# Kaggle Challenge: Diagnose Schizophrenia Using Multimodal Features From MRI Scans

Ao Li (liao@indiana.edu),
Chia-Hsuan Chou(chou5@umail.iu.edu),
Jing Wang(jw220@indiana.edu).

April 30, 2016

## 1 Objectives and Significance

Schizophrenia, as one of severe mental disorders, has a huge impact on patients' lives by affecting how them think and behave for a life long time. The diagnostics of schizophrenia is still largely dependent on subjects and focused on eliminating the symptoms, which are usually subtle and indistinguishable from other mental illnesses such as bipolar disorder so far. Several causes are considered by the scientists for schizophrenia. One major cause of this disease is abnormal brain structure and chemistry.

With the development of modern electromagnetic techniques, brain imaging data are widely used in clinical and cognitive neuroscience to help doctors to diagnose various brain disorders. Data mining methods based on the brain data from these techniques become extremely attractive to researchers over the past few years. As the objectives of

this project, we focus on taking advantage of brain MRI scans and applying data mining methodologies to build automatic diagnose model for individuals with schizophrenia. The significance of our work is to fit the gap between schizophrenia's diagnosis and its automatic prediction. We aim at developing data mining methods to construct a thorough and less subjective diagnosis for this disease. In our task, the data set is based on the Kaggle MLSP 2014 Schizophrenia Classification Challenge (Kaggle, 2014). Multimodal features of disorder individuals versus controls can be extracted from two kinds of MRI modalities, functional and structural MRI. Useful information like functional network connectivity (FNC) and source-based morphometry (SBM) can be further refined from brain maps that are obtained from MRI using independent component analysis (ICA). FNC are time wise correlation values that quantify the overall connection between independent brain maps. SBM are weights of brain maps derived from the application of ICA on the grey-matter portion of the brain, where signal transduction happens and forms the computational ability of a brain. Both data sets are considered relevant to schizophrenia detection, which could enhance the prediction accuracy of automatic diagnostics if properly used.

However, how to train a program using the most efficient algorithm to maximize prediction accuracy based on the two data sets is still a blank area. The competition offers a good opportunity for us to implement data mining tools we learned from the class. Before modelling and analyzing, we conduct exploratory data analysis onto data sets for a better understanding and representing of the brain. Since we notice that the data is highly correlated with each other in the exploring step, feature selection is applied to reduce the dimensionality of the data. In the final classification task, both logistics regression and support vector machine (SVM) with Gaussian radial basis are deployed with the selected features. Based on what we learned from the class, we are able to conduct better analyses about this data set with more understandable feature selection techniques and with more reasonable classification methodologies with a similar or higher prediction accuracies.

# 2 Background

The symptom of schizophrenia can last for years or be lifelong with various disabilities. Approximately 1.1 percent of US adult population are reported to have a 12-month prevalence of this disorder disease. So far, this brain disorder disease can not be cured because the causes of it are still largely unknown. Two major aspects are considered as contributors to the risk of developing schizophrenia. One is genes and environment and the other one is brain chemistry and structure. It is known that schizophrenia happens within families but no genetic information is found yet. Scientists are still trying to figure out key genes that cause the disorder. Researchers also think that some complex reactions of brain chemicals may play a role in schizophrenia development. More importantly, the brain structure and activity may be slightly/significantly changed during the disorder progress that can be detected by neuroimaging and magnetic resonance imaging technique (National Institute of Mental Health, 2016).

There are quite a few submissions for this Kaggle challenge. The team who won the second place of this competition implemented feature selection and classification methods for this MRI data. Firstly, it calculated the feature importance based on mean decrease of the Random Forest by Gini index criteria. In order to do the feature selection, it introduced a random vector to the features' set. Any feature whose importance are less than this dummy variable would no longer be considered as an important feature for the classification task. Second, a support vector machine with Gaussian radial basis kernel was applied to the training set. The final Area Under Curve (AUC) achieved by the above proposed approach is 0.923 (Lebedev, 2014).

For the team which won the third prize in the competition, they applied Distance Weighted Discrimination (DWD) on their model. Applying Support Vector Machine (SVM) to the High Dimension Low Sample Size (HDLSS) data can be tricky since SVM has data piling problem at the margin. Therefore, the winner used DWD to

solve the problem which were all based on finding the penalty cost. Before doing the feature selection, they used ten-fold cross validation on all features and determined the penalty parameter so that they could use the parameter in their DWD model. After drawing the performance under the Receiver Operating Characteristic (ROC) curve and finding out when the parameter reaches to 300, they get the maximum performance (Koncevicius, 2014).

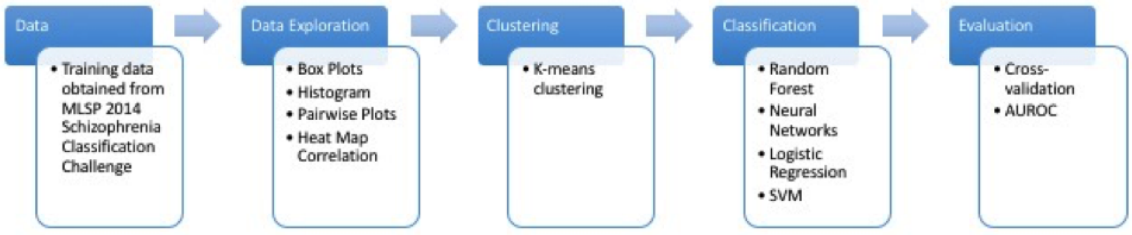
Previous work just combined the structural MRI and functional MRI together as potential features in this classification challenge. However, these are two different measurements on the same brain neuron network. Simply combing these two data sets might result in a multicollinearity representation of our brain image network. Different from previous work on this problem, our goal is to investigate more carefully about the sMRI and fMRI data sets in order to understand the relationship between them by unsupervised machine learning approaches for a better understanding and representing for features of brains. Based on the knowledge of the selected features, supervised machine learning techniques including logistics regression and SVM are applied for the classification task.

# 3 Methods

The data of this project are obtained from the Kaggle challenge: MLSP 2014 Schizophrenia Classification Challenge. This challenge aims at classifying healthy people from the patients with schizophrenia and schizo-affective disorder. The analysis is conducted based on the data from the magnetic resonance imaging (MRI), which is widely applied in understanding human brains in neuroscience nowadays. The training set of this challenge consists of 46 individuals with disease and 40 healthy people, who are diagnosed by Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, also

known as DSM-IV criteria.

The overall pipeline of our project can be found in the figure below.




## 3.1 Data Exploration

There are many ways to explore the data when we do not have any clue and assumption about the given data. Here we choose boxplot, histogram, scatterplot and heat map for data visualization. These visualization methods are applied to both FNC features and SBM features to depict the data graphically for our data exploration.

Firstly, we draw boxplot of the features, which is an easy and convenient way to know the quartiles of one or more sets of data intuitively. Using boxplot, we can obtain the inter-quartile range, first quartile, mean and third quartile within features. Moreover, boxplot can clearly tell us the outliers, which may influence the statistical values of the data and the variance, which can measure how the data points spread out.

Secondly, we draw histogram to display distributions of the numerical data from the features. Histogram can roughly estimate the probability distribution of a continuous variable by binning the range of value. Also, we display the density estimation to enable us comparing the distributions of different features clearly and easily.

Thirdly, in order to view the correlation between any two features from the data set,

we use scatter plot to visualize it. The scatter plot displays as a collection of points and the point's cluster in a band can tell us whether the two features are positive or negative correlated. Here notice that the correlation between two features does not imply the causality. In other words, we just want to know the relationship between different features.

Finally, drawing a heap map is the best way to examine the data by showing the correlation matrix intuitively. The heap map not only can show us the degree of correlation by colors but also can roughly present the cluster of features in the given data. Based on that information, we can obtain more clues for further data analysis.

## 3.2 Unsupervised machine learning analysis—Clustering

After exploring the data, we only have little picture about what our results should look like. Therefore, in order to address the problem, we derive a structure by clustering the data based on relationships among features in the data. We use the most common clustering method called K-means clustering. The advantage of K-means clustering is that the computational time of K-means clustering is not expensive compared to other clustering algorithms.

The goal of K-means clustering is to partition the observations into k different number of clusters. In order to assign a cluster to each data point, K-means clustering aims to find the centers $\mu_i$, $i = 1, \cdots, k$ of the clusters. The algorithm minimizes the L2 norm distance from the data points to the cluster center. The minimization objective function is written as,

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2,$$

where $c_i$ is the set of points which belong to cluster $i$. Here the K-means clustering

uses Euclidean distance for calculation. The algorithm will continue iterating until it converges to certain criteria, which means the assignments of the means do not change from one iteration to the next iteration by some threshold. Other distance measures or similarity measures are discussed in the results section.

## 3.3 Supervised machine learning analysis—Classification

### 3.3.1 Logistic Regression

Logistic regression is a widely used method for classification. It is able to classify observations by estimating the probability that an observation is in a particular category. The dependent variable of logistic regression is binary (class 0 and class 1) rather than continuous. In logistic regression, the relationship between one dependent variable and several independent variables (features) is measured using the sigmoid function. The idea behind the logistic function is to use sigmoid transformation to convert continuous values into probabilities. And we utilize 0.5 threshold on the values of probabilities to get binary classification results. The logistic regression formula is

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + x\beta.$$

After estimating these $\beta$ values by gradient descent methods, the probability $p$ that we are interested in can be rewritten as

$$p = \frac{1}{1 + exp\{\beta_0 + x\beta\}}.$$

### 3.3.2 Random Forest

We have more number of features than the data can present. In order to avoid the curse of high dimensionality, random forest as decision tree methods is implemented to give us the importance of features based on gini index. Curse of high dimensionality happens when the dimensionality increases. The volume of the space increases so fast that the available data become sparse which the scenario meets our case completely. Here we utilize Random Forests to rank both FNC and SBM features to obtain most important features and apply them for further classification.

Random forest can be thought as an arbitrary number of decision trees and by inputting every features in our data, those trees can vote for every feature and ultimately give us the most popular features. Since decision trees always over fit the data, random forest overcomes the over fitting problem by applying bootstrap aggregation to multiple decision trees by a training set $X = x_1, \cdots, x_n$ with responses $Y = y_1, \cdots, y_n$. Given $X$ and $Y$, we select a random sample with replacement of the training set and fit decision trees to these samples repeatedly. Below shows the formula of Random Forest Prediction.

$$s = \frac{1}{K} \sum_{K=1}^{K} K^{th},$$

where $K$ runs over every decision trees in the forest. The reason we average the prediction is to avoid the over fitting and the dependence problem of the decision trees, which may cause the sensitivity of noise. However, it is not enough to only apply the bagging algorithm to decision trees. Feature bagging is also significant since some features in the data might be very strong predictors for the target output and these features also have higher chance to be selected in many decision trees every time. It results in the correlation between one tree and another tree and indirectly causes the high sensitive noise in the training set. In order to rank the importance of features, Random forest has to be calculated with the out-of-bag error for each data point and

averaged over the forest. After the computation, we can obtain the importance score for the i-th feature and based on the ranking, we know the importance level of every feature in the data. This method is really efficient and significant for selecting features and avoiding curse of dimensionality. We are able to implement other supervised learning models on the data after the feature selection.

### 3.3.3 Neural Networks

Neural Networks (NNs) is a model inspired by biological neural networks and it is a very significant method for the supervised process since NNs can learn from the previous experience/examples and get better output. Suppose we have several input neurons (features), the input neurons will connect to other interconnected neurons (hidden layers) so that they can exchange messages between each other. In order to assure that the model learns from previous experience, we add numeric regression weights on the connection which can be tuned based on experiences and allow the system to learn from the data. Then, by keeping updating the weights, we are able to obtain the optimal output. The network function can be described as,

$$f(x) = h(\sum_i (w_i g_i(x))),$$

where the $h(\cdot)$ is the sigmoid function for this two-class classification problem, and $g(\cdot)$ can be different activation functions for NNs. In other words, the NNs initially give us a random guess about what it might be, then calculate the equation above and see how far the answer was from the actual one so that the model are able to adjust the connected weight appropriately and ultimately generates an optimal outcome. It is said that NNs can approximate the data well given wisely chosen weights and enough hidden layers.

NNs is a very efficient in our case since the advantages of NNs are to deal with large numbers of features and diversity of the data. However, the way that NNs handle the relationship between variables is still not easy for us to understand. Here our input neurons are the features and our output neuron is whether the person has Schizophrenia or not. By repeated learning and cross-validation, the results are improved and the accuracy of our prediction increases as well.

### 3.3.4 Support Vector Machine

Support Vector Machine (SVM) is one of the most powerful supervised learning models for data classification and regression analysis especially for small data sets. It looks for a projection of features into a certain hyper plane. Within this hyper plane, we can successfully linearly classify all training vectors into classes. For a given training data set, the SVM algorithm is able to generate a model that assigns new data points into one of the classes. The best separation would be the hyper plane that leaves the maximum margin from both classes. The points lying on the boundaries of the hyper plane are called support vectors.

Other than linear classification, SVM is also able to perform non-linear classification using the kernel techniques, which maps the input into high-dimensional spaces. In our project, we implement SVM with different kernels and decide to use Gaussian kernel as it gives highest prediction accuracy.

## 3.4 Evaluation

Specifically, cross-validation combined with area under the receiver operating characteristic curve (AUROC) is used to evaluate model performance. Cross-validation

is a model validation technique used for assessing model accuracy. One round of k fold cross-validation involves three steps: partitioning the data into k equal subsets; performing analysis on k-1 subsets; validating using the left 1 subset. In order to reduce variability, k rounds of validation are performed and the results are averaged over all rounds. ROC curve is a classical graph that plots a predictor's true positive rate (y-axis) over the false positive rate (x-axis). The area under the curve is named as AUROC. It measures discrimination, which is the ability of the model to correctly classify the binary class. The larger the AUROC is, the better the predictor behaves. The formula to calculate AUROC is as below:

$$A = \int_{\infty}^{-\infty} TPR(T)FPR'(T)dT,$$

where TPR stands for true positive rates and FPR means false positive rates.

# 4 Results

## 4.1 Data exploration

For these 86 people data, there are 379 features in the Function Network Connectivity (FNC) and 33 features in the Source-Based Morphometry (SBM) loadings. There are more features than the data points. The first 15 features of SBM distribute more similar to each other than the FNC features as shown in Figure 1. And the first three loadings are less correlated to each other, which means that the features are more independent from each other and measuring different scopes of the brain shown in Figure 2. The heat map plot (Figure 3) based on the Pearson's correlations also shows that the first 15 features of the SBM are not related to each other.

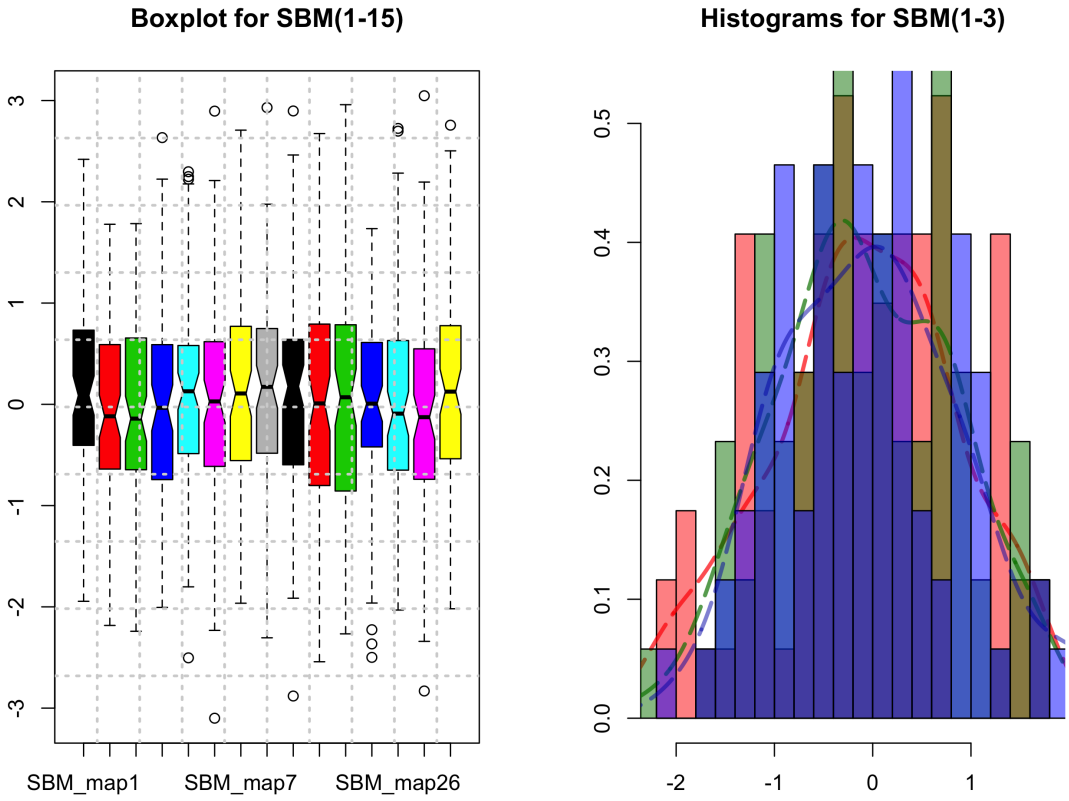



Figure 1: Boxplots and histograms of SBM features

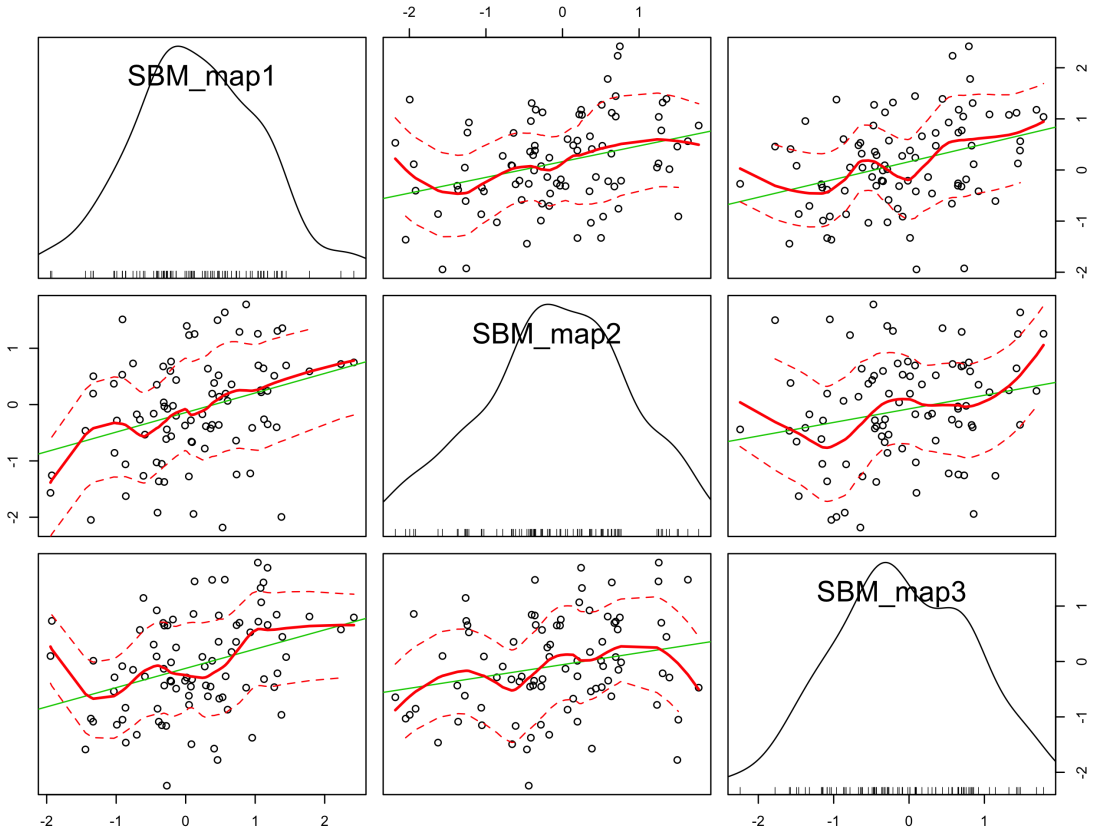



Figure 2: Scatter plots of SBM(1-3)

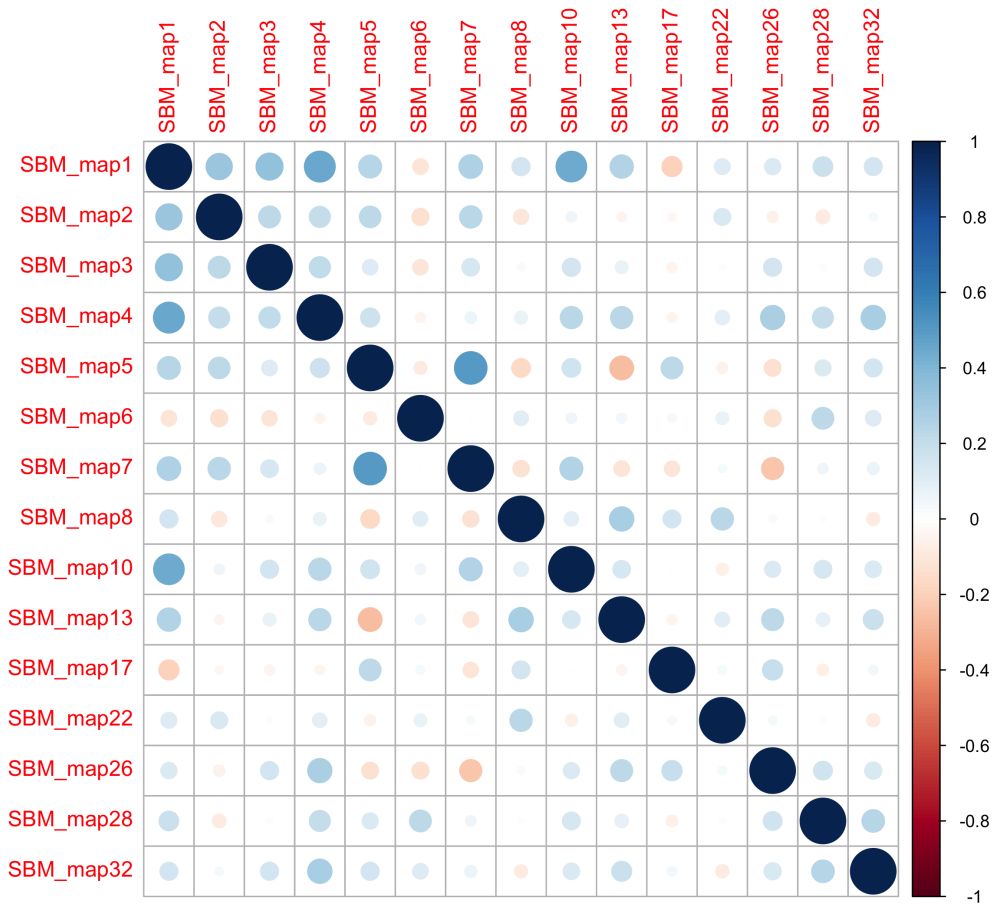



Figure 3: Heat plot of correlation matrix

Then, we present the boxplots for the first 15 features of it and histograms of the first three features of FNC in Figure 4. We can see that inputs of FNC is from $-1$ to 1 since they are calculated by correlations between regions of brain. Also, the histograms and density estimations of them look different from each other. According to the boxplots and histograms, we fail to figure out the difference between healthy people and people with schizophrenia.

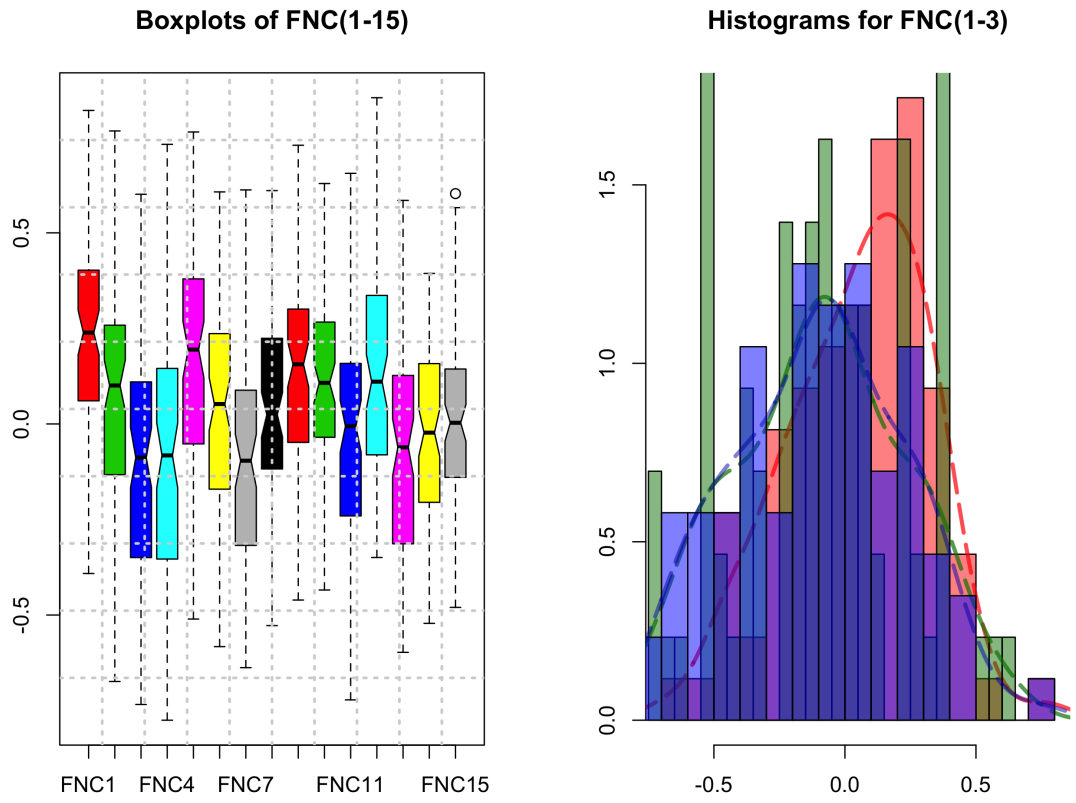



Figure 4: Boxplots and histograms of FNC features

Too many features will bring us the curse of high dimensionality. Based on the scatter plots of these 379 loadings of the FNC in Figure 5, we can see that the first three features are linearly correlated to each other. Then, we also calculate the Pearson's correlations between the first 15 FNC features. The heat plot in Figure 6 of the correlation matrix shows that the first seven FNC features are related to each other. It seems that the first seven features are more similar to each other than the rest eight features. This suggests that we need to select features before analyzing the relationship between the schizophrenia and these features from the brain image data. Before analyzing the data by both unsupervised and supervised machine learning techniques, we combine the FNC and SBM features as 411 features for 86 people.

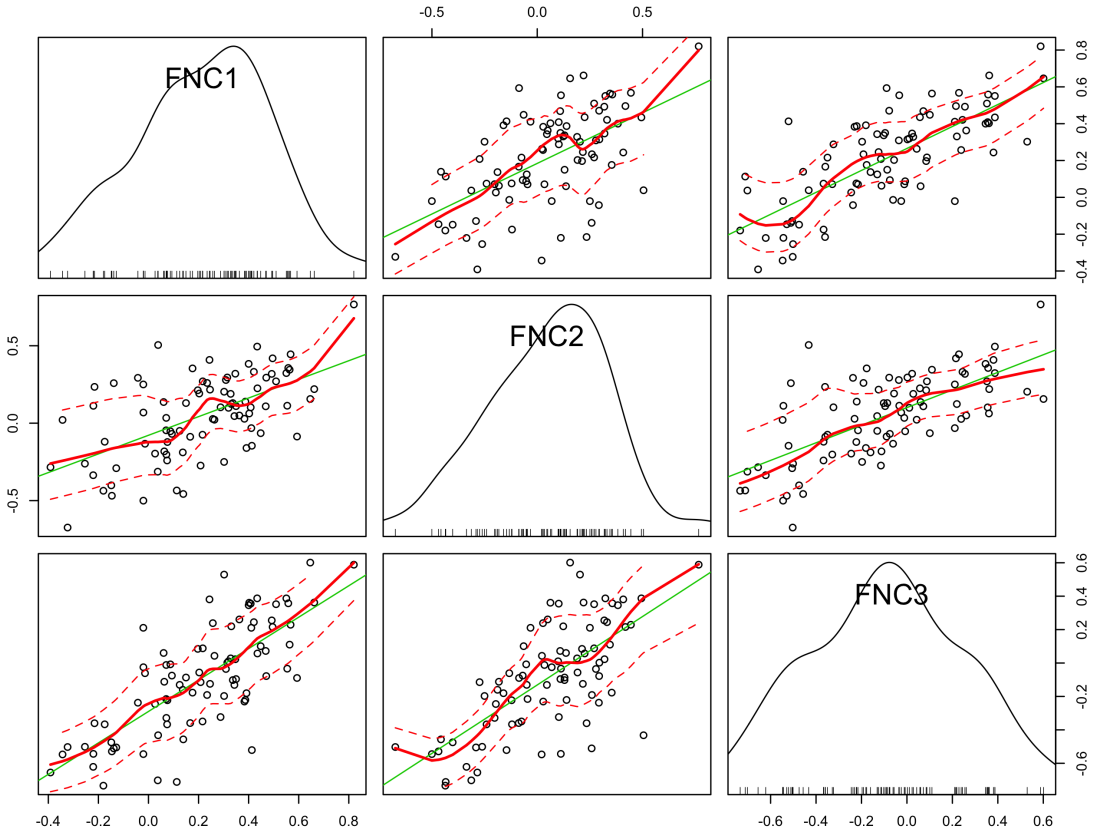



Figure 5: Scatter plots of FNC(1-3)

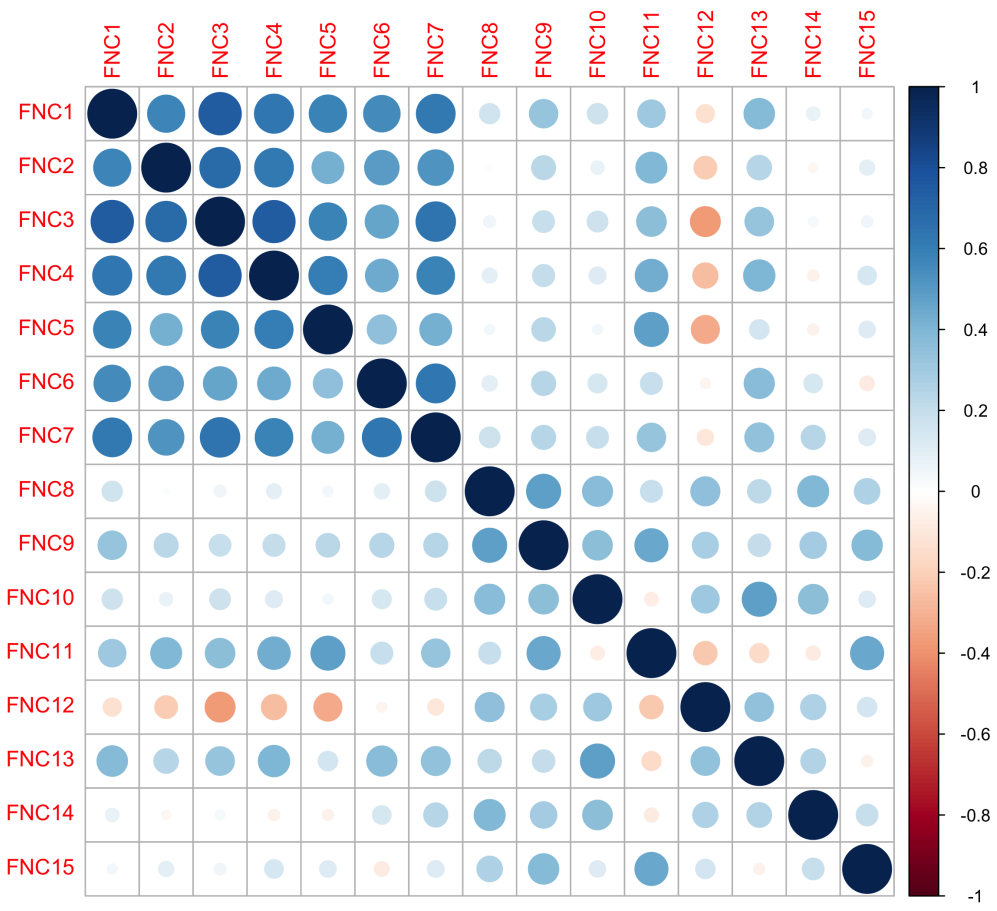



Figure 6: Heat plot of correlation matrix

## 4.2 Clustering

This section provides a basic k-means clustering analysis for unsupervised machine learning. The k-means clustering algorithm finds clusters that points in the same cluster are more similar and data in different clusters are dislike to each other. Firstly, we calculate the distance between 411 features within these 86 people by the Euclidean metric. We calculate 20 iterations of the within groups sum of squares and try to find a L-curve for Figure 7. However, it is hard to decide the number of clusters based on the below plot. Then, we try to make the number of clusters equal to 6. The clusters

for features are as the following:

```
Cluster 1:
[1] "FNC87"  "FNC95"  "FNC99"  "FNC100" "FNC102" "FNC110" "FNC116"
[8] "FNC123" "FNC136" "FNC243" "FNC247" "FNC248" "FNC249" "FNC256"
[15] "FNC264" "FNC270" "FNC271" "FNC273" "FNC285" "FNC339" "FNC342" ...

Cluster 2:
[1] "SBM_map1"  "SBM_map2"  "SBM_map3"  "SBM_map4"  "SBM_map5"
[6] "SBM_map6"  "SBM_map7"  "SBM_map8"  "SBM_map10" "SBM_map13"
[11] "SBM_map17" "SBM_map22" "SBM_map26" "SBM_map28" "SBM_map32" ...

Cluster 3:
[1] "FNC28"  "FNC29"  "FNC30"  "FNC31"  "FNC32"  "FNC33"  "FNC35"
[8] "FNC54"  "FNC55"  "FNC57"  "FNC79"  "FNC80"  "FNC81"  "FNC82"
[15] "FNC86"  "FNC88"  "FNC96"  "FNC103" "FNC104" "FNC105" "FNC109" ...

Cluster 4:
[1] "FNC3"  "FNC4"  "FNC7"  "FNC11" "FNC13" "FNC19" "FNC21"
[8] "FNC24" "FNC27" "FNC44" "FNC51" "FNC63" "FNC67" "FNC69"
[15] "FNC76" "FNC78" "FNC91" "FNC92" "FNC93" "FNC114" "FNC115" ...

Cluster 5:
[1] "FNC2"  "FNC5"  "FNC6"  "FNC8"  "FNC9"  "FNC10" "FNC12"
[8] "FNC14" "FNC15" "FNC17" "FNC18" "FNC20" "FNC26" "FNC40"
[15] "FNC41" "FNC42" "FNC43" "FNC45" "FNC46" "FNC52" "FNC53" ...

Cluster 6:
[1] "FNC1"  "FNC16" "FNC22" "FNC23" "FNC25" "FNC34" "FNC36"
```

[8] "FNC37" "FNC38" "FNC39" "FNC47" "FNC48" "FNC49" "FNC50"
[15] "FNC56" "FNC58" "FNC60" "FNC61" "FNC62" "FNC64" "FNC72" ...

We can see that all of the SBM features follow a single cluster and most of the clusters are formed by nearby features. Like the findings in the previous section, this also reflects that the nearby features are related to each other. Based on the clustering Figure 8 between FNC 2 and FNC 5, we realize that finding clusters is very hard for high dimensional data. It is almost impossible to tell the difference among the clusters of the FNC features. In the future, we will consider embedding these feature vectors before applying clustering analysis. Moreover, we can carry out experiments with different measures of distance or similarity including min-max and cosine similarity measures with normalized data. Because it is hard for us to explain the results of k-means clustering for this high dimension features, the future direction is to communicate with the experiments in brain image for further explanations and clarifications.

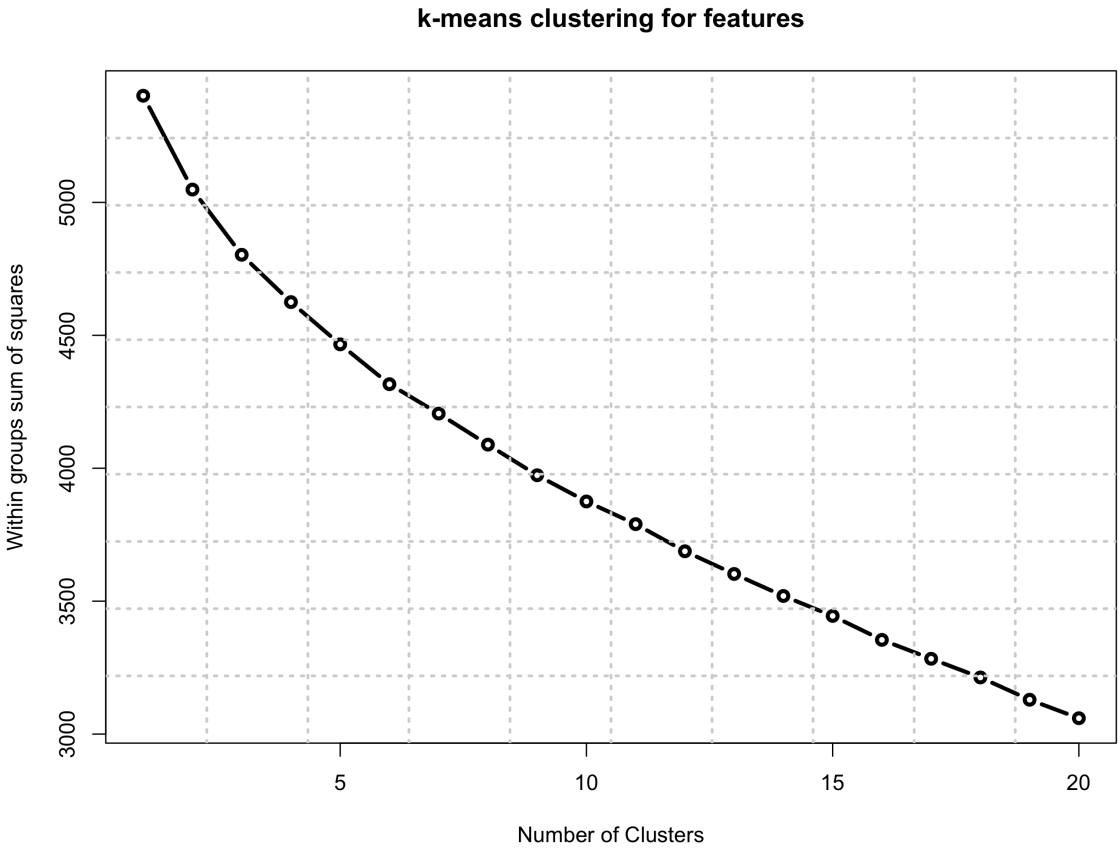



Figure 7: WSS plot of k-means clustering

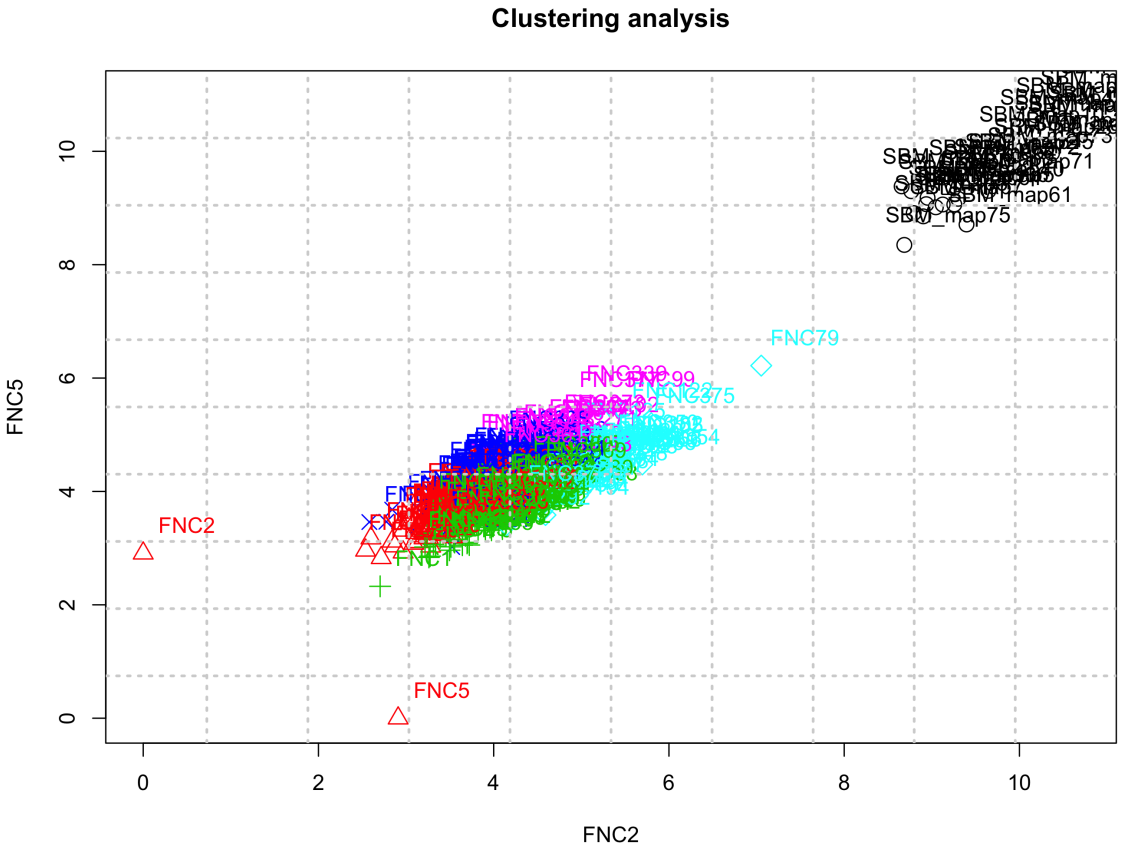



Figure 8: k-means clustering based on FNC 2 and FNC 5

## 4.3 Classification

We randomly pick 75% of the data as training set and the left 25% as the testing set. In the training 64 people's data, there are 30 people with the disease and 34 people are healthy. Similarly, there are 10 people have Schizophrenia and 12 of them are as control group.

### 4.3.1 Logistic Regression

For this binary classification problem, logistic regression is a competitive tool. We implement a logistic regression model with all of the features from both FNC and SBM. Also, we use 0.5 as the threshold to decide the classes of data. Then, the trained logistic regression model is tested with the testing data. The prediction accuracy for testing data set is 0.6818181818, and AUC is as large as 0.6833333333. Based on the ROC plot (Figure 9) and prediction and AUC values, we are not satisfied with our results. A ROC curve is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the total actual positives versus the fraction of false positives out of the total actual negatives, at various threshold settings. Then, we start investigating the data for the reasons that we fail to obtain better prediction accuracies. From the previous data exploratory and clustering analysis, one of the reasons we think is that the brain image data is highly correlated between features. Also, the logistic regression also warns us that the number of dimensions is higher than the number of data. From all of these reasons, we adapt random forest technique to decide the importance of features for future investigation.

```
> roc.curve(threshold, print=TRUE)
          Predicted
Observed  0  1
        0 8  4
        1 3  7
           FPR             TPR
0.3333333333  0.7000000000
> mean(glm.pred==y.test)
[1]  0.6818181818
> auc(M.ROC[1,],M.ROC[2,])
```

```
[1] 0.6833333333
```

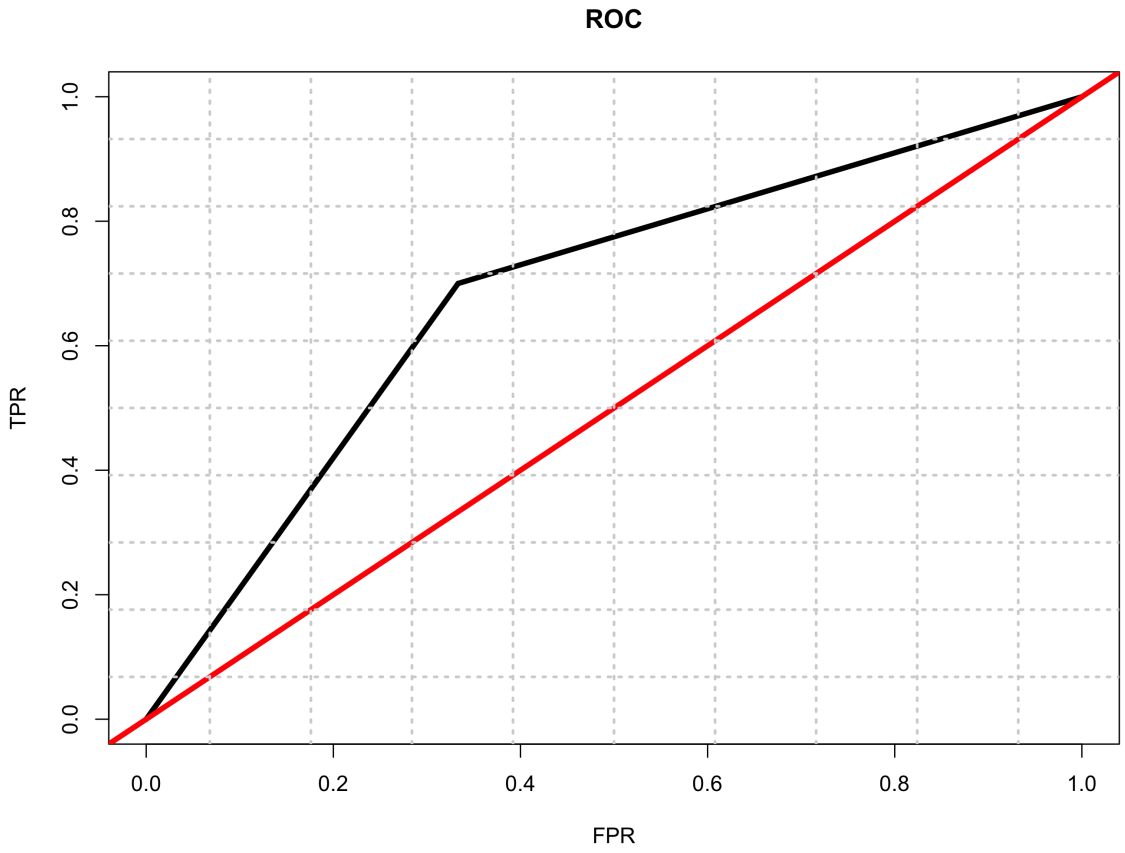



Figure 9: ROC plot for logistic regression

### 4.3.2 Random Forest

Random forest provides an improvement over decision trees for this brain image data (Lebedev, 2014). The reason that we implement random forest to decide the importance of features is that decision trees give us an attractive and easily interpreted result. In the left part of Figure 10, we plot the most important features by random forest of the brain image data. Also, based on the right figure, we notice that there exists a

drop after the 13th most important feature indicated by the orange line. We decide to pick the following important features

[1] ‘‘FNC244’’ ‘‘SBM_map67’’ ‘‘FNC295’’ ‘‘SBM_map61’’ ‘‘FNC226’’
[6] ‘‘FNC183’’ ‘‘FNC33’’ ‘‘FNC302’’ ‘‘FNC220’’ ‘‘FNC243’’
[11] ‘‘FNC37’’ ‘‘FNC289’’ ‘‘SBM_map36’’

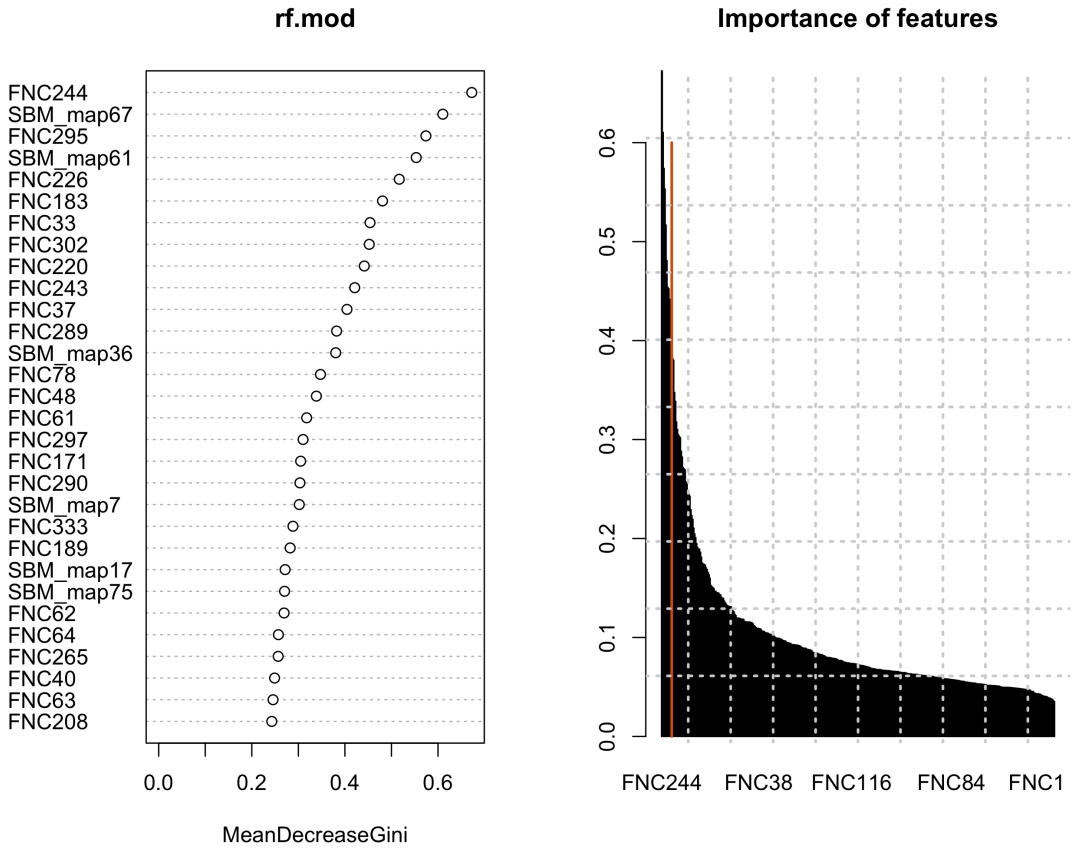



Figure 10: Random Forest plots of important features

### 4.3.3 Neural Networks

Nowadays, neural networks have become much more popular in the area of deep learning. Since there are only 64 people left in the training data set, we only pick the first 13 most important features for the Neural Networks. Before fitting a neural network, it is always required to normalize the data. Otherwise, it will lead to a difficult training process without normalization. There are different normalization methods to scale the data. Here, we choose to use the min-max method. The reason that neural networks have not always been popular, partly because they were, is that there is no fixed rule to choose number of layers and neurons. As a simple experiment, we choose to use 3 hidden layers with the configuration as 13:9:6:4:1. The input layer has 11 inputs with features as FNC244, SBM_map67, FNC295, SBM_map61, FNC226, FNC183, FNC33, FNC302, FNC220, FNC243, FNC37, FNC289, and SBM_map36. The three hidden layers have 9, 6, and 4 neurons and the output layer is our probability of schizophrenia index. Also, we use 0.5 as the threshold for the probability outcomes. For any probability larger than 0.5, we predict 1 value, which means that the person has the schizophrenia. Then, we use this model to make prediction on the testing data set and achieve a 72.73% prediction accuracy.

```
> table(pr.nn_,y.test)
        y.test
pr.nn_   0  1
      0 11  5
      1  1  5
> mean(pr.nn_==y.test)
[1] 0.7272727273
```

Note that the the neural network method fits the model by gradient descent. Then, we apply a 4-fold cross validation in order to achieve a better fitting model on the testing

data set. The training set is equally split into 4 folds with each fold has 16 data inputs. At each time, we randomly pick three folds to train the model and test on the left over data set. The least prediction error on the training set based on the 4-fold cross validation is 0.3125. We apply this least prediction error model to the testing data set. The prediction accuracy is raised to 0.8181818182. A Receiver operating characteristic (ROC) plot is present and the area under curve (AUC) of the plot is 0.85.

```
block
 1  2  3  4
16 16 16 16
> cv.error
[1] 0.3125 0.3750 0.5000 0.3125
> cv.error[which.min(cv.error)]
[1] 0.3125
> roc.curve(threshold,print=TRUE)
          Predicted
Observed   0   1
        0 10   2
        1  2   8
          FPR             TPR
0.1666666667 0.8000000000
> mean(pr.nn==y.test)
[1] 0.8181818182
> auc(M.ROC[1,],M.ROC[2,])
[1] 0.85
```

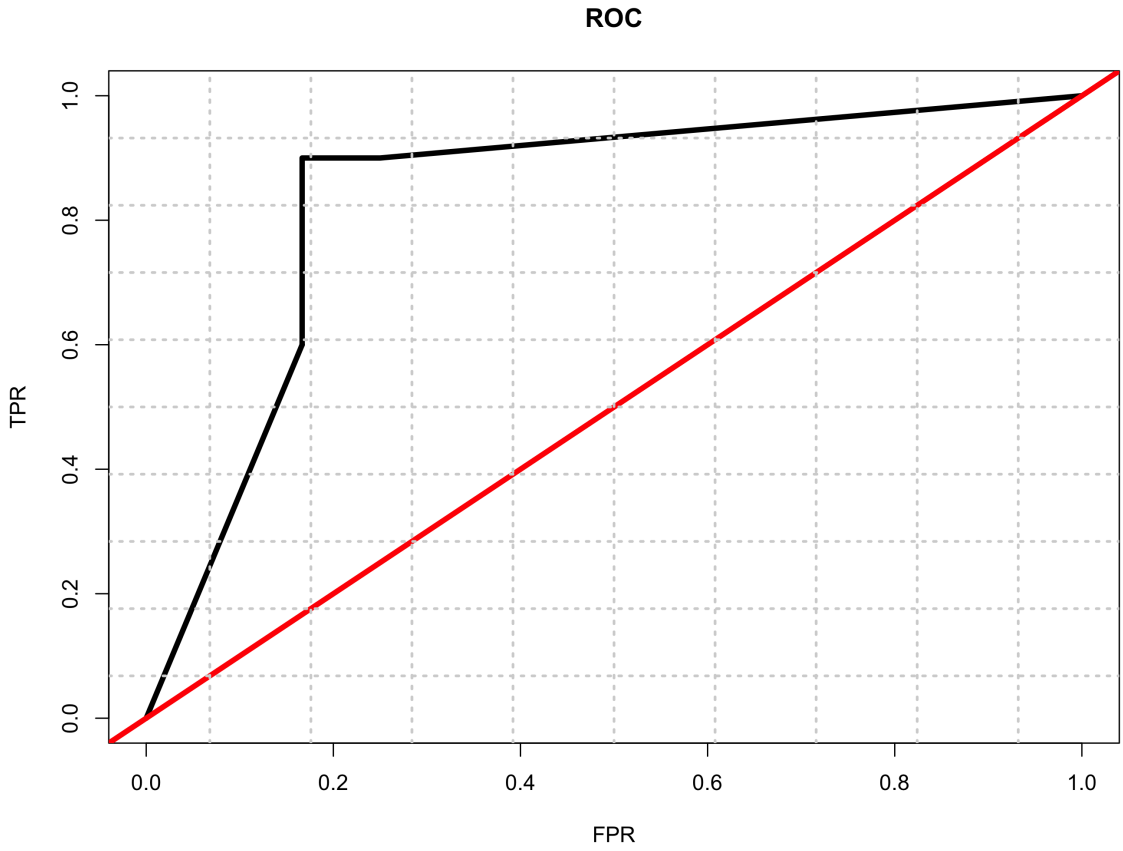



Figure 11: ROC plot for neural network

Although there is no fixed rule to choose number of layers and neurons, we conclude that applying cross validation technique increases the accuracy of the prediction on the testing data set. For future study of applying deep learning to brain image data, a recurrent neural network structure will be investigated since it directly models the internal memory to process arbitrary sequences of inputs.

### 4.3.4 Support Vector Machine

Support vector machine usually produces nonlinear boundaries by constructing a linear boundary in a transformed version of the feature space. In this project, we try different kernels for the projection of features and find that the Gaussian kernel behaves relatively well. In order to fit a soft margin separation, we use "cost" to specify the cost of a violation to the margin. Also, we apply ten-fold cross-validation to choose the shrinkage parameters. Based on the training set, the classification accuracy of the learned model for testing data is 0.9090909091. We also compare the ROC curve in Figure 12.

```
> roc.curve(threshold, print=TRUE)
          Predicted
Observed   0   1
       0  11   1
       1   1   9
             FPR                TPR
0.08333333333  0.90000000000
> mean(svm.pred==y.test)
[1] 0.9090909091
```

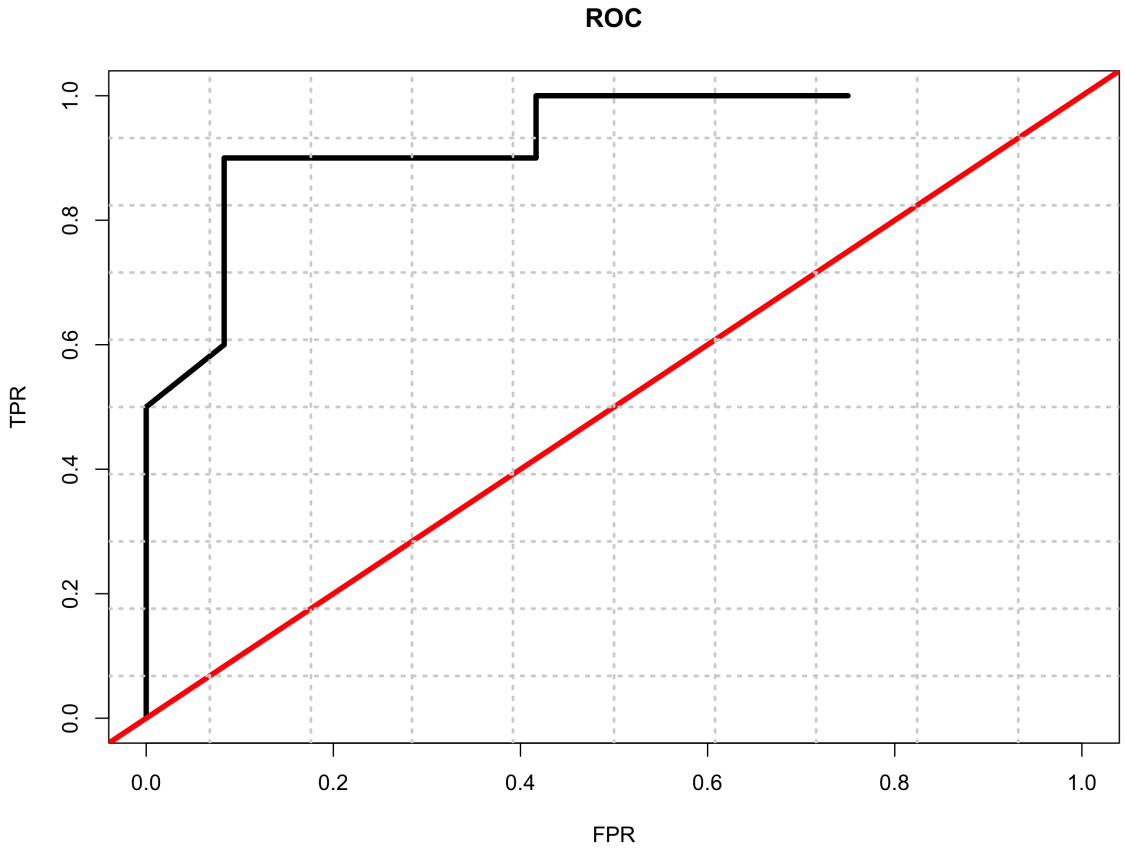



Figure 12: ROC plot for Support Vector Machine

# 5 Conclusions

The diagnosis of schizophrenia used to be largely dependent on doctors' experience and bio-markers since it does not show huge difference from other mental disorders. In this data mining project, we target at fitting the gap between schizophrenia's diagnosis and its automatic prediction including several learning techniques. Data mining methods including logistic regression, random forest, neural networks, and support vector machine are implemented to construct a thorough and less subjective diagnosis for this

disease. The brain imaging data FNC and SBM are treated as potential features to predict schizophrenia through different classifiers.

In the beginning of data exploratory, we draw boxplots and histograms in order to intuitively learn more about the features of data before conducting any modeling and analyzing steps. By the scatter plots and heat maps of the correlation matrix, we notice that there are more features than what the data can present and the features are highly correlated to each other. In the unsupervised machine learning section, k-means clustering is utilized to find reasonable clustering of the brain image data. However, it is hard for us to explain the clusters of features. Different measurements of distance or similarity will have an impact on the clustering analysis. For future studies, distance measurements that are implicitly designed for this spatial temporal correlated data will be studied. We also explore and investigate carefully about the predictors in order to understand the relationship between brain disorders and brain imaging data.

For the classification task, logistic regression with all features are first implemented. Prediction accuracy, ROC curve and AUC values are calculated. Since the logistic regression behaves poorly with the full features, we decide to conduct feature selection in order to avoid the curse of high dimensionality. Then, random forest is applied to decide the importance of features. With the important features selected from random forest, we compare and discuss different algorithms learned from the data mining class including neural networks and support vector machine. Based on the results of prediction accuracy and AUC, we conclude that SVM performs the best among these methods. But, we only implement the SVM with a basis Gaussian kernel for predictors. For future studies, SVM with appropriate kernels should be discussed for this data set. We know that the brain imaging data is collected with spatial-temporal correlation. A specific kernel based on this correlation matrix will be investigated to increase specificity and model accuracy.

# 6 Individual task

I (Ao Li) am in charge of 1 objectives and significance, 4 results, and 5 conclusions. In the first section of this paper, I list all the objectives of this Kaggle challenge. After reading more materials about the brain disorder schizophrenia, I discover the significance of applying data mining to the diagnosis of schizophrenia. In the beginning of data exploratory, I draw boxplots and histograms in order to intuitively learn more about the features of data before conducting any modeling and analyzing steps. By the scatter plots and heat maps of the correlation matrix, I notice that there are more features than what the data can present and the features are highly correlated to each other. In the unsupervised machine learning section, k-means clustering is utilized to find reasonable clustering of the brain image data. However, it is hard to explain the clusters of features. I also explore and investigate carefully about the predictors in order to understand the relationship between brain disorders and brain imaging data. Logistic regression with all features are first implemented. Prediction accuracy, ROC curve and AUC values are calculated. Since the logistic regression behaves poorly with the full features, I decide to conduct feature selection in order to avoid the curse of high dimensionality. Then, random forest is applied to decide the importance of features. With the important features selected from random forest, I compare and discuss different algorithms learned from the data mining class including neural networks and support vector machine. Based on the results of prediction accuracy and AUC, we conclude that SVM performs the best among these methods. But, we only implement the SVM with a basis Gaussian kernel for predictors. I majorly try different unsupervised and supervised models with several set ups to pursue more interpretable and reasonable prediction results. In the conclusions, I summarize the findings of this project and some thoughts about the future directions for further investigation. Chia-Hsuan Chou and Jing Wang are in charge of 2 background and 3 methods. They provide details of the techniques that I applied in the results section. More background of the

schizophrenia and MRI technology is collected and reviewed by them.

I (Chia-Hsuan Chou) am in charge of part of the 3 methods which are data exploration, unsupervised machine learning clustering, and supervised machine learning Random Forest and Neural Network. When researching and exploring the methods, I find out that there are lots of ways to train the data and those methods all have their pros and cons. Some data might fit the method after pre-processing and some method might fit the data easily but the result we get is not quite good and reasonable. Therefore, it is hard to tell what method will give us the best prediction and fit our data well and it is also important to decide what method should be implemented first because the order will matter the result a lot. Moreover, it is quite necessary that we try error and then study and modify the process of the method in the same time since every subtle modification can cause the accuracy of our prediction. Here we find out that feature selection is the key point to infuence the accuracy of the prediction since our data points are much less than the features the research given. Thus, after considering the reduction of the dimension, the performance increases and it gives us a better result.

I (Jing Wang) am in charge of the 2 background part and part of the 3 method (logistic regression, support vector machine and evaluation). In the step of background information gathering, I searched on NCBI about the disease itself and available NIH websites for its potential causes and treatments. Also, previous team work on this task on the Kaggle website are carefully reviewed and summarized. In the method part, different machine learning methods are proposed to do the automatic schizophrenia prediction. Logistic regression is a perfect model for binary outcome prediction. SVM is a classical data mining method we learnt in the class, which is also proposed by other existing teams on this task. However, neither method can generate a perfect model for this task since various methods can be applied to do feature selection, which largely affects the outcome. In the evaluation step, the classical cross-validation is proposed. AUROC is calculated as required by the task description. Our model gives high accuracy on this

task but further feature selection criteria can be proposed to make the model better.

# References


1. Alexander V. Lebedev, The 10th Annual MLSP Competition: Second Place, 2014.

2. Kaggle, MLSP 2014 Schizophrenia Classification Challenge, 2014.

3. Karolis Koncevicius, MLSP 2014 Schizophrenia Classification Challenge: 3rd position (solution), 2014.

4. National Institude of Mental Health, Schizophrenia, 2016.

5. Richiardi, Jonas, et al., Machine learning with Brain Graphics: Predictive Modeling Approaches for Functional Imaging in Systems Neureoscience. IEEE Signal Processing Magazine, 2013, vol.30, no.3, p. 58-70.