Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 8, 2020

ANALYSIS OF EEG DATA AND PREDICTION OF SCHIZOPHRENIC CHARACTERISTICS BASED ON EEG SIGNALS

S. Saravanan[1], S. Bhutada[2], Y. Shishodia[3]

[1,2,3]**Computer Science and Engineering, SRM IST, Kattankulathur, India**
E-mail: [1]Saravans2@srmist.edu.in, [2]Sampannam_purshottam@srmuniv.edu.in, [3]Yashkumar_sudhir@srmuniv.edu.in

Received: 08.05.2020 ***Revised: 06.06.2020*** ***Accepted: 30.06.2020***

Abstract
Schizophrenia is a mental illness in which the people suffering from it misinterpret reality. It causes delusions or hallucinations which causes disruption in regular thinking. The patient sometimes has problems understanding whether what they are experiencing is actual reality or part of a hallucination. People with schizophrenia require treatment that spans their entire life. Timely treatment generally helps getting the symptoms under control prior to the development of serious complications and help improve long-term outlook. A number of mental disorder studies have been using Machine Learning approaches for examining Electroencephalography (EEG) data. Machine Learning Algorithms applied on reasonably large EEG datasets can help in faster diagnosis, early risk prevention, and possibly prevention of the disease. We will be making efficient use of Artificial Neural Networks (ANN) algorithm to train a prediction model. Previous results show that accuracy of neural networks algorithm has been consistently high. EEG data from 32 control and 49 schizophrenic patients is contained in the dataset. The dataset will be split to make a training and testing set. Both sets contain control subjects and patients. The analysis will be done in RStudio. The analysis will help in better understanding of the effects on schizophrenia on the brain and pinpointing the specifics of the disease. It will also enable us to identify if a patient has schizophrenia by looking at the EEG report of a patient.

Keywords --- Artificial Neural Network, Data Analysis, Schizophrenia, EEG, Machine Learning

DOI: http://dx.doi.org/10.31838/jcr.07.08.219

INTRODUCTION

Data analysis is a technique of amending, modifying, and examining data which helps the user unfold useful understanding, generate conclusions, and aid making decisions.

It has many different approaches, comprising diverse techniques. It is used extensively in many different domains like the field of science, social sciences and business.

Data mining, however is a type of technique for analyzing data that puts focus on knowledge discovery and modeling for predictive purposes, whereas business intelligence (BI) is about data analysis that is focused on aggregating data by using business data.

There are three ways in which we can categorize statistical application- descriptive analysis, exploratory analysis and confirmatory analysis.

Exploratory analysis are about uncovering new patterns in the data. Confirmatory analysis helps us to test an already existing hypothesis. Predictive analysis is used for classification of the data or for forecasting data by implementation of statistical models.

Data integration is usually treated as a precursor to analyzing data, which ultimately leads to data visualization. Data mining and data modeling are terms which are usually used interchangeably.

Data Analysis Process consists of the following iterative phases –

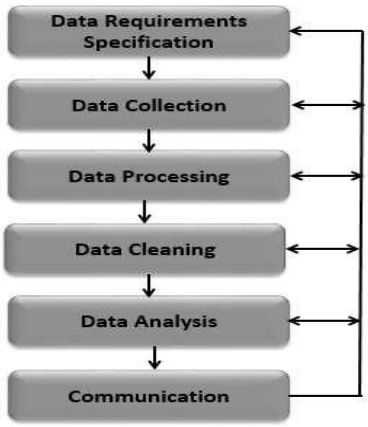

Figure 1. Data Analysis Process

Data Requirements Specification

The first and the most important thing to do is to understand what question we want answered when we perform data analysis. That activity is called specifying our data requirement. Then we use the requirement to identify which inputs we require and what sort of output we seek from it.

Data Collection

Collecting data is the next step that we take. We collect information on the variables have we have selected in the step of identifying requirements. We must ensure that the data is as accurate and without bias as possible. That is what will ensure that the process of our data collection will ultimately yield successful results or not. Accurate data collection can help us achieve high levels of accuracy.

Data can be collected by self experimentation, obtained from databases available on web pages ,etc. Data can be structured or

unstructured, or it might consist of information that we do not require for our purposes. This is why processing and cleaning the data are very important steps.

Data Processing
Since the data collected can be in format, one step has to be about putting the data in a structure so that we can apply methods of analysis to them easily. It also improves the readability of the data and makes it more understandable

Data Cleaning
Even after the processing of the data, it is often seen that the data might contain errors, double entries or missing values. Thus in the process of cleaning we remove such errors so that our analysis is not affected too much because of it.

Data Analysis
We can now use the data which has been processed and cleaned for further work. Different kinds of techniques are available for analysis of data are used for understanding, interpreting, and deriving conclusions depending on the requirements. Some tools are also available for Data Visualization that can be used for examining the data in a graphical format. This will help us in obtaining additional insights into the data.

Many different techniques, namely Correlation and Regression can be used to help us understand the relation underlying in the data. These techniques help us in analyzing and hence share the findings further.

Communication
The analysis results should be reported in a format as specified by the users to facilitate their decision making. The feedback collected from the users might involve additional analysis.

The analysts can select data visualization techniques which will help them communicate the message clearly to the users. Some techniques commonly used are tables and charts. There is a option of highlighting the important information in different colours and formats so as to aid in understanding.

Problem Definition
Humans (and many other animals) can reduce or suppress their brains' responses to sensory consequences that are a result of their own actions. The nervous system manages this task by using a "corollary discharge forward model system". According to this system, a copy of an expected motor plan is forwarded from motor to sensory cortex.

The neurons dispense a "corollary discharge" of the expected sensory effects of the expected motor act. A simple example of this would be that when we move our eyes from left to right, our brain cells know the environment is not shifting.

Schizophrenia is a major mental disorder that affects approximately 1% of people around the world. One possible explanation for some of the symptoms of schizophrenia is that one or more problems with the corollary discharge process in the nervous system makes it difficult for patients to judge the difference between stimuli generated internally and externally.

Therefore, studying this process and its relationship to symptoms in the illness might allow us to better understand abnormal brain processes in patients with this diagnosis.

Artificial Neural Networks
Artificial neural networks or ANN are mainly computing systems alternately known as connectionist systems which are inspired by, but not the same as biological neural networks that comprise mainly animal brains. These computing systems "learn" to perform tasks by examples, generally independent of the task rules. For example, the system of image recognition can get conditioned in order to identify pictures containing cats by analysing images that have been manually labelled as "cat" or "no cat" and using these findings to recognise cats when exposed to various other images.

They carry this out without any prior exposure to knowledge of cats, like they have fur, tails and whiskers. Instead, they generate results by identifying characteristics from the examples that they process automatically.

ANN are based on a group of nodes or units called artificial neurons. Each connection can communicate a signal to other neurons as the synapses in a biological brain. An artificial neuron thus accepts a signal then synthesizes it and signals the other neurons connected to it.

METHODS USED
Modules
This research was divided into two modules, Exploratory Data Analysis (EDA) and Predictive analysis.

The first module consists of two subparts.
i. Attempt to figure out "Which part of the brain is most affected in functioning due to Schizophrenia?" through analysis and visualization.
ii. Attempt to figure out the effect of Schizophrenia on humans of different age groups and comparing this to other age groups to identify patterns.

The second module consists of three subparts.
i. Prediction and analysis using Linear Regression.
ii. Prediction and analysis using Logistic Regression.
iii. Prediction and analysis using Artificial Neural Networks.

Tools Used
i. R Studio Version 1.2.5033
ii. Tableau 2019.1.0

Dataset Definition
The dataset was obtained from a publicly available source. It was published in a research paper "Did I Do That? Abnormal Predictive Processes in Schizophrenia When Button Pressing to Deliver a Tone" [15].

EEG data obtained from 32 control patients and 49 Schizophrenic was obtained and compiled. A simple button pressing task in which subjects either (1) pressed a button to immediately generate a tone, (2) passively listened to the tone, or (3) pressed a button without generating a tone to study the corollary discharge in people with schizophrenia and comparison controls.

It was found that comparison controls suppressed the N100, a negative deflection in EEG brain wave 100 milliseconds after the onset of a sound, when they pressed a button in order to produce a tone comparable to passive playback, however schizophrenic patients did not.

The dataset was split to make a training and testing set. Both sets contained control subjects and patients. The accuracy of neural networks algorithm was found to be the highest. The analysis was done in RStudio.

The analysis will help in better understanding of the effects on schizophrenia on the brain and pinpointing the specifics of the disease. It will also enable us to identify if a patient has schizophrenia by looking at the EEG report of a patient.

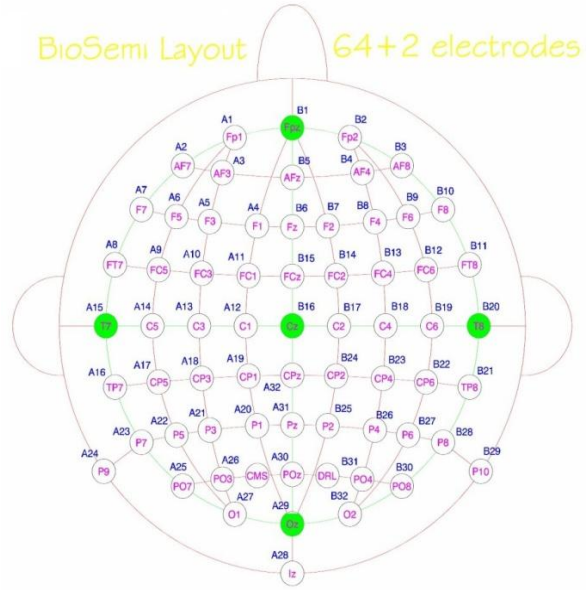

Figure 2. Electrodes layout

LITERATURE SURVEY

[1] The paper shows the capability of electroencephalogram data as a biomarker of the syndrome of schizophrenia. EEG data was recorded while schizophrenic patients viewed natural scenes. The average EEG activity at the viewing of the image was observed. After 500 ms of the viewing of the image, significant differences were found between the healthy controls and patients in occipital areas. The classifier was trained by using these differences to discriminate between the schizophrenic patients and the controls. The best classifier had a sensitivity of 81% for patient detection and 59% specificity for control detection. The overall accuracy was 71%. Statistical Analysis was used in this paper was done in MATLAB. We intend to use deep learning algorithms which have been proven to give higher levels of accuracy.

[2] They then assessed the within and between-site generalizability of a machine learning framework which exhibited excellent performance in an earlier study by using two different datasets of resting-state fMRI collected from various sites. They generated the generalizability in the main data through cross-validation. Then, the model which was trained on the main data set was tested by using the other data set using external validation. Finally, they recognized the poor between-site performance generalization, modified and enhanced the unsupervised pre-existing algorithm to check if transfer learning through additional unlabelled data could help to improve the performance. The main limitations of this paper are that it is based on functional MRI data, but is not unable to give satisfactorily high accuracy results which are required for all medical diagnosis.

[3] It was found that the schizophrenic group exhibited lower strengths of interregional connectivity as compared to the control group. There was irregular connectivity observed in both intra- and inter-hemispherical connections. The number of irregular connections in the left hemisphere was more than the right hemisphere in intra-hemispherical connections. They revealed that during exploring the large regions, serious discontinuities mostly occurred on occipital and temporal regions in big-region connections. While irregularity of connectivity was observed on the connection from temporal region to other regions in the connections in-big-region. This paper is focused on using exploratory analysis on fMRI data to understand the disease better. We go one step further and use it to train a prediction model so we can predict the results and thus help in diagnosis process. We also use EEG data as opposed to fMRI data for its merits of providing higher level of accuracy.

[4] A part of this paper is very valuable to us as it involves establishing the importance of using EEG data, and the relation of EEG data with mental health disorders like Schizophrenia. This paper provides detailed reasoning as to why EEG data can serve as biomarkers, and thus forms the platform for future studies to be done involving EEG data and mental health disorders. We have built on to this by using EEG data to train a predictive model by using ANN (Artificial Neural Networks) algorithm.

[5] This paper uses analysis of EEG data of patients of "Epileptic Seizures". It performs wavelet transform and classification using ANN algorithm and Logistic regression algorithm. Wavelet transform is specifically more efficient in visualizing different aspects of signals such as patterns, irregularities and trends than other approaches. By using wavelet transform method, transient features can be captured with minimum inaccuracy and made local in the context of both time and frequency. In this paper, ANN was applied to EEG data for detection and prediction of "Epileptic Seizures". This paper was successful and paved the way for similar analysis of different mental health disorders. Thus, we have built on this idea and used it for the disorder called "Schizophrenia"

[6] Using Random Forest algorithm, they initiated a very efficient classification model system, which could exclude a probable schizophrenic disorder while diagnosing a set of probably exposed population. Their model has followed the following steps for pre-processing: ICA and Spectral Analysis with normalization. After using Random Forest on this and validating it with different parameters and random states and a ten fold cross validation, they could exclude Schizophrenia 100%. The drawback as stated was that this model has not been tested on patients who might be suffering from other ailments.

[7] A SRC model was designed by the authors. Their aim was to pick out the single nucleotide polymorphisms (SNPs). It then put forward an SRC model and a sliding window model for fMRI data selection. Two types of data were then combined to classify schizophrenic patients and healthy patients by using LSV machine to help with a better diagnosis of schizophrenia. But the paper takes a small sample size and does not consist of a prediction mechanism.

[8] They analyse different ways of decomposition associated with the spectrum of EEG, measures that increment magnitude, time-frequency analyses and phase data from the EEG separately, consequences of the choice of parameters on the concluded and analysed findings. Afterwards, they compare the extent of phase locking in the gamma band response, which evokes about 50 ms and followed by presenting the standard tones in healthy and controls with schizophrenia. These procedures facilitate the naming which generally act as a source of indecision and hesitation in researching. These measures use the magnitude and phasing data formulated from detailed data extracted from the EEG.

[9] They propose to see ERPs and EEG data as the time/frequency irregularities of the actual processes. The process uses ICA, time/frequency analysis, and visualization. It attempts to measure EEG dynamic without the requirement of a model. The EEG observations focus either on the changes which are induced or on the peaks that are evoked during the experimental events. These measures, despite being almost complementary, cannot completely model all the event related dynamic of data, and is unable to isolate signals in the cortical area.

[10]On doing t-SPM, they saw that, on comparison, schizophrenics had greater slower activity in the parietal-occipital region and faster activity in the occipital regions. In contrast, alpha 2 activity reduced considerably in the occipital regions and this reduced activity encompassed large area of the head. These findings indicated cerebral hypofunction along with excitability in acute unprocessed schizophrenia. This is an earlier research paper on Schizophrenia and consists of primitive and inefficient methods.

[11]They analyse 40 controls and patients using EEG signals with the aim to classify them into groups. Techniques like LDA and AB were used as classifiers. An accuracy of 86-90% proves that EEG signals are efficient in classifying the patients and controls. This validation allows us to use EEG signals as the basis for our analysis.

[12]They analysed the transfer of information between cortical areas in case of a schizophrenic patient. Used 20 patients and controls. Estimated A-CMI to characterise dynamic cortical areas and establish AMI. This study shows hypo temporality and intra hemispheric over connectivity in patients. This proves the presence of abnormal lateralization in cortical areas in patients. We intend to expand upon this study.

[13] This study aimed to identify if quantitative EEG analysis can assess the cortical disturbances. They used Lempel-Ziv complexity under subjects at rest and at performing mental tasks. An additional group of depression patients was used along with patients and controls. They showed that schizophrenics have a lower LCZ than both depression and normal controls in both rest and mental task state. This proved the usefulness of EEG analysis in schizophrenic patients.

[14] They conducted a literature review to conclude if EEG abnormalities were consistent enough to be used as a diagnostic test for schizophrenia. A total of 53 papers were reviewed by the authors. The results showed that most of the abnormalities are replicated and the consistency of results in increased preponderance of rhythms. This strengthens the case for the use of EEG signals in schizophrenia diagnosis.

[15] The authors aimed to generalize patterns in communication to hear a tone due to a button press. 48 patients and controls were tested by comparison of N1 to tones sent to N1 of playback. Relation between N1 and LRP was assessed. Results suggested strong efference is associated with strong corollary discharge in both groups. This directly correlates with our task which is like this and validates the effectiveness of button pressing task to distinguish patients and controls.

SURVEY TABLE

Sl. No.	Topic	Techniques used	Limitations	Objective
1	"EEG classification during screen free-viewing for Schizophrenia detection"	MATLAB was used for all statistical analysis and classifier assessments.	Uses MATLAB which provides lower accuracy.	The data was used for statistical analysis and then used for training a classifier with an accuracy of 71%.
2	"Generalizability of machine learning for classification of schizophrenia based on resting-state functional MRI data"	MATLAB was used for performing ICA and applying LDA modeling.	It is based on functional MRI data, but is not unable to give satisfactorily high accuracy results which are required for all medical diagnosis.	ICA and LDA modeling were applied to the main set for internal validation. 73% accuracy was observed from internal validation. The resulting external validation was of an accuracy of 70%.
3	"Machine learning technique reveals intrinsic characteristics of schizophrenia: an alternative method"	MATLAB was used for applying "Kendall Tau method" for feature collection. T-test and Fisher's score was also applied.	fMRI has lower levels of accuracy as compared to EEG.	Pearson correlation, statistical parameter mapping (SPM), Kendall Tau method, T-test and Fisher's score were the techniques used. The study effectively proves that schizophrenic group shows a lower strength of inter-regional connectivity.
4	"Neurobiology of Schizophrenia: Electrophysiological Indices"	19- channel EEG for recording data, spectral analysis and microstate analysis	Establishes usefulness of EEG data but does not utilize the data.	Spectral analysis was done on the EEG data. Microstate analysis was also performed. The main conclusion drawn was the importance of the centroid of the brain due to its sensitivity to developmental EEG changes. The result showed that both the patients and the controls showed significantly different field configurations.
5	"Classification of EEG signals using neural network and logistic regression"	Wavelet transform analysis, LR and ANN	Done for Epilepsy. We implemented it for Schizophrenia. Very small amount of test subjects.	Wavelet transform method was used for capturing transient features and shows the energy distribution of the EEG signals. LR and ANN was used to train predictive modeling which provide a accuracy of 93% and the logistic regression-based classifier

				was accurate upto 89%.
6	"High-performance exclusion of schizophrenia using a novel machine learning method on EEG data"	Python 3- pandas, numpy, MNE and scikit-learn packages were used. ICA, Spectral analysis, and normalization was done. Random forest was the prediction algorithm used.	Not tested on patients with other disorders	Independent component analysis (ICA) was performed, followed with spectral analysis. Normalization was used for preprocessing the data Random forest algorithm was then applied. The overall rate of accuracy is 71.43%.
7	"Classification of Schizophrenia Patients with Combined Analysis of SNP and fMRI Data Based on Sparse Representation"	Sparse representation clustering (SRC) model, Linear support vector machines (SVM), Leave one out (LOO) cross validation method	Takes a small sample size and does not consist of a prediction mechanism.	LOO validation was performed using SVM. A combination of voxels and SNP subsets was normalized and used for classification and validation. Classification using only SNPs gave a 100% accuracy, whereas SVM using voxels gave a 92.5% accuracy.
8	"Event-Related EEG Time-Frequency Analysis: An Overview of Measures and An Analysis of Early Gamma Band Phase Locking in Schizophrenia"	Field-trip software in MATLAB for time-frequency analysis	Small Dataset and use of inefficient methods.	The standard tones were analyzed with a complex decomposition in MATLAB. The results clearly showed that the early-evoked gamma band phase is asynchronized in schizophrenic patients.
9	"Mining event-related brain dynamics."	ICA, time frequency analysis and trial-by-trial analysis of EEG signals.	Cannot completely model all the event related dynamics of data, and cannot isolate the signals in the cortical area.	The EEG observations focus either on the changes which are induced or on the peaks that are evoked during the experimental events.
10	"Computerized EEG in Schizophrenic patients."	T-statistical significance probability mapping (t-SPM).	Primitive and inefficient.	On doing the T-statistical significance probability mapping (t-SPM) schizophrenics had greater slower activity in the parietal-occipital regions, and faster activity (beta 1) in the occipital regions.
11	"Entropy and complexity measures for EEG signal classification of schizophrenic and control participants."	Linear discriminant analysis and adaptive boosting.	They have accurately been able to classify patients. We intend to predict on the basis of classifications.	Techniques like linear discriminant analysis and adaptive boosting are used as classifiers. An accuracy of 86-90% was achieved which was improved to 91-92% by computational adjustments.
12	"EEG in schizophrenic patients: mutual information analysis."	Analysis of A-CMI and auto mutual information (AMI).	Small dataset, primitive methodology and low accuracy.	Results: The patients had lower complexity than controls.
13	"Abnormal EEG complexity in patients with schizophrenia and depression."	Analysis of EEG data using Lempel-Ziv Complexity.	Low accuracy and tasks don't pertain to an external stimuli.	An additional group of depression patients was used along with patients and controls. They showed that schizophrenics have a lower LCZ than both depression patients and normal controls in both rest and mental task state.
14	"The status of spectral EEG abnormality as a diagnostic test for schizophrenia."	Verify Usefulness of EEG data as a diagnosis test for Schizophrenia.	Only a survey. Establish the use of EEG data but don't specify the areas.	A total of 53 papers were reviewed by the authors. The results showed that the abnormalities replicate and the consistency of results in increased preponderance of rhythms.

15	"Did I Do That? Abnormal Predictive Processes in Schizophrenia When Button Pressing to Deliver a Tone."	N1 suppression and Lateral Readiness Potential (LRP) amplitude relationship.	They don't use any ML algorithms and have a small sample size	The authors aimed to generalize patterns in communication to hear a tone due to a button press. 48 patients and controls were tested by comparison of N1 to tones sent to N1 of playback. Relation between N1 and LRP was assessed. Results suggested strong efference is associated with strong corollary discharge in both groups.

MODULES

Module 1

i. Attempt to figure out "Which area of the brain is most affected in functioning due to Schizophrenia?" through analysis and visualization.

Schizophrenic Patients

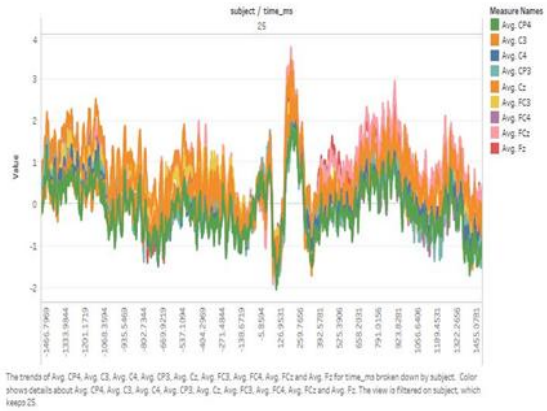

Figure 3. Distortion in EEG amplitude

Control Patients

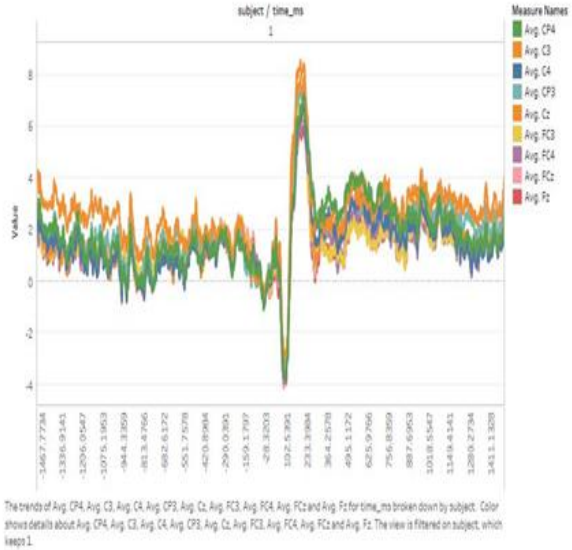

Figure 4. Distortion in EEG amplitude

Module 2

Attempt to figure out the effect of Schizophrenia on humans of different age groups and comparing this to other age groups to identify patterns.

Schizophrenic Patients

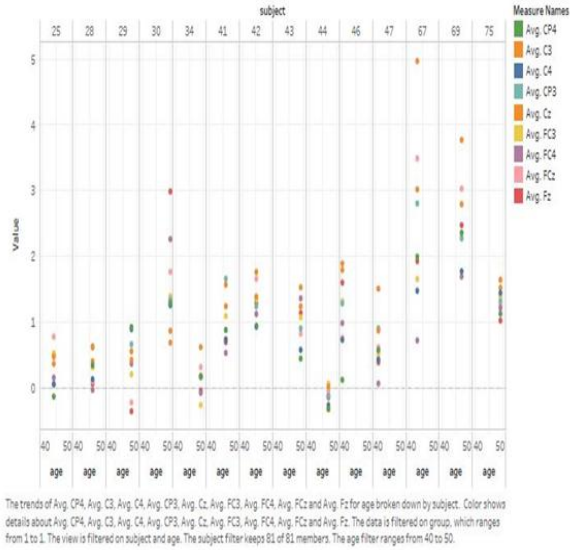

Figure 5. Average amplitude by age

Control Patients

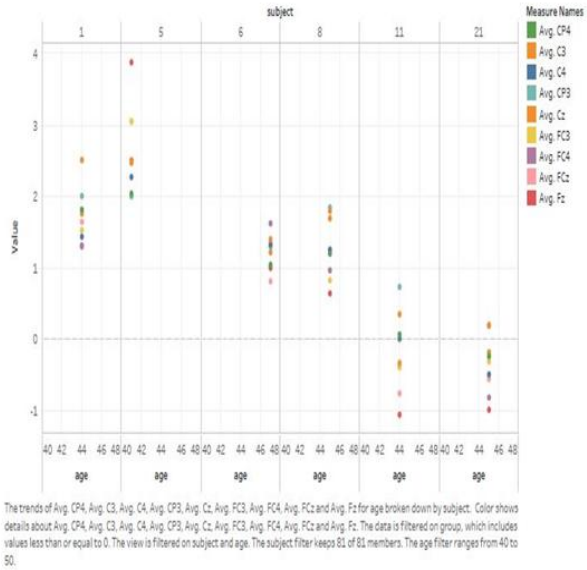

Figure 6. Average amplitude by age

Module 2

Attempt to predict on the basis of current test data, the presence of Schizophrenia in a subject by analyzing the EEG data for the subject.

Linear Regression
Model Summary

```
> print(summary(model))

Call:
lm(formula = group ~ Fz + FCz + Cz + FC3 + FC4 + C3 + C4 + CP3 +
    CP4 + time_ms, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3884 -0.1976 -0.1762 -0.1397  0.9908

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.853e-01  5.901e-04 313.961  < 2e-16 ***
Fz           1.256e-02  4.781e-04  26.276  < 2e-16 ***
FCz         -2.855e-03  7.711e-04  -3.703 0.000213 ***
Cz           1.872e-02  7.248e-04  25.825  < 2e-16 ***
FC3         -5.701e-03  7.385e-04  -7.719 1.17e-14 ***
FC4         -2.974e-02  6.997e-04 -42.499  < 2e-16 ***
C3          -2.075e-02  9.168e-04 -22.630  < 2e-16 ***
C4           1.673e-02  8.529e-04  19.612  < 2e-16 ***
CP3          2.119e-02  7.670e-04  27.632  < 2e-16 ***
CP4         -1.694e-02  6.600e-04 -25.671  < 2e-16 ***
time_ms     -2.282e-06  6.255e-07  -3.648 0.000264 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3868 on 522536 degrees of freedom
Multiple R-squared:  0.008719,  Adjusted R-squared:  0.0087
F-statistic: 459.6 on 10 and 522536 DF,  p-value: < 2.2e-16
```

Figure 7. Linear regression model summary

Plots

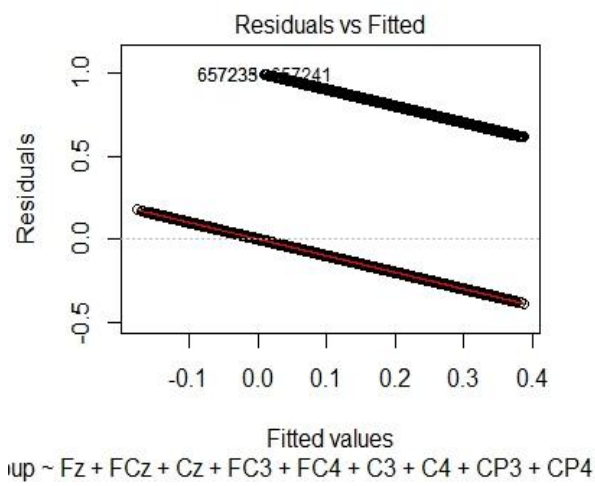

Figure 8. Fitted values vs Residuals(Linear regression)

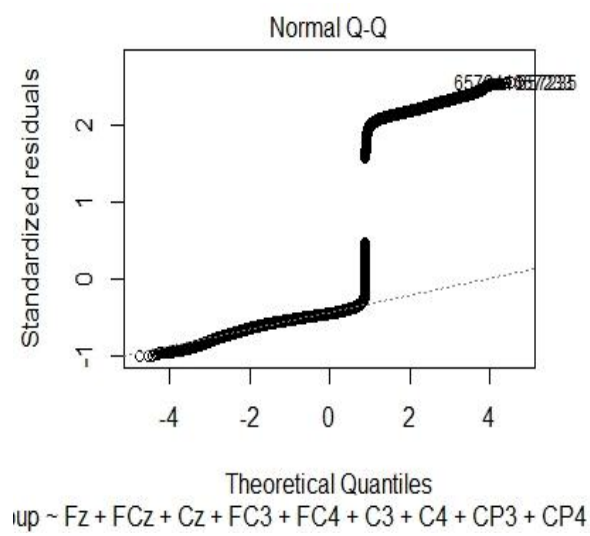

Figure 9. Residuals vs Quantiles

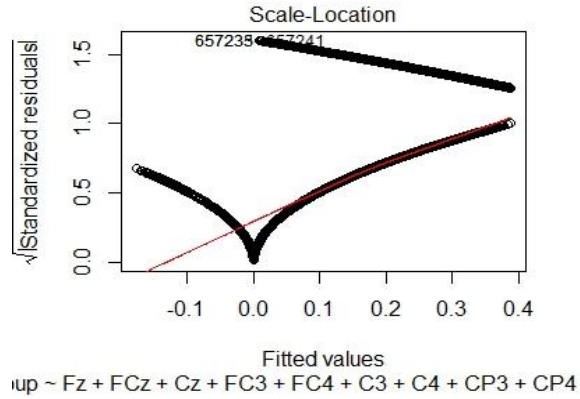

Figure 10. Standard residuals vs Fitted values

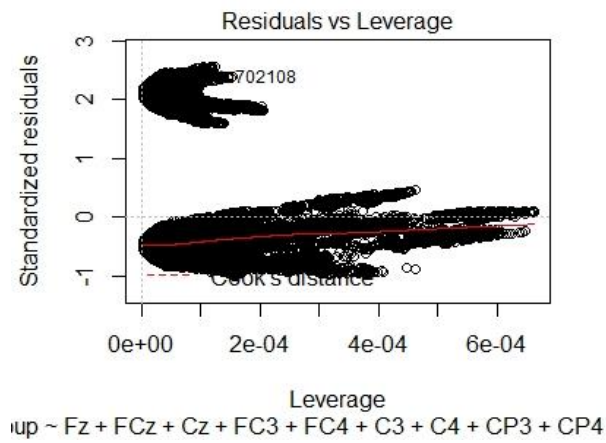

Figure 11. Standard Residuals vs Leverage

Linear Model Statistics

Residual Standard Error	0.3848 on 522536 degrees of freedom
Multiple R squared	0.008719
Adjusted R squared	0.0087
F-statistic	459.6 on 10 and 522536 degrees of freedom
p-value	<2.2e-16

Logistic Regression

```
> summary(log.model)

Call:
glm(formula = group ~ ., family = binomial(link = "logit"), data = train)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-1.457e-03  -2.000e-08  -2.000e-08  -2.000e-08   2.429e-03

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.106e+03  2.780e+03  -0.757    0.449
subject      3.168e+01  4.186e+01   0.757    0.449
condition   -2.822e-01  3.798e+01  -0.007    0.994
Fz          -9.216e-02  4.406e+01  -0.002    0.998
FCz          1.937e+00  6.380e+01   0.030    0.976
Cz          -5.691e-01  3.867e+01  -0.015    0.988
FC3         -1.016e+00  3.729e+01  -0.027    0.978
FC4          7.041e-01  3.668e+01   0.019    0.985
C3          -6.638e-01  3.415e+01  -0.019    0.984
C4          -1.525e+00  4.932e+01  -0.031    0.975
CP3          6.222e-01  2.961e+01   0.021    0.983
CP4          5.923e-01  3.581e+01   0.017    0.987
time_ms      1.632e-04  3.028e-02   0.005    0.996

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5.0077e+05  on 522546  degrees of freedom
Residual deviance: 1.7205e-03  on 522534  degrees of freedom
AIC: 26.002

Number of Fisher Scoring iterations: 25
```

Figure12. Logistic regression model summary

Logistics Model Statistics

Median Deviance	-2.000e-08
Null Deviance	5.0077e+05 on 522536 degrees of freedom
Residual Deviance	1.7205e-03 on 522534 degrees of freedom
Akaike information criterion	26.002
Model Accuracy	61.28%

Artificial Neural Networks

The accuracy of the model is 0.9009 and the loss was equal to 0.3243. Two layers of 5 and 1 perceptrons were used. Epoch was equal to 80 and the batch size was 20.

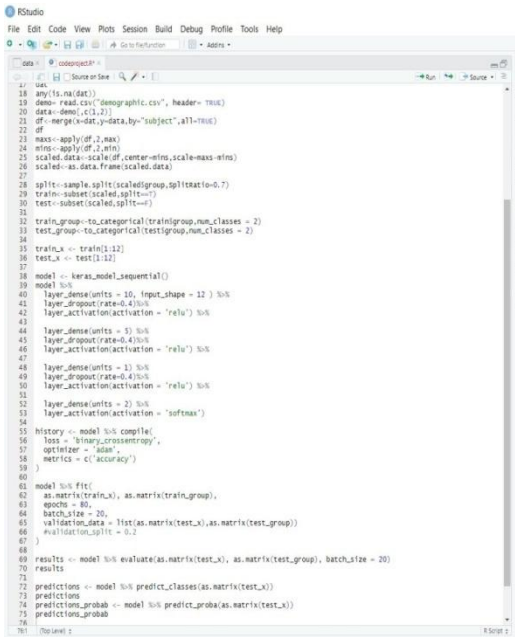

Figure 13. Sample code- ANN implementation

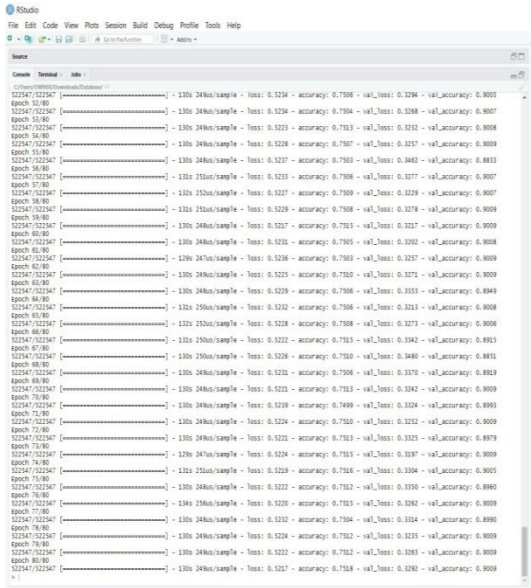

Figure 14. Screenshot of Epochs (ANN)

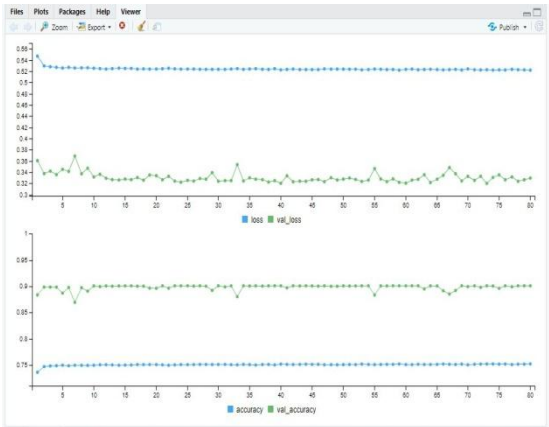

Figure 15. Graph of the model while training

Figure 16. Accuracy Screenshot of ANN model

Random Forest

The error rate of the classifier was 5.75%. The accuracy of the model was found to be 0.7103642. The number of trees was set to default to 500.

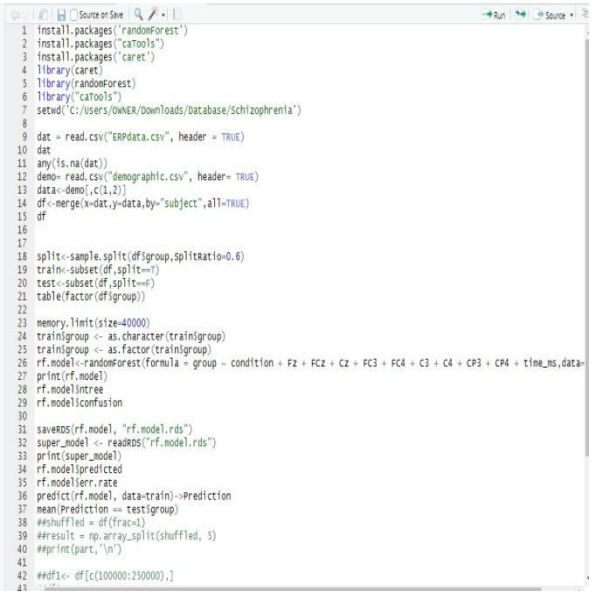

Figure17. Random Forests Code Screenshot

```
df<-merge(x=dat,y=data,by="subject",all=TRUE)
df
shuffled = df(frac=1)
result = np.array_split(shuffled, 5)
print(part,'\n')

df1<- df[c(100000:250000),]
df1
df2<- df[c(562178:709633),]
df2

split<-sample.split(df1$group,SplitRatio=0.6)
train<-subset(df,split==T)
test<-subset(df,split==F)
table(factor(df1$group))
table(factor(df2$group))
df1$group <- as.character(df1$group)
df1$group <- as.factor(df1$group)
rf1.model<-randomForest(formula = group ~ condition + Fz + FCz + Cz + FC3 + FC4 + C3 + C4 + CP3 + CP4 + time_ms,data
print(rf1.model)
df2$group <- as.character(df2$group)
df2$group <- as.factor(df2$group)
rf2.model<-randomForest(formula = group ~ condition + Fz + FCz + Cz + FC3 + FC4 + C3 + C4 + CP3 + CP4 + time_ms,data
print(rf2.model)
memory.limit(size=40000)
rf.model<-randomForest(formula = group ~ condition + Fz + FCz + Cz + FC3 + FC4 + C3 + C4 + CP3 + CP4 + time_ms,data=
print(rf.model)
train$group <- as.character(train$group)
train$group <- as.factor(train$group)
```

```
> memory.limit()
[1] 8075
> memory.limit(size=40000)
[1] 40000
> split<-sample.split(df$group,SplitRatio=0.6)
> train<-subset(df,split==T)
> test<-subset(df,split==F)
> train$group <- as.character(train$group)
> train$group <- as.factor(train$group)
> rf.model<-randomForest(formula = group ~ condition + Fz + FCz + Cz + FC3 + FC4 + C3 + C4 + CP3 + CP4 + time_ms,data=tra
in)
> print(rf.model)

Call:
 randomForest(formula = group ~ condition + Fz + FCz + Cz + FC3 +      FC4 + C3 + C4 + CP3 + CP4 + time_ms, data = train)

               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 5.75%
Confusion matrix:
       0      1 class.error
0 156250  20697  0.11696723
1   5035 265915  0.01858276
>
```

Figure 18. Error Rate Random Forest

```
> rf.model$ntree
[1] 500
> rf.model$confusion
       0      1 class.error
0 156250  20697  0.11696723
1   5035 265915  0.01858276
> predict(rf.model, data=train)->Prediction
> mean(Prediction == test$group)
[1] 0.7103642
warning messages:
1: In `==.default`(Prediction, test$group) :
  longer object length is not a multiple of shorter object length
2: In is.na(e1) | is.na(e2) :
  longer object length is not a multiple of shorter object length
>
```

Figure 19. Accuracy- random forest

RESULT FROM THE ANALYSIS

In our research, in the first module, it is clear that the patients of Schizophrenia and control patients can be easily distinguished on the basis of EEG data. The advent of Schizophrenia in relation to age is also shown in the second part of the first module.

In the second module, four different algorithms were applied to the dataset for classification and prediction. The accuracy of the logistic regression model is 61.28%. The Random Forest model has a low error rate of 5.75%, and a predictive accuracy of 0.7103642, a significant improvement from logistic regression. By far, the best results have been obtained from the Artificial Neural Network model, which has shown a high accuracy of 0.9009 and a low loss of 0.3243.

From our findings, we can sufficiently claim that EEG data can be used successfully for diagnosis of Schizophrenia and that the ANN predictive model provides the highest accuracy

CONCLUSION

From the descriptive analysis, we can sufficiently claim that EEG data can be used to distinguish between schizophrenic and control patients. Along with that there is also a proof seen that the advent of Schizophrenia affects the aged people more.

From the predictive analysis, we can understand that the although random forests help us with the classification well (5.75% error rate), it is not sufficiently accurate for predicting the onset of Schizophrenia. Neural Networks give us the best possible results. Thus, Neural Networks should be used for making predictions for further studies.

FUTURE WORK

One of the main aspects would be to externally validate the models by collecting data from various sources. This will help the model train better and thus produce better results when applied in the practical world.

Another important aspect is to understand the effects of other underlying diseases on the diagnosis of Schizophrenia, and what effect other diseases have which could lead to false diagnosis.

REFERENCES

1. EEG Classification during Scene Free-Viewing for Schizophrenia Detection. Christ Devia 2019.
2. Generalizability of machine learning for classification of schizophrenia based on resting state functional MRI data. Xin-Lu Cai, Kristoffer H. Madsen, Raymond C. K. Chan 2019.
3. Machine learning technique reveals intrinsic characteristics of schizophrenia: an alternative method. Junhua Li, Yu Sun, Rongjun Yu 2019.
4. Neurobiology of Schizophrenia: Electrophysiological Indices. Martha Koukkou, Thomas Koenig, Anja Bänninger, Kathryn Rieger, Laura Diaz Hernandez, Yuko Higuchi, Tomiki Sumiyoshi, Annarita Vignapiano, Giulia Maria Giordano, Antonella Amodio, Armida Mucci 2019.
5. Classification of EEG signals using neural network and logistic regression. Abdulhamit Subasi, Ergun Ercelebi 2019.
6. High-performance exclusion of schizophrenia using a novel machine learning method on EEG data. Ricardo Buettner, Michael Hirschmiller, Kevin Schlosser, Manfred Rössle, Marc Fernandes, Ingo J. Timm 2019.
7. Classification of Schizophrenia Patients with Combined Analysis of SNP and fMRI Data Based on Sparse Representation. Dongdong Lin, Hongbao Cao, Yu-Ping Wang, Vince Calhoun 2011.
8. Event-Related EEG Time-Frequency Analysis: An Overview of Measures and an Analysis of Early Gamma Band Phase Locking in Schizophrenia. Brian J. Roach, Daniel H. Mathalon 2008.
9. Mining event-related brain dynamics. Makeig S, Debener S, Onton J, Delorme A 2004.
10. Computerized EEG in Schizophrenic patients. Miyauchi T, Tanaka K, Hagimoto H, Miura T, Kishimoto H, Matsushita M 1990.
11. Entropy and complexity measures for EEG signal classification of schizophrenic and control participants. Malihe Sabeti, Serajeddin Katebi, Reza Boostani 2009.
12. EEG in schizophrenic patients: mutual information analysis. Sun Hee Naa, Seung-Hyun Jina, Soo Yong Kima, Byung-Joo Hamb 2012.
13. Abnormal EEG complexity in patients with schizophrenia and depression. Yingjie Lia, Shanbao Tong, Dan Liua, Yi Gaia, Xiuyuan Wang, Jijun Wang, Yihong Qiub, Yisheng Zhub 2008.
14. The status of spectral EEG abnormality as a diagnostic test for schizophrenia. Nash N. Boutros, Cynthia Arfken, Silvana Galderisi, Joshua Warrick, Garrett Pratt, William Iacono 2007.
15. Did I Do That? Abnormal Predictive Processes in Schizophrenia When Button Pressing to Deliver a Tone. Judith M. Ford, Vanessa A. Palzes, Brian J. Roach, and Daniel H. Mathalon 2013.