A Multi-channel EEG Biomarker Weighted Spectral Clustering Model for MDD Prediction

Shreeya Garg ¹ and Urvashi Prakash Shukla ²

¹Banasthali Vidypith
²Affiliation not available

October 30, 2023

Abstract

The article designs an approach for the detection of MDD using spectral clustering. The raw EEG is pre-processed, and then three quantitative biomarkers: band power (beta, delta, and theta) along with three signal-extracted signals: Detrended Fluctuation Analysis (DFA), Higuchi’s Fractal Dimension (HFD), and Lempel-Ziv Complexity (LZC) are extracted. The data is scrutinized in both inter and intra hemispheric regions. A weighted graph is constructed based on the data that is clustered to obtain the condition of the subject. The results showcase both the efficiency and efficacy of the designed approach. The accuracy achieved for the left hemisphere is 98% and CEP of 2.5%, while 97% accuracy and 3.3% CEP for the right hemisphere.

In this case, channel-wise maximum accuracy has been achieved for Fp1 and F8

Hosted file

A Multi-channel EEG Biomarker Weighted Spectral Graph Clustering Model for MDD Prediction.zip available at https://authorea.com/users/686031/articles/679553-a-multi-channel-eeg-biomarker-weighted-spectral-clustering-model-for-mdd-prediction
A Multi-channel EEG Biomarker Weighted Spectral Clustering Model for MDD Prediction

Shreya Garg*, Urvashi Prakash Shukla†
*Department of Electronics, Banasthali Vidyapith, India
†Department of Computer Science, Banasthali Vidyapith, India

Abstract

Major depressive disorder (MDD) is a global issue and every year the number of people suffering, is increasing at an alarming rate. The role of electroencephalography (EEG) in diagnosing MDD has shown a potential surge. Many studies have been carried out for designing an automated approach to the diagnosis of MDD through EEG as a primary tool of analysis. However, most of the methodologies depend on machine learning and the application of deep neural network tools. These heavily depend on the annotated EEG signals for training, which requires trained professional for data generation. In addition, the time and memory complexity of its implementations are huge. With these challenges, the article designs an approach for the detection of MDD using spectral clustering. The raw EEG is pre-processed, and then three quantitative biomarkers: band power (beta, delta, and theta) along with three signal-extracted signals: Detrended Fluctuation Analysis (DFA), Higuchi’s Fractal Dimension (HFD), and Lempel-Ziv Complexity (LZC) are extracted. The data is scrutinized in both inter and intrahemispheric regions. A weighted graph is constructed based on the data that is clustered to obtain the condition of the subject. The results showcase both the efficiency and efficacy of the designed approach. The accuracy achieved for the left hemisphere is 98% and CEP of 2.5%, while 97% accuracy and 3.3% CEP for the right hemisphere. In this case, channel-wise maximum accuracy has been achieved for Fp1 and F8.

Index Terms

Major depressive disorder (MDD), Electroencephalographic (EEG), EEG Biomarkers, Detrended Fluctuation Analysis (DFA), Higuchi’s fractal dimension (HFD), Lempel Ziv Complexity (LZC), Spectral Clustering.

I. INTRODUCTION

MDD is a complex and diverse mood disorder that inculcates the persistent feeling of sadness, loneliness, hopelessness, and the inability to connect with the outer world. This psychological effect is persistent for a protracted period, which eventually turns out into a serious diseases. Depression is the first rank mental disorder illness as per World Health Organisation (WHO) [1]. One out of every 15 adults gets affected by it every year. Nearly 264 million people worldwide are suffering from this silent killer. As per the National Mental Health Survey conducted in India; one in 20 Indians suffer from depression, that is nearly about 56 Million Indians.

MDD is coined rightly as “The Mental Health Epidemic” [2]. India has one of the highest numbers of cases of mental illnesses globally. Also, at present 14% of India’s population requires active mental health interventions. COVID-19 pandemic had a significant impact on the lives of millions around the world. The different measures imposed by the government like containment, social isolation to immobilize the virus, and the lockdown conditions. These conditions have led to the lack of physical access to friends, peers, limited opportunity for outdoor movement, work-personal imbalance. Eventually limited scope of privacy, have adversely affected the mental well-being of each individual. These factors have accelerated mental health instability. The researchers conducted a post-COVID-19 survey and reported that more than two-fifths of the population is suffering from anxiety and depression in India [3].

The diagnostic methods employed for MDD usually depend on the different scales or scores like the Montgomery-Asberg Depression Rating Scale (MADRS) [4], Hamilton depression rating scale (HAMD-17) [5], Patient Health Questionnaire-9 (PHQ-9) [6], and so on. These scales are heavily dependent on the physical therapist who conducts these assessments. It is a challenging task to facilitate these services in small towns and cities. As researchers, exploring methods that significantly reduce the burden and aid the process of detection of MDD automatically in its early stage has great significance.
In general, Electroencephalography (EEG) is incorporated for investigating brain activation, especially under physiological conditions [7]. EEG measures brain activity that predominately reflects the neuron dynamics between the nervous system and states of the brain [8]. It provides an insight into the functionality and dynamics of the brain [9]. These recordings serve as a crucial tool for studying the various functional states of the brain and thereby diagnosing and monitoring neurological diseases [10], such as epilepsy, insomnia, seizure, etc. The fast, accurate, cost-effective and reliable for analyzing the functionality of the brain has attracted researchers to explore EEG. These signals are non-stationary, with inherent non-linearity, and visually complex, making it imperative for fetching the features and need complex dynamic methods to extract the necessary clinch. The other challenges faced are the low magnitude range of the signal, which makes the signal vulnerable to noises (e.g., physiological artifacts or non-physiological artifacts).

Thus, different EEG-based biomarkers can effectively address the above issues, enabling to distinguish between the healthy and MDD subjects under the various treatment responses [11]. The majority of analysis carried out for the classification using Machine Learning (ML) approaches. The requirement of a large size dataset in ML approaches is needed to validate its success, so cross-validation techniques are applied in limited EEG data to estimate the performance of the proposed algorithm; which leads to overfitting and data leakage [12]. Thus, this shortcoming can be tackled by selecting a set of discriminate EEG features. The accuracy achieved by the deep learning approaches vastly relies on the abundance of labelled training data [13]. Labelling EEG signals requires the expertise of the pathologist and also is an extremely time-consuming procedure [14]. On the contrary, analysis of unlabelled data can be a boon. Clustering is a primitive way of exploring data with little or no prior knowledge [15]. Also, commonly known as unsupervised classification is applied to data analysis in absence of labels [16]. The goal of clustering algorithms is the separation of the dataset into finite and discrete sets, exploring the hidden data structure [17]. In this article, the author has evaluated the performance of the designed unsupervised method for decoding the primary screening of depression using an EEG signal. It is tested on a well-known public database. The analysis is carried out with the quantitative EEG biomarkers that can effectively mitigate the challenges and issues faced in operating on raw EEG signals. The framework design is a complete system from pre-processing of the data, followed by extraction of six minutely chosen EEG-based biomarkers to represent the signal effectively in both spatial and temporal information. This data is applied with weighted spectral graph clustering to decode the subjects into two categories. The complete outline of the designed framework is shown in Fig.1

Fig. 1: The complete proposed framework for detection of MDD.

The major contributions made are as follows:

- The framework is a complete unsupervised technique has been designed to predict MDD. This eliminates the need for labeling and direct dependency on the expert. Also, the time and computational complexity are greatly reduced. The reliance on the large data set to address the issues like overfitting and data leakage is effectively reduced. The framework can easily be integrated to identify new or hidden trends in the other mental issues.
The designed framework relies on the derived Quantitative EEG biomarkers and the signal extracted features. The complexity associated with operating on a direct EEG signal is meritoriously addressed. The neural information captured from different brain sites at a different frequency can be quantified into cortical activity through band power. Also, non-linear and non-stationary characteristics of EEG can be effectually investigated using signal extracted features.

A comprehensive analysis is done for understanding the relation and inter-dependency between the inter-hemispheric and intra-hemispheric electrode pairs. Also, the role of the individual channel is exploited to comprehend spatial and temporal reliance.

The time and computational complexity of the designed framework is less as compared to the existing deep neural network method. The overall computational complexity of the spectral clustering is $O(C^3)$, where $C$ is the number of channels in the study.

The organization of the paper is as: An extensive survey of various methods employed for MDD prediction is presented in Section II. The complete overview of the designed framework is elaborated in Section III. Section IV, provides the experimental setup, along with an in-depth discussion on the results achieved. Followed by the conclusion and future scope in Section V.

II. LITERATURE SURVEY

The behavioral symptoms majorly depend on the age and history report of the patient that are not present in most of the cases. This affects the diagnosis process and thereby prevents the early treatment of the person. Thus, the objective neuroimaging biomarkers (or “neuro markers”) have the advantage of being clinical unibias. This motivates the researchers to utilize EEG data of a subject as a tool to screen various mental disorders such as Alzheimer’s [18], depression [19], anxiety [20], seizures [21].

The majority of the work were carried out on analyzing depression through EEG signals, has generally incorporated training models either through raw, preprocessed EEG signals, or by the method of features extraction. A DeprNet [22], CNN model based on resting-state EEG recordings with a sample size of 33 subjects (18 Normal and 15 MDD). The maximum accuracy attained by this convolution model is 99% record-wise split data and 91% subjectwise split data. Similar works based on 30 subjects were designed for single-channel EEG data from the right and left hemispheres of the brain and tested on CNN-LSTM [23]. The author reported major drawback was computational complexity though the accuracy achieved was 97.66%. On the same dataset, [24] proposed deep learning CNN model with 13 convolution layers with an accuracy of 95.49%. A more efficient and model with less complexity is reported in [25] with 6 CNN layer structure. A complete survey on various deep neural network strategies applied is studied in [26].

The next set of research works reported is based on exploring the raw EEG data for various linear or non-linear features for the detection of MDD. A study on 213 subjects (121 Normal and 92 MDD) [27], was conducted on resting-state EEG signals. The features extracted included time-domain features (peak variance, skewness, kurtosis, and Hjorth parameter), frequency-domain features (centroid frequency, relative and absolute power), and non-linear features such as C0-complexity, Kolmogorov Entropy, Shannon Entropy, Power-Spectral Entropy, etc. were comprised for the training of classification model, namely, KNN (K-nearest Neighbour), SVM (Support Vector Machine), Classification Tree (CT), and Artificial Neural Net (ANN). The accuracy reported for the four different classification algorithms was approximately in the range of 67% to 74%.

An inclusive survey on the various biomarkers, engaged in the extraction from an EEG signal, and the different trends for various depressive disorders are described in [28]. Another study [29], with fifteen non-linear features such as Fractal Dimensions (FD), Detrended Fluctuation Analysis, Hurst’s Exponent, etc, presented the Depression Diagnosis Index (DDI) that helped the detection of depression based on a single numerical value, calculated with the features, ranked highest and are common to both the hemispheres. Further, the author also discussed five different classification models - SVM, KNN, NB, PNN, and Decision Tree, attaining the highest classification accuracy of 98%.

A method designed in [30], used the power of the signal, DFA, HFD, and correlation dimensions as the four EEG biomarkers for the training classification model. In [12], extensively explored the band power of an EEG signal and Fractal dimensions (HFD and KFD) over a comparatively large dataset of 400 subjects (200 Normal and 200 MDD). A comprehensive survey of all existing work especially models designed on the quantitative biomarkers EEG and features extracted are tabulated in TABLE I.
TABLE I: A comprehensive survey on all recent existing works done using Quantitative EEG Biomarkers.

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Dataset</th>
<th>Features</th>
<th>Model Description</th>
<th>Accuracy (%)</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seal A, et al. [22]</td>
<td>2021</td>
<td>33</td>
<td>18 15</td>
<td>Raw resting - stage EEG data, filtered HPF (0.1-100 Hz), 50Hz notch filter</td>
<td>CNN-DeprNet( 18 layers)</td>
<td>91.40&lt;br&gt;Training Time = 37.6 mins and Testing Time 1 sample is 0.036 sec, the system is designed for 19 channels cannot be utilized for 1 channel</td>
</tr>
<tr>
<td>Chien Te Wu, et al. [12]</td>
<td>2021</td>
<td>400</td>
<td>200 200</td>
<td>EEG spectral bandpower, Coherence, HFD, Katz’s Fractal Dimension (KFD)</td>
<td>Classifiers : KNN, Linear discriminant analysis (LDA), SVM, CK-SVM</td>
<td>84.16&lt;br&gt;Efficiency achieved by CK-SVM with 51 and 62 feature set is 84.16% and 80.83% respectively</td>
</tr>
<tr>
<td>Ay B, et al. [23]</td>
<td>2019</td>
<td>30</td>
<td>15 15</td>
<td>Raw EEG signals</td>
<td>CNN-LTSM model (Left and Right Hemisphere)</td>
<td>99.12&lt;br&gt;1 epoch = 52 sec (Right Hemisphere Training)</td>
</tr>
<tr>
<td>Cai H, et al. [27]</td>
<td>2018</td>
<td>213</td>
<td>92 121</td>
<td>C0-complexity, Kolmgorov Entropy, Shannon Entropy, Correlation Dimension, Power Spectral Entropy</td>
<td>Classifier : SVM, KNN, CT, ANN</td>
<td>79.27&lt;br&gt;Highest accuracy claimed with features extracted is 79.27% is with bandpower in KNN</td>
</tr>
<tr>
<td>Bachmann M, et al. [31]</td>
<td>2018</td>
<td>26</td>
<td>13 13</td>
<td>Spectral assymetry index, Alpha Power Variability, Relative Gamma Power, HFD, DFA, LZC.</td>
<td>Logistic Regression</td>
<td>88.00&lt;br&gt;Number of subjects taken in account for the analysis was small and might hinder the accuracy achieved, when incresed.</td>
</tr>
<tr>
<td>Mumtaz W, et al. [32]</td>
<td>2017</td>
<td>64</td>
<td>30 34</td>
<td>Wavelet Transformation(WT), STFT and EMD, Coherence</td>
<td>Logistic Regression (LR)</td>
<td>87.50&lt;br&gt;For each subset= 100 times algorithm (Runs)</td>
</tr>
<tr>
<td>Mumtaz W, et al. [33]</td>
<td>2016</td>
<td>63</td>
<td>30 33</td>
<td>EEG spectral bandpower and frontal alpha asymmetry</td>
<td>Classifiers - LR, SVM, Naive Bayesian(NB)</td>
<td>98.40 For each subset= 100 times algorithm (Runs)</td>
</tr>
<tr>
<td>Acharya, et al. [29]</td>
<td>2015</td>
<td>30</td>
<td>15 15</td>
<td>Fractal Dimensions(FD), Largest Lyapunov Exponent (LLE), Detrended Fluctuation Analysis(DFA), Sample Entropy, Hurst’s Exponent, High Order Spectra and Recurrence Quantification Analysis (RQA).</td>
<td>Classifiers - SVM, KNN, NB, PNN and Decision Tree(DT) (Left and Right Hemisphere)</td>
<td>98.00&lt;br&gt;Training time= 199.07 sec (1 epoch Left Hemisphere) and 198.68 sec (1 epoch Right Hemisphere)</td>
</tr>
<tr>
<td>Hosseinifard B, et al. [30]</td>
<td>2013</td>
<td>90</td>
<td>45 45</td>
<td>Bandpower, DFA, HFD, Correlation Dimension, LLE.</td>
<td>Classifiers : LDA, LR, KNN.</td>
<td>90.00&lt;br&gt;LR classifier gave better accuracy as compared to KNN and LDA classifiers.</td>
</tr>
<tr>
<td>Ahmadlou M, et al. [34]</td>
<td>2012</td>
<td>24</td>
<td>12 12</td>
<td>KFD, HPD in frontal brain, Wavelet analysis</td>
<td>Enhanced probabilistic neural network (EPNN)</td>
<td>91.30&lt;br&gt;EPNN are slow and require more memory space</td>
</tr>
</tbody>
</table>

III. METHODOLOGY

The designed unsupervised framework is needed as the fact that EEG data varies from person to person, also there is variability observed between channels with the non-stationary characteristics of the signal. Thus, to combat all the shortcomings author has designed this framework based on the EEG biomarkers and signal extracted features.
A. Preprocessing

EEG signals are mainly affected by minor eye movement, chewing, muscular, or also because of the electrode signal interference. These signals along with the main recordings get captured. This unwanted voltage is known as artifacts. To remove these artifacts, the signal is sampled at 256Hz. Any sudden upshoot or spike is filtered using a Discrete Fourier Transform (DFT) filter using a hamming window as defined mathematical in Eq.1. Filter frequency range is (0.5-40 Hz) with one-pass, zero-phase, and non-casual.

\[
\omega(n) = \begin{cases} 
\alpha_0 - \alpha_1 \cos \frac{2\pi n}{N} & 0 \leq n \leq M \\
0 & \text{otherwise}
\end{cases}
\]  

(1)

Here, \(n\) are the samples of EEG discrete signal with a window size of \(M\) for total \(N\) samples over which the DFT has been performed in one recording of a subject. In the equiripple sense, the optimal values for the coefficients are \(\alpha_0 = 0.53836\) and \(\alpha_1 = 0.46164\).

B. Biomarkers Extraction

Quantitative analysis of EEG can be effectually characterised through various biomarkers. A parametric analysis for the complete process implemented is presented with their respective values have been explained in TABLE II.

- **BANDPOWER**

EEG signals are random or can be said in-deterministic in nature. Thus, approximation of band power of an EEG signal cannot be performed using conventional methods. Therefore, power spectral density is calculated on the small chunks of data assuming the signal to be stationary in that segment through a windowing technique. The band power of an EEG signal is normally calculated by segregating the signal into its major frequency component which is generally coined as the relative bandpower of that frequency band. The indepth analysis of EEG spectrum, can be precisely done by employing Welch’s method as it uses Fast Fourier Transform (FFT) as it is a computationally efficient method. Through the spectral analysis the goal is to decompose the EEG signal into a weighted sum of sinusoids of each EEG segment (or windowed chunk). This allows a easy assess the frequency content of EEG signal.

Hence, Welch’s estimation of Power Spectral Density (PSD) is given as:

\[
S_x(\omega_n) \triangleq \frac{1}{M} \sum_{k=0}^{M-1} P_{x,k}(\omega_n)
\]  

(2)

Here, \(P\) is the periodogram for the \(k^{th}\) window. Through the PSDs calculation for each channel under observation, relative bandpower of any frequency component can easily be estimated as [35]:

\[
RP = \frac{\int_{\text{Area}} [S_x(\omega_n)]_{\text{frequency band}}}{\int_{\text{Total Area}} [S_x(\omega_n)]}
\]  

(3)

Fig. 2: Comparative trends observation in different bandpower in MDD and Normal Subjects (a) Beta (b) Delta (c) Theta.

In this study, three frequency bands - beta, delta and theta, have been analyzed through their respective relative bandpower. The observations made from the Fig.2, suggest that for MDD subjects, Beta, Delta and Theta bands
are significant discriminators as compared to alpha band. Increased Delta bandpower in right hemisphere and increased Beta bandpower have been observed in MDD subjects while same trend has been observed with Theta band. A similar trend has been reported in [33], for theta and alpha bands in MDD subjects.

**SIGNAL EXTRACTED FEATURES**

EEG signal, owing to its non-stationary, non-linear, and fluctuating behaviour, accommodates various statistical and non-linear features describing its degree of complexity or randomness, dimensions and noisy behaviour. The following are the signal extracted features taken to study the characteristics of the raw EEG signal.

- **Detrended Fluctuation Analysis**: Signals or objects can be considered self-similar if there are parts that are exactly or approximately similar to the object or the part are statistically similar at different scales. With this knowledge we can easily formulate DFA, as a scaling analysis that we use to compute long-range correlation parameter of a non-stationary, fluctuating signal. Complete DFA analysis of a noisy EEG signal has been explained in Algorithm 1.

\begin{algorithm}
\begin{enumerate}
\item **Input**: EEG time-series data \(x = \{x_1, x_2, ..., x_N\}\).
\item Secondary series \(y\), called the detrended data by removing the mean \(\bar{x}\).
\item Divide \(y(r)\) in non-overlapping segments of equal length \(s\).
\item Compute least-square fit line \(y_n(r)\) (local trend), for \(n^{th}\) window.
\item Compute the detrended time-series \(Y_s(r)\) as the difference \(y(r) - y_n(r)\).
\item Calculate variance \(F_s^2(n)\) for all the \(2N_s\) segments of \(Y_s(r)\).
\item **Output**: Deterrended Fluctuation \(F(s)\) of the given EEG time-series data,
\end{enumerate}
\end{algorithm}

\[ F(s) = \sqrt{\frac{1}{2N_s} \sum_{n=1}^{2N_s} F_s^2(n)} \]  

In the present study, the window length \(s\) varies dynamically in logarithmic fashion with changing input size. Moreover, the detrended data computed was fitted using first degree polynomial. Further parametric analysis of DFA is discussed in TABLE II.

- **Higuchi’s Fractal Dimension**: Fractal dimension (FD) is the statistical parameter of complexity, that compares the change in details in a complex time series, as the measuring scale changes. Higuchi’s Fractal dimension is one of the method of calculating dimension of a time series fractal curve [36]. EEG signals, owing to its continuous, non-stationary and fluctuating behaviour can extensively be considered as a fractal curve. Application of Higuchi’s algorithm on EEG signals has been explained in Algorithm 2.

\begin{algorithm}
\begin{enumerate}
\item **Input**: EEG time-series data \(x\).
\item Set \(k_{max}\), compute new sub series \(y_b^a\), such that \(b \in \{1, ..., k_{max}\}\) and \(a \in \{1, ..., b\}\).
\item Compute length \(l_a(b)\) of each new time series, \(y_b^a\).
\item Compute length of the curve, \(l(b)\)
\item Calculate slope \(H\) of the best-fit straight line, passing through the points, \((ln(1/b), ln(l(b)))\).
\item **Output**: Higuchi’s Fractal Dimension \(H\).
\end{enumerate}
\end{algorithm}

In presented study, HFD values for MDD and Normal subjects, obtained after averaging the channel-wise values for a patient, showed minimum divergence. The value of \(k_{max}\), was changed dynamically according to the length of the data array being passed for each subject. For a particular subject, HFD values were calculated for all the 20 channels and then bifurcated in differed Data Matrices \(D\) explained in Section IV.

- **Lempel-Ziv Complexity**: The randomness or complexity of a EEG signal exists as a result of one’s scalp activity, it can be determined
through Lempel-Ziv Algorithm [37] explained in Algorithm 3, works on binary sequences. It links the complexity to the rate of occurrence of the distinct sub-sequences along the given sequence. The given EEG channel data \( x \), for each patient is first modified in a binary sequence \( P : \{p_1, p_2, p_3, ..., p_n\} \) as per the rule

\[
p_i = \begin{cases} 
0, & \text{if } x_i < \Theta, \\
1, & \text{otherwise}
\end{cases}
\]  

(5)

where, \( \Theta \) is selected as the median of the EEG data \( x \). [38], [39], [40]

**Algorithm 3 Lempel-Ziv Algorithm**

1: **Input**: Binary Sequence \( P \).
2: Let SQ be the two sub-sequences of \( P \) and \( SQ_n \) is generated from SQ, after omitting the last term.
3: Take a complexity counter \( c(n) \) and set it to 1 and scan \( P \) for new sequences.
4: If \( Q \notin v(SQ_n) \), increment \( c(n) \) and renew \( S \) by combining \( S \) and \( Q \), else renew \( Q \) while keeping the \( S \) unchanged and check again.
5: Normalize \( c(n) \) as per [37], [38] to compute \( C_{LZ} \).
6: **Output**: LZC value of the given sequence, \( C_{LZ} \).

![Fig. 3: Comparative trends observation in different signal feature extracted in MDD and Normal Subjects (a) DFA, (b) HFD (c) LZC.](image)

DFA has been seen with a potential of discriminator, along with HFD and LZC. Lower values of DFA are observed for depression along with a significant dip visible at Channels C3 and Pz. This is evidently from the Fig.3a.Similar conclusion have also been drawn in study carried out in [41]. An opposite trend is examined in HFD, an increased value is seen in MDD subjects as seen in Fig.3b. Although a spike at channels C3 and T3 is seen in Fig.3c.A same line of examination has been carried out in study done in [42].

In the next step, after the various EEG biomarkers and signal extracted features are calculated. These are stored in various DataMatrix like hemisphere wise, channel wise, and mean. Then, a \( L_2 \) normalization is carried out to provide an unbiased dataset to the designed clustering model.

**C. Data Clustering**

Application of graph approach has been widely used in identification and characterization of various abnormalities including MDD. This helps in defining brain in terms of specific regions (electrodes) and their corresponding connections (the weighted edges) [43]. This provides an additional avenue for exploring connections across the brain. Spectral clustering uses the first few singular vectors of the Laplacian matrix \( L \), which are the inherent global quantities and the local sensitive information may be rendered unattended. Thus, the clusters formed after spectral clustering are re-clustered only once to generate the final clusters. A undirected weighted graph \( G = (V, E) \) with vertex set \( V = C_1, ..., C_{20} \) is constructed in our case. The weights are non-negative thus, \( W_{ij} > 0 \). The weighted adjacency matrix \( W = (w_{ij})_{i,j} \). This relation value \( W_{ij} = 0 \) in case there is no correlation found between the channels from the derived biomarkers [44]. The authors have adopted \( \epsilon \) neighbourhood approach for constructing a similarity matrix. Then, the normal steps are executed to obtain the unnormalized spectral clustering as given in Algorithm 4.
<table>
<thead>
<tr>
<th>Procedure</th>
<th>Parameters</th>
<th>Value</th>
<th>Equations</th>
<th>Package used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preprocessing (FIR filter and Normalization)</td>
<td>Window</td>
<td>Hamming Window</td>
<td>(1)</td>
<td>MNE - python</td>
</tr>
<tr>
<td></td>
<td>Stopband attenuation</td>
<td>33dB</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Low passband edge</td>
<td>0.5Hz</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>High passband edge</td>
<td>40Hz</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Filter length</td>
<td>1691 samples (6.605sec)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative Bandpower</td>
<td>Sampling frequency</td>
<td>256Hz</td>
<td>(2), (3)</td>
<td>YASA</td>
</tr>
<tr>
<td></td>
<td>Window size</td>
<td>4 seconds</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Method for averaging periodogram</td>
<td>Median</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bands</td>
<td>Delta, Theta, Beta</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detrended Fluctuation Analysis (DFA)</td>
<td>Minimum window size</td>
<td>4 seconds</td>
<td>Algorithm 1</td>
<td>eeglib</td>
</tr>
<tr>
<td></td>
<td>Maximum window size</td>
<td>signal size/4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximum number of window sizes</td>
<td>log2(size)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fit-degree</td>
<td>1 (default)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Higuchi’s Fractal Dimension (HFD)</td>
<td>$k_{max}$</td>
<td>By default window size/4</td>
<td>Algorithm 2</td>
<td>eeglib</td>
</tr>
<tr>
<td>Lemple-Ziv Complexity (LZC)</td>
<td>Threshold</td>
<td>median of the data</td>
<td>Algorithm 3</td>
<td>eeglib</td>
</tr>
<tr>
<td>Spectral Clustering</td>
<td>Number of clusters</td>
<td>2</td>
<td>Algorithm 4</td>
<td>MATLAB 2021a</td>
</tr>
<tr>
<td></td>
<td>Number of iterations</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-means</td>
<td>Number of clusters</td>
<td>2</td>
<td>(6)</td>
<td>MATLAB 2021a</td>
</tr>
<tr>
<td></td>
<td>Number of iterations</td>
<td>1000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE II: Parametric Overview of the Complete Designed Framework.

**Algorithm 4** Unnormalized Spectral Clustering

1: **Input**: DataMatrix $D$, with channels $C$ and Biomarkers $B$, the $\epsilon$
2: Construct a similarity matrix $W_{ij} \in R^{C \times C}$
3: Compute the unnormalized Laplacian $L$
4: The first $z$ Eigenvectors $u_1, \ldots, u_z$
5: Let $U \in R^{C \times z}$ be the matrix containing the eigenvector columns
6: Cluster these columns using K-means algorithm

$$K_j = \arg i = \min||W - \mu_j||^2$$

7: **Output**: Clusters $Z_1$ and $Z_2$

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Environmental Settings

In this study, the MNE-Python package [45], eeglib package [46], and Yet Another Spindle Algorithm (YASA) [47] packages in the Google Colab platform are used with Python 3.6.9 for the pre-processing and extracting the necessary features from the raw EEG signals. While the clustering part has been implemented on the MATLAB 2021a. The system details are tabulated in Table.III.

B. Dataset

The dataset used in this article is collected by Mumtaz in this study [48]. It consists of in total of 120 subjects with 62 MDD patients and the rest 58 are healthy. The age group is between 12 to 77 years. It was gathered from
**Table III: System Specifications.**

<table>
<thead>
<tr>
<th>Name</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>System RAM</td>
<td>8GB</td>
</tr>
<tr>
<td>CPU</td>
<td>Intel(R) Core(TM) i7 CPU 3630QM @ 2.70 GHz</td>
</tr>
<tr>
<td>Graphic Processor</td>
<td>Intel(R) HD Graphics</td>
</tr>
<tr>
<td>Platform</td>
<td>Google Colaboratory and MATLAB 2021a</td>
</tr>
<tr>
<td>Language</td>
<td>Python 3.6.9</td>
</tr>
</tbody>
</table>

Hospital Universiti Sains Malaysia (HUSM). The data is collected using 20 channels. The authors have utilized EEG dataset with eyes closed (EC) and eyes open (EO) only. To collect the data participants were requested to keep their eyes closed for 5 minutes with minimal head movement and blinking.

**C. DataMatrix Formation**

Data matrices of dimension $120 \times 6$ were formed containing feature data for all the subjects. EEG files acquired in .edf format, containing sampled data of each of the 20 channels were transformed into the form of a matrix and normalized for feature extraction as described in Section.III. For analysis, three separate data matrices are formed - mean data, channel-wise data, and hemispherical data. Channel-wise data is analyzed in 20 different matrices, one
Fig. 6: Correlation analysis among the various features extracted in (a) Channel-wise (Fp1) and (b) Mean datamatrix.

Fig. 7: Clustering results for MDD detection of the Left hemisphere datamatrix for actual labels and different methods employed.

for each channel, containing feature data of all the 120 subjects for that respective channel. The mean data matrix is formed by averaging the feature data of a subject for all the channels and tabulating the acquired data for all the subjects. The mean data, in particular, showed a minimal difference between two subjects, whether depressed or non-depressed, this minute difference is challenging for especially the unsupervised approaches like clustering. Hemispherical data is further taken into consideration for the analysis of changes occurring in the two hemispheres. Data Matrices for both the hemispheres are formed by averaging each subject channel-wise data ($F_{p1}-T_{5}$) for the left and ($F_{p2}-T_{6}$) for the right, assuming EEG channels are labelled according to the 10-20 system as shown in Fig. 4.

D. Parameters for Analysis

The evaluation of clustering algorithms heavily depends on the characteristics of input datasets. Thus, to decipher the clusters formed in a cohort for the given dataset, clustering validity indexes are incorporated with two important characteristics: compactness and separability. The datamatrix ($D$) for any of the three cases as defined as described in Subsection.C. This data is clustered into two $Z_1$ and $Z_2$. The evaluation of these clusters is done using the following indices.

- **Rand Index (RI)**
  The Rand measure or more popularly known as Rand Index [49] it measures the relationship between points in the datasets rather than exploring relationship of comparison between the true labels and assigned. It can
Fig. 8: Clustering results for MDD detection of the Right hemisphere datamatrix for actual labels and different methods employed.

Fig. 9: Box-and-whisker diagram for exploring the biomarkers distribution in (a) mean datamatrix, (b) Left and (c) Right hemisphere of the brain.

be used in absence of labels also, thus suitable for unsupervised methods. Eq.16 describes intuitively the relationship among the $T_p + T_n$ as number of agreements and $F_n + F_p$ are the number of disagreements between the the two clusters formed.

$$RI = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$  \hspace{1cm} (7)

Rand index is the frequency of agreements over the total pairs matched. Its value lies between $RI \in [0, 1]$ , with 0 indicating that the clustering achieved is not identical to that with the labels. The value more towards 1 suggest the more similarity.

• **Adjusted Rand Index**

$RI$ suffers from the problem of instability when the data is partitioned into two clusters. In order to solve this problem Adjusted Rand Index was introduced it considers a random model with generalized hypergeometric distribution. The major advantage of using $ARI$ is, there are no assumptions made on the cluster structure and can be used effectively and efficiently to compare the results of K-means (normally assumes isotropic blob shapes) with results of spectral clustering algorithms which can find cluster with “folded” shapes. $ARI$ is mathematically stated as in Eq.17 where $X = [(T_p + F_p)(T_p + F_n) + (F_n + T_n)(F_p + T_n)]$
\[ ARI = \frac{\binom{N}{2} (T_p + T_n)) - X}{\binom{N}{2} - X} \]  \hspace{1cm} (8)

- **Classification Error Percentage (CEP)**
  The number of labels mis-classified by the designed model for a given clustering algorithm is given by CEP. The label assigned by the clustering algorithm is compared with the desired output. Mathematically, CEP is defined as in Eq. 18 [50]

\[ CEP = \frac{(T_p + T_n) - (F_n + F_p)}{N} \]  \hspace{1cm} (9)

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Time (sec)</th>
<th>Testing time of one sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-DeprNet [22]</td>
<td>2256</td>
<td>3.600 msec</td>
</tr>
<tr>
<td>CNN-LTSM [23]</td>
<td>52</td>
<td>5.000 msec</td>
</tr>
<tr>
<td>S-SVM [51]</td>
<td>21888</td>
<td>15.6515 sec</td>
</tr>
<tr>
<td>H-KNN [27]</td>
<td>36956</td>
<td>13.0665 sec</td>
</tr>
<tr>
<td>H-DBN [52]</td>
<td>29604</td>
<td>9.6442 sec</td>
</tr>
<tr>
<td>S-EMD [53]</td>
<td>18920</td>
<td>5.6340 sec</td>
</tr>
<tr>
<td><strong>PROPOSED MODEL</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-means Clustering</td>
<td>0.0</td>
<td>1.392 msec</td>
</tr>
<tr>
<td>Spectral Clustering</td>
<td>0.0</td>
<td>2.005 msec</td>
</tr>
</tbody>
</table>

**TABLE IV:** Comparative study of time complexity with the other existing methods in literature employed on the same dataset.

**E. Results and Discussion**

The model accuracy and performance is investigated under various aspects. A correlation among the various biomarkers extracted from the raw EEG signals is examined for signal channel data (Fp1) and the mean data matrix (in each the number of subject is 120). The correlation among the biomarkers have evidently changed in both the cases. Also, the conclusion drawn from the Fig. 6.(a) are HFD is uncorrelated, while LZC has non-linear relationship with Beta bandpower. Rest all the features are exhibiting no relationship among themselves. On the other hand, for mean data matrix Fig. 6.(b) Beta bandpower is uncorrelated and HFD has manifested a non-linear relationship. Thus, is suitable in providing a non overlapping clustering data segregation.

Through the Fig. 9, one can observe that Fig. 9.(a) has many outliers as compared to the hemisphere data. And, correspondingly the accuracy obtained for mean data is 90% as compared to 95% and 97% for Left and right hemisphere respectively. The right hemisphere data is more skewed as compared to the left. The heavy tailed data distribution affect the nodes in spectral partitioning which can be observed in the results from the Table.VII. Here, higher ARI and RI are reported for left as compared to right in spectral clustering. Also, as the data is not normally distributed the K-means algorithm is not able to provide the accuracy being less computationally complex as compared to the spectral clustering.

The role of each channel in the prediction of MDD needs to be examined. In this study, the clustering is performed channel wise of each subject to understand the role play by the waves captured by the electrodes. In Table.VI, the accuracy along with sensitivity and specificity are investigated. Similarly, the clustering parameter are tabulated in Table.VIII. The conclusion drawn are : maximum accuracy has been achieved for channel Fp1, and on the other hand the ARI is 0.8389 with CEP = 0.0417. This is the channel that is involved with cognition, working memory and perception. Also, the least accuracy obtained is for the F4 with a negative ARI and CEP of 0.480. The presence of electrode artifacts can be one of potential reason for low accuracy reported. The clustering results are shown for mean dataset in Fig. 5, where although the data is overlapping on the two selected feature axis. The
TABLE V: Model accuracy comparison with the other existing methods.

k-means provides a hard partitioned cluster in place of the overlapping scenario as per the original labels. While in case of left hemisphere as visible in Fig.7 and right in Fig.8.

The methodology designed is also compared with the other existing methods in literature. These work also have taken the same dataset of EEG for detection of MDD. The comparison is tabulated in Table.V. Achieving an accuracy of 98% for left and 97% for right hemisphere, with 0 training time and just 2.005 msec for 120 subjects. The time complexity is also compared with the same methods to understand the time against accuracy accomplished in Table.IV.

TABLE VI: Model accuracy for different biomarkers for individual channels of brain electrodes.

V. Conclusion

EEG signals have been explored to understand the brain activities and understand the conditions of human psychology. In this study, the authors have designed a complete framework that reduces the cost of annotation of EEG signals, along with all the time and space complexity associated with training huge networks is addressed. Through the method small sample size data without labels can also be exploited to understand the hidden patterns
even for a new problem or aliment. The framework extracts a few prominent biomarkers along with the signal extracted features of raw EEG signals. Then, different datamatrix are made to realize the effect of brain activity in different regions and their correlation. The results achieved has been able to prove the efficiency and also being one of its kind of the flow designed. The overall accuracy for left hemisphere achieved is 98% while for right is 97%. While, in case of channel wise the highest accuracy obtained is for channel Fp1 which is 96%.

In the future, authors would like to explore other biomarkers like Spectral Asymmetry Index (SASI), Hjorth Complexity and, Hurst Exponent. Through which more insight of the EEG signal can be studied. The effect of each biomarker on the overall accuracy along with understanding of finding more prominent and less complex way of spectral partitioned also needs to be studied.

VI. ACKNOWLEDGEMENT

The research work is carried out with the support of Department of Computer Science and Department of Electronics at Banasthali Vidyapith, India.
REFERENCES


Shreya Garg is an undergraduate student, currently completing her Bachelor’s degree specializing in Electronics and Communication from Banasthali Vidyapith. Her research interests revolve around Artificial Intelligence, Machine Learning, Digital Signal Processing and related fields.
Dr Urvashi Prakash Shukla is at present working as an Assistant Professor in the Department of Computer Science, in Bansanthali Vidhyapith. She received her bachelor’s degree in Electronics Communication Engineering in 2009. She pursued her M.Tech in 2013 in the field of Compressive sensing. She completed her Ph.D. from NIT, Jaipur, and Rajasthan, India on “Clustering Algorithms Based on Social Spider Optimization for Hyperspectral Image Analysis”. She has published many articles in reputed journals and conferences. An active reviewer of IEEE Access and Elsevier journals. Her research interests revolve to obtain the solution for problems associated with medical signal processing, Hyperspectral Images, Data mining, and Artificial Intelligence.