the cell adhesion molecules (CAMs) pathway.

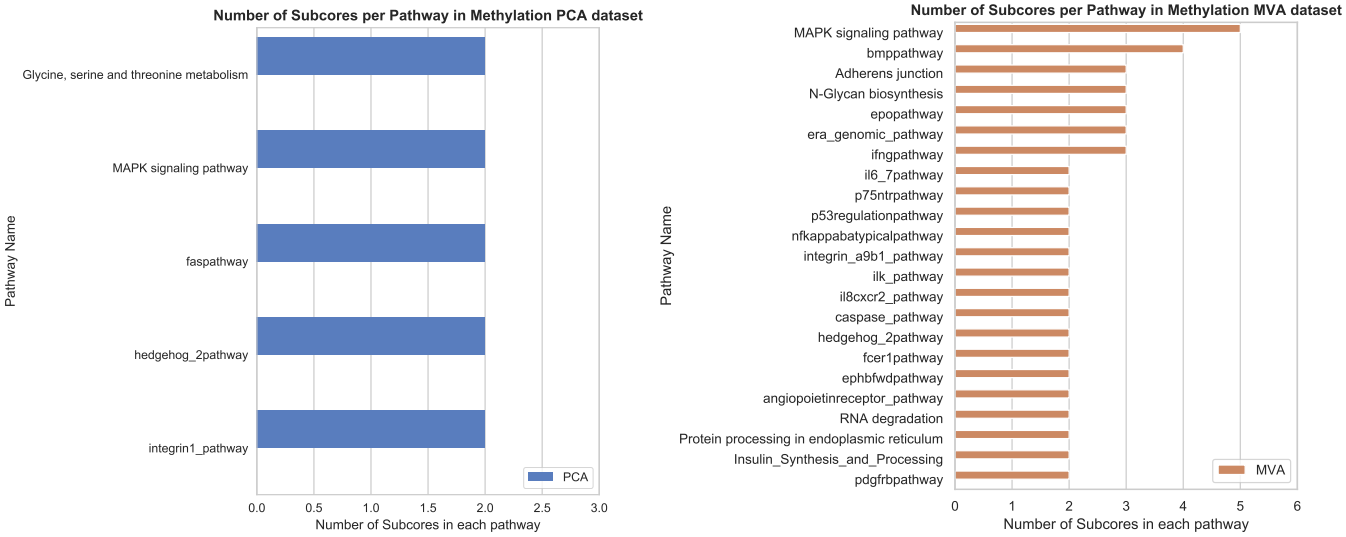

Figure 9: Significant Subcores per Pathway: PTSD Methylation.
Left: PTSD Methylation PCA Pathway Names vs. Number of Subcores. Pathways with two or more subcores.
Right: PTSD Methylation MVA Pathway Names vs. Number of Subcores. Pathways with more than 2 subcores.

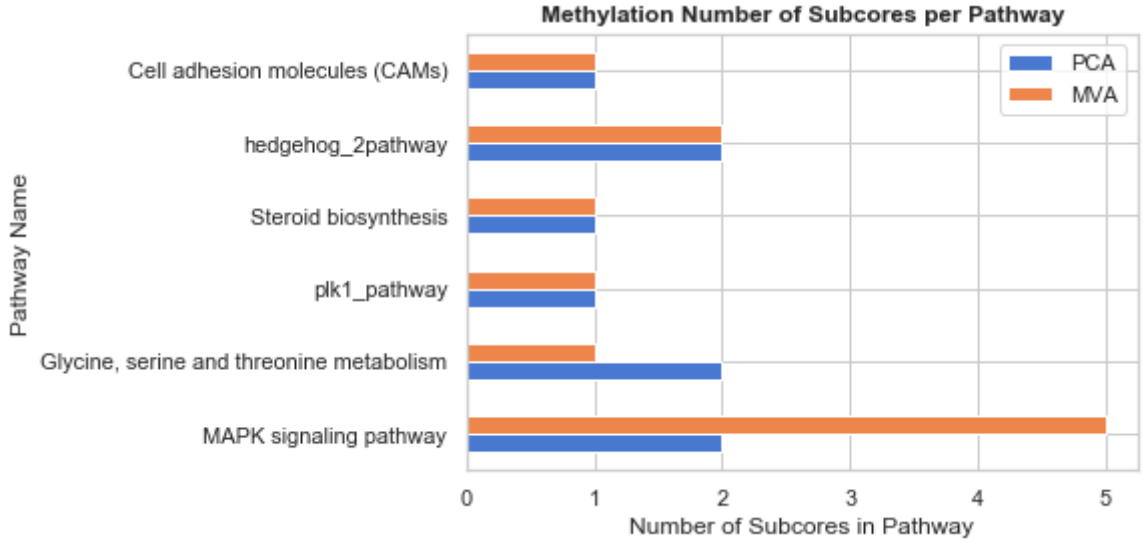

Figure 10: PTSD Common Pathways between PCA and MVA approaches. Pathways with ≥ 1 significant subcores either in the PCA or MVA approach.

5 Conclusion

5.1 Summary and Discussion

The results presented in the previous section support the idea that it is possible to predict whether a patient has PTSD or fibromyalgia from gene expression or DNA methylation data. As shown in the prediction accuracy tables for both conditions, it was possible to achieve a similar prediction accuracy rate to the one obtained using all of the expression and methylation samples in the datasets. The results also show that the method used in this project dos not provide sufficient improvement in the accuracy of the prediction of either condition. The maximum prediction accuracy rates remained around 70% in both conditions which could indicate that there is still a lot of irrelevant features in the subset of subcores returned by the algorithm.

This suggestion is supported by the small discrepancy between the accuracy rates of different classification algorithms used despite the fact that each algorithm weights features differently. For example, because the features in the dataset are not independent of each other, the Naive Bayes approach was expected to have a low accuracy rate. However, for both conditions the maximum achieved accuracy rate with other algorithms was not too different from Naive Bayes. Naive Bayes has been proven to result in high accuracy rate predictions for non-feature-independent datasets given that they cancel each other out or are distributively evenly [51]. In this case, it is necessary to assess the distribution of the features to have a more informed idea as to why the Naive Bayes approach matched the prediction rate of other classification algorithms. This could determine whether the feature sample was poor for all algorithms or the Naive Bayes approach performed extraordinarily well.

The accuracy rate results of the QDA algorithm were also expected to be low given the correlation between the features on the datasets. Based on the results of the analysis by Naes et al. that colinearity seems to decrease the overall classification accuracy rate of a dataset by cancelling out colinear features, it might seem that the prediction on this algorithm would rely on isolated subcores, those that do not have colinear counterparts or other subcores within the same pathway [49]. The sci-kit learn function indicated that there was colinearity between variables meaning that the algorithm could not clearly assess the importance of individual features on the dataset [49]. Colinearity was expected due to the nature of the generation of the subcores by Algorithm 1. Different subcores with a similar set of genes starting on different nodes within the same pathway are expected to have similar t-scores prompting the colinearity result. One way in which this colinearity can be avoided is to use cores instead of subcores as features. A core would be the connected set of subcores that have been found to be statistically significant.

Finally, despite the often reported comorbidity between PTSD and fibromyalgia, a biological link between the conditions is still unclear [52]. Some pathways such as the MAPK signaling pathway and the Bone Morphogenic Protein (BMP) pathway appear in the list of pathways with largest number of significant subcores in both conditions. However, because of the nature of these pathways and

their large number of functions associated to them, it is hard to pinpoint specific reasons as to why these pathways are prominent in both conditions [53, 54]. A better approach would be to compare the differentially expressed subcores between both conditions and pinpoint the specific role of such subcores in the function of the pathway.

5.2 Future Work

There is still a lot of room for research in the field of biological bases for psychosomatic diseases. Furthermore, the issue can be approached from many different points of view.

From the biological perspective it is imperative to continue to map interaction networks between genes, proteins and complexes with higher accuracy. In this case, the CePa package provides a robust set of biological pathways. But pathways do not exist in isolation and many of them are connected to each another. Improving the accuracy of pathway information can help us make more sense of the signals traveling through the pathways and the effect they have on the classification of a patient as healthy or diseased. By having knowledge of the attributes of the connected network of pathways we can draw conclusions about the ways in which disease signals travel.

A quick analysis on the characteristics of the integrated pathway network of CePa package shows that there is a large number of pathways with common nodes. The combined network of pathways contains 16995 nodes and 37714 edges between them and although there are a total of 216 weakly connected components, 16156 nodes are within one of such components which could allow us to map several pathways with differentially expressed subcores together so we can find the starting node of a differentially expressed signal.

It is more complicated to draw conclusions on the behavior of signals given the seemingly sparse overall network. The clustering coefficient of the largest weakly connected component on the integrated CePa network is only 0.08 indicating this is a very sparse network. This contrasts with the individual pathways which have been shown to display small-world phenomena [24]. However there is the possibility that the sparsity of the network is due to missing nodes that are not yet biologically corroborated. In such case, more biological advances in interaction pathways are necessary for better disease status predictions.

One possibility is use the directionality in the biological pathways to find the node that starts the differentially expressed signal. As mentioned above, many of the pathways in the CePa package have nodes in common but the genome datasets do not always contain information about every possible node in the pathways making it difficult to analyze the disease signal propagation within the integrated network of pathways (see Figure 11). One idea is to find the shortest path connecting all of the significant subcores even if it passes through nodes for which we have no information. For example, if we could prove that the combination of pathways forms a small-world network it would be easier to understand the characteristics of disease propagation throughout the network.

Similarly, it is possible to attempt to improve the accuracy of the classification of healthy versus diseased patients by using alternative machine learning algorithms or improving the parameters of the ones used in this project. Due to the time constraints of this project I was unable to test the effects of many different parameters for each classification algorithm tested, however, it is imperative to do so in order to assess the efficiency of the feature generation algorithm more obtusely. In the K-neighbors algorithm we would have to investigate how changing the value `k` affects the prediction

score of the dataset. For the RBF SVM approach changing the parameter γ usually has a large effect on the prediction accuracy rate so varying such parameter can have large consequences in the prediction accuracy rate of the algorithm.

Furthermore, given the results obtained by other algorithms it is imperative to analyze the relationship between features in the dataset to further inform the process of classification algorithm selection for these types of datasets.

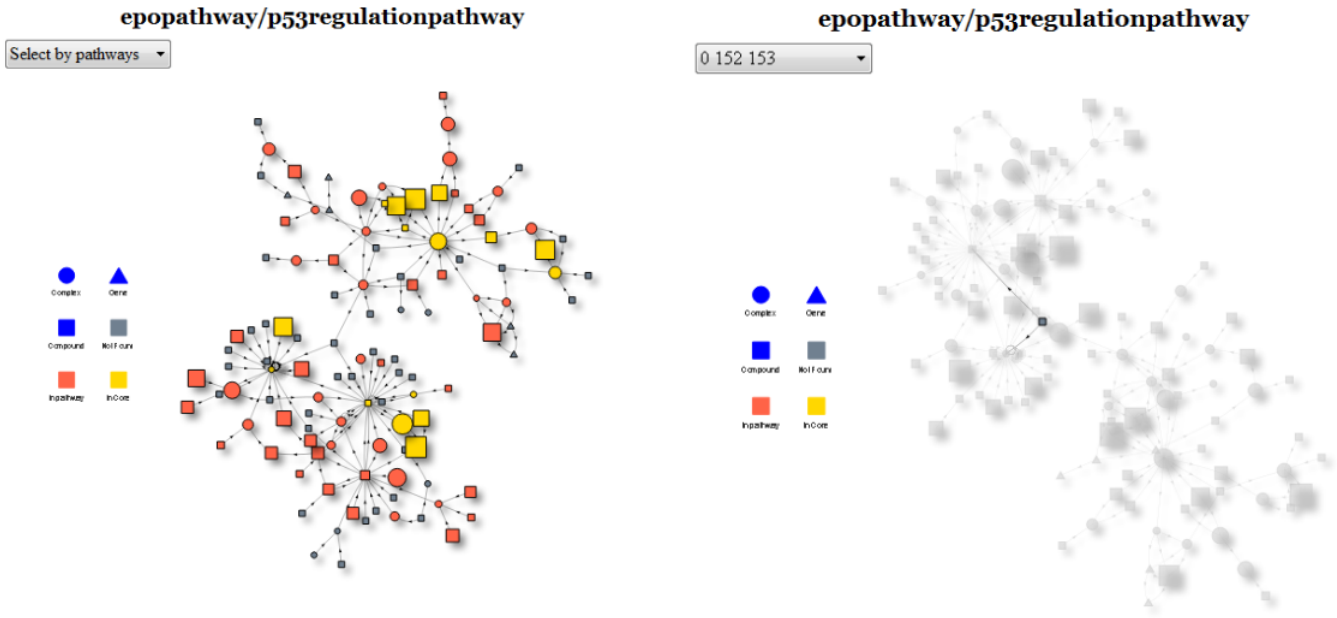

Figure 11: Example of Connected Pathways. The figure on the left shows two pathways connected by a single node for which there is no information. The figure on the right shows the connection between these two pathways (the image is flipped).

6 Acknowledgments

First and foremost, I would like to thank my parents for teaching me that working hard is the only way to achieve one's dreams. Thanks to all my friends and family members for all the support you provided me through these times.

Thank you to Dean Francis J. Doyle III for allowing me into his lab and providing me with resources for researching a project that was meaningful to me.

Thank you so much to Ms. Kelsey Dean and Dr. Burook Misganaw for putting up with me throughout the length of this project and for teaching me all I know about this research field. Thank you specially for all the time you spend on helping me finish my thesis. I could have not done it without your help.

Thanks also to the other members of the Doyle lab for being around with a smile and encouraging me to keep working.

Finally thank you to Prof. Finale Doshi-Velez and Prof. Yaron Singer for teaching me so much about computer science through your respective classes and for agreeing to read through this thesis. My only wish is that I had taken your classes sooner. Thank you also for inspiring me to keep learning about the computer science even when it gets hard.

Bibliography

[1] S.-A. Lee and K.-C. Huang, "Epigenetic profiling of human brain differential dna methylation networks in schizophrenia," *BMC Medical Genomics*, vol. 9, no. 3, p. 68, 2016.

[2] C. A. Castellani, B. I. Laufer, M. G. Melka, E. J. Diehl, R. L. O'Reilly, and S. M. Singh, "Dna methylation differences in monozygotic twin pairs discordant for schizophrenia identifies psychosis related genes and networks," *BMC Medical Genomics*, vol. 8, no. 1, p. 17, 2015.

[3] Y. Li, J. Xu, H. Chen, Z. Zhao, S. Li, J. Bai, A. Wu, C. Jiang, Y. Wang, B. Su, and X. Li, "Characterizing genes with distinct methylation patterns in the context of protein-protein interaction network: application to human brain tissues," *PloS One*, vol. 8, no. 6, p. e65871, 2013.

[4] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, no. 1, p. 140, 2007.

[5] R. A. Kallivayalil and V. P. Punnoose, "Understanding and managing somatoform disorders: Making sense of non-sense," *Indian Journal of Psychiatry*, vol. 52, no. Suppl1, p. S240, 2010.

[6] W. Katon, A. Kleinman, and G. Rosen, "Depression and somatization: a review: Part i," *The American Journal of Medicine*, vol. 72, no. 1, pp. 127–135, 1982.

[7] W. Katon, A. Kleinman, and G. Rosen, "Depression and somatization: a review: part ii," *The American Journal of Medicine*, vol. 72, no. 2, pp. 241–247, 1982.

[8] S. Patel, J. Kai, C. Atha, A. Avery, B. Guo, M. James, S. Malins, C. Sampson, M. Stubley, and R. Morriss, "Clinical characteristics of persistent frequent attenders in primary care: case–control study," *Family Practice*, vol. 32, no. 6, pp. 624–630, 2015.

[9] J. F. Peres, A. L. Gonçalves, and M. F. Peres, "Psychological trauma in chronic pain: implications of ptsd for fibromyalgia and headache disorders," *Current Pain and Headache Reports*, vol. 13, no. 5, pp. 350–357, 2009.

[10] H. Cohen, L. Neumann, Y. Haiman, M. A. Matar, J. Press, and D. Buskila, "Prevalence of post-traumatic stress disorder in fibromyalgia patients: overlapping syndromes or post-traumatic fibromyalgia syndrome?," in *Seminars in Arthritis and Rheumatism*, vol. 32, pp. 38–50, Elsevier, 2002.

[11] R. A. Hawkins, "Fibromyalgia: a clinical update," *The Journal of the American Osteopathic Association*, vol. 113, no. 9, pp. 680–689, 2013.

[12] M. Pongratz and D. Sievers, "Fibromyalgia-symptom or diagnosis: a definition of the position," *Scandinavian Journal of Rheumatology*, vol. 29, no. 113, pp. 3–7, 2000.

[13] B. Bondy, M. Spaeth, M. Offenbaecher, K. Glatzeder, T. Stratz, M. Schwarz, S. de Jonge, M. Krüger, R. R. Engel, L. Färber, D. E. Pongratz, and M. Ackenheil, "The t102c polymorphism of the 5-ht2a-receptor gene in fibromyalgia," *Neurobiology of Disease*, vol. 6, no. 5, pp. 433–439, 1999.

[14] K. D. Jones, T. Gelbart, T. C. Whisenant, J. Waalen, T. S. Mondala, D. N. Iklé, D. R. Salomon, R. M. Bennett, and S. M. Kurian, "Genome-wide expression profiling in the peripheral blood of patients with fibromyalgia," *Clinical and Experimental Rheumatology*, vol. 34, no. 2 Suppl 96, p. 89, 2016.

[15] R. Yehuda, "Post-traumatic stress disorder," *New England Journal of medicine*, vol. 346, no. 2, pp. 108–114, 2002.

[16] R. Yehuda, L. J. Siever, M. H. Teicher, R. A. Levengood, D. K. Gerber, J. Schmeidler, and R.-K. Yang, "Plasma norepinephrine and 3-methoxy-4-hydroxyphenylglycol concentrations and severity of depression in combat posttraumatic stress disorder and major depressive disorder," *Biological Psychiatry*, vol. 44, no. 1, pp. 56–63, 1998.

[17] P. Kuan, M. A. Waszczuk, R. Kotov, S. Clouston, X. Yang, P. Singh, S. T. Glenn, E. C. Gomez, J. Wang, E. J. Bromet, and B. J. Luft, "Gene expression associated with ptsd in world trade center responders: An rna sequencing study," *Translational Psychiatry*, vol. 7, no. 12, p. 1297, 2017.

[18] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[19] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS Computational Biology*, vol. 4, no. 11, p. e1000217, 2008.

[20] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[21] T. M. Mitchell, "Machine learning and data mining," *Communications of the ACM*, vol. 42, no. 11, pp. 30–36, 1999.

[22] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[23] R. Yang, B. J. Daigle Jr, L. R. Petzold, and F. J. Doyle III, "Core module biomarker identification with network exploration for breast cancer metastasis," *BMC Bioinformatics*, vol. 13, no. 1, p. 12, 2012.

[24] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, p. 56, 2011.

[25] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart, "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm," *Bioinformatics*, vol. 26, no. 12, pp. i237–i245, 2010.

[26] E. O. Paull, D. E. Carlin, M. Niepel, P. K. Sorger, D. Haussler, and J. M. Stuart, "Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (tiedie)," *Bioinformatics*, vol. 29, no. 21, pp. 2757–2764, 2013.

[27] K. Mitra, A.-R. Carvunis, S. K. Ramesh, and T. Ideker, "Integrative approaches for finding modular structure in biological networks," *Nature Reviews Genetics*, vol. 14, no. 10, p. 719, 2013.

[28] R. Hammamieh, N. Chakraborty, A. Gautam, S. Y. Muhie, R. Yang, D. P. Donohue, R. V. S. R. Kumar, B. J. Daigle Jr, Y. S. Zhang, D. A. Amara, S. A. Miller, S. Srinivasan, J. D. Flory, R. Yehuda, L. Petzold, O. M. Wolkowitz, S. Mellon, L. Hood, F. J. Doyle III, C. Marmar, and M. Jett, "Whole-genome dna methylation status associated with clinical ptsd measures of oif/oef veterans," *Translational Psychiatry*, vol. 7, no. 7, p. e1169, 2017.

[29] Z. Gu and J. Wang, "Cepa: an r package for finding significant pathways weighted by multiple network centralities," *Bioinformatics*, vol. 29, no. 5, pp. 658–660, 2013.

[30] C. Jacquemont and T. Taniguchi, "The fanconi anemia pathway and ubiquitin," *BMC Biochemistry*, vol. 8, no. 1, p. S10, 2007.

[31] "prcomp function | r documentation." `https://www.rdocumentation.org/packages/stats/versions/3.5.1/topics/prcomp`. (Accessed on 11/03/2018).

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[33] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

[34] P. A. Jaskowiak and R. Campello, "Comparing correlation coefficients as dissimilarity measures for cancer classification in gene expression data," in *Proceedings of the Brazilian Symposium on Bioinformatics*, pp. 1–8, Brasília, 2011.

[35] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[36] "1.4. support vector machines — scikit-learn 0.21.dev0 documentation." `https://scikit-learn.org/dev/modules/svm.html#kernel-functions`. (Accessed on 12/02/2018).

[37] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[38] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[39] N. Cristianini and J. Shawe-Taylor, *Support Vector Machines*, p. 93–124. Cambridge University Press, 2000.

[40] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[41] "Liblinear – a library for large linear classification." `https://www.csie.ntu.edu.tw/~cjlin/liblinear/`. (Accessed on 12/02/2018).

[42] "1.7. gaussian processes — scikit-learn 0.20.1 documentation." `https://scikit-learn.org/stable/modules/gaussian_process.html#gaussian-process-classification-gpc`. (Accessed on 12/03/2018).

[43] "1.10. decision trees — scikit-learn 0.20.1 documentation." `https://scikit-learn.org/stable/modules/tree.html#tree`. (Accessed on 12/02/2018).

[44] "1.11. ensemble methods — scikit-learn 0.20.1 documentation." `https://scikit-learn.org/stable/modules/ensemble.html#random-forests`. (Accessed on 12/03/2018).

[45] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[46] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[47] "Multi-class adaboosted decision trees — scikit-learn 0.20.1 documentation." `https://scikit-learn.org/stable/auto_examples/ensemble/plot_adaboost_multiclass.html`. (Accessed on 12/03/2018).

[48] "1.9. naive bayes — scikit-learn 0.20.1 documentation." `https://scikit-learn.org/stable/modules/naive_bayes.html#multinomial-naive-bayes`. (Accessed on 12/03/2018).

[49] T. Næs and B.-H. Mevik, "Understanding the collinearity problem in regression and discriminant analysis," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 15, no. 4, pp. 413–426, 2001.

[50] A. L. Blum and P. Langley, "Selection of relevant features in machine learning," in *Proceedings of the AAAI Fall Symposium on Relevance*, vol. 184, pp. 245–271, 1994.

[51] H. Zhang, "The optimality of naive bayes," *AA*, vol. 1, no. 2, p. 3, 2004.

[52] K. G. Raphael, M. N. Janal, and S. Nayak, "Comorbidity of fibromyalgia and posttraumatic stress disorder symptoms in a community sample of women," *Pain Medicine*, vol. 5, no. 1, pp. 33–41, 2004.

[53] "Kegg pathway: hsa04010." `https://www.genome.jp/dbget-bin/www_bget?hsa04010`. (Accessed on 11/21/2018).

[54] R. N. Wang, J. Green, Z. Wang, Y. Deng, M. Qiao, M. Peabody, Q. Zhang, J. Ye, Z. Yan, S. Denduluri, O. Idowu, M. Li, C. Shen, A. Hu, R. C. Haydon, R. Kang, J. Mok, M. J. Lee, H. L. Luu, and L. L. Shi, "Bone morphogenetic protein (bmp) signaling in development and human diseases," *Genes & Diseases*, vol. 1, no. 1, pp. 87–105, 2014.