Grenze International Journal of Engineering and Technology, January Issue

EEG Signal Classification using Unsupervised and Supervised Machine Learning Algorithm

Sanjay S Menon[1], Abhay P[2], Vishnu Narayanan[3], Muhammed Rashid[4], Hema P Menon[5] and Davidson Devasia[6]
[1-5]Department Of Computer Science Engineering, Sreepathy Institute of Management and Technology, Vavanoor, India
Email: sanjay.s.menon@simat.ac.in, abhay.p@simat.ac.in, muhammed.rashid@simat.ac.in, vishnu.narayanan@simat.ac.in, hemapmenon@simat.ac.in
[6]Department of Pediatric Neurology, Aster Medicity, Kochi, India
Email: davidson.devasia@asterhospital.com

***Abstract*— This work aims to develop a robust classification system for EEG signals to differentiate between ictal, inter-ictal, and pre-ictal stages, essential for accurate epilepsy diagnosis and management. Leveraging advanced machine learning techniques and signal processing methodologies, the work focuses on preprocessing and analyzing EEG data to extract pertinent features that capture the unique patterns associated with each stage. Utilizing classification algorithms such as Random Forest and K-means clustering, EEG segments are categorized into the corresponding stages based on the extracted features. The proposed classification system is subjected to comprehensive evaluation using performance metrics, including accuracy, precision, and recall, to assess its effectiveness and reliability in stage detection. The outcomes of this project hold substantial potential for enhancing the accuracy and efficiency of epilepsy diagnosis. Identifying the pre-ictal stage will make it easy for predicting the on-set of a seizure, contributing to improved patient care and quality of life.**

***Index Terms*—EEG signals, Epilepsy Diagnosis, Random Forest, K-means clustering.**

I. INTRODUCTION

EEG is technique used to record electrical activity of the brain. It plays a pivotal role in diagnosing various neurological disorders, including epilepsy. Epilepsy is a chronic neurological disorder characterized by recurrent seizures, which are sudden and unprovoked electrical disturbances in the brain. Identifying and classifying different stages of epileptic activity, such as ictal (during a seizure), interictal (between seizures), and preictal (before a seizure), is crucial for effective diagnosis, treatment planning, and seizure prediction.
Classification of EEG signals into these distinct stages is a challenging due to the complex and noisy nature of EEG data. Collecting the data is hectic and expensive. The work utilizes signal processing techniques and feature engineering, and machine learning algorithms to analyze and classify EEG data recorded from patients with epilepsy. The goal is to create a reliable and efficient classification model that can assist clinicians in identifying the On-set of the Seizure in the epileptic patients, A seizure is a sudden and uncontrollable burst of electrical activities in the brain. It can cause for changes in behavior, movements, and level of consciousness of the patient. When two or more seizure occurs within 24 hours without any known cause then it is considered to be as epilepsy.

Grenze ID: 01.GIJET.11.1.535

Manual analysis of EEG signals poses limitations in accuracy, efficiency, leading to delay in seizure prediction and giving medical care to the patient. The study tries to understand the different features of seizure and EEG signal and the variations in the signal during, before, and after the seizure to classify the signal.

II. LITERATURE SURVEY

Epilepsy is a neurological disease caused by the recurrent seizures due to the sudden development of synchronous neuronal firing in the cerebral cortex. Most of the people affected was not ready for proper medication, thus it leads to shorten their lifespan. Leon D. Iasemidis [1] aimed to use the advanced signal processing techniques to determine the precursors of seizure from the brain signal. EEG measures brain activity to detect signs of impending seizures. Traditional visual analysis is time-consuming and may lack accuracy, prompting the development of automated systems.

Larbi Boubchir et. all.,[2] proposed a system that extract features from EEG signals in time, frequency, or time-frequency domains and use machine learning techniques like neural networks and support vector machines to classify these signals as seizure or non-seizure. Detecting epileptic spikes in EEG recordings using a Kalman Filter-based approach in this the proposed method involves a two-step process: first, pre-highlight the spikes by applying a time-varying autoregressive model (TVAR), and second, using a thresholding procedure to detect these spikes, it was done by Alexandros T. Tzallas et. all.,[3].

EEG signals change over time and are unpredictable, so different statistical methods are needed to analyze them effectively. Techniques like linear and non-linear methods, Fourier Transform, and Wavelet Transform are used to identify problems like epileptic seizures and enhance brain-computer interfaces are explained by Er. Jasjeet Kaur et. all.,[4].

The study by Haidar Khan et. all.,[5] looks at how to predict seizures that start in a specific part of the brain by using data from scalp EEGs and a deep learning model called a convolutional neural network (CNN). The model learns to recognize patterns in the brain's electrical activity that show three stages: normal activity, the period just before a seizure, and the seizure itself. It found that it could predict a seizure about 10 minutes before it starts with 87.8% accuracy and a low rate of false alarms. FATHI E. ABD EL-SAMIE et. all.,[6] did a review on spike detection algorithms for EEG and MEG signals focusing more on EEG signals, it shows how automated analysis of these signals help in seizure detection and prediction.it investigates feature-based methods, methods based on artificial intelligence, convolution and correlation-based methods, periodicity-based methods, and others.

An epileptic seizure detection system based on classifying raw EEG signal recording was proposed by Ahmed M. Abdelhameed et. all.,[7] which employs a mixing of unsupervised and supervised deep learning, and k-fold cross validation.

Paper by Jitendra Kumar Jaiswal et. all.,[8] suggests Random Forest algorithm is the well working when it comes to classification problems which is being considered in this paper. Shi Na et. all.,[9] suggests an improved k-means clustering algorithm to address inefficiencies in the standard approach. In traditional k-means, each data point recalculates its distance which slows down the process. The improved method reduces this by storing cluster assignments and distances from previous iterations. This results in faster execution and better accuracy so can be used in identifying the spike clusters.

Fan Liu et. all.,[10] explains how we can use Elbow method to determine unknown targets in the open world which can be utilized in identifying the clusters of spikes. A CNN based model for automated seizure detection was proposed by Wei X, Zhou L et. all.,[11] in which a multi-channel EEG signal was analyzed, temporal and spatial information were used. The developed 3D CNN model outperformed the 2D CNN model by having accuracy of more than 90%.

Itaf Ben Slimen et. all.,[12] proposed a way to predict ictal, pre-ictal stages of seizure with the help of the CHB-MIT database and stages were characterized based on spike rate variations according to shapes, spikes, and the amplitude of the EEG signal. The model showed a 92% accuracy in seizure prediction.

III. MATERIALS AND METHODS

The proposed system integrates both unsupervised and supervised machine learning algorithms to classify the EEG signal to different stages of seizure which will help to identify and predict seizure.

Work uses unsupervised algorithm to label the pre-ictal, ictal, inter-ictal stages EEG signal and use the labelled data to train a Supervised model (Random Forest). Spike detection is implemented to identify the abnormal brain activity during the seizure and then the identified spikes are clustered using K-Means algorithm to label them.

This is where the dynamic Identification of the value of K is necessary for this the Elbow method is used. The labelled data undergoes feature extraction and is saved into a CSV file, which is then used to train the Random Forest model. The overall workflow is depicted in Fig. 1.

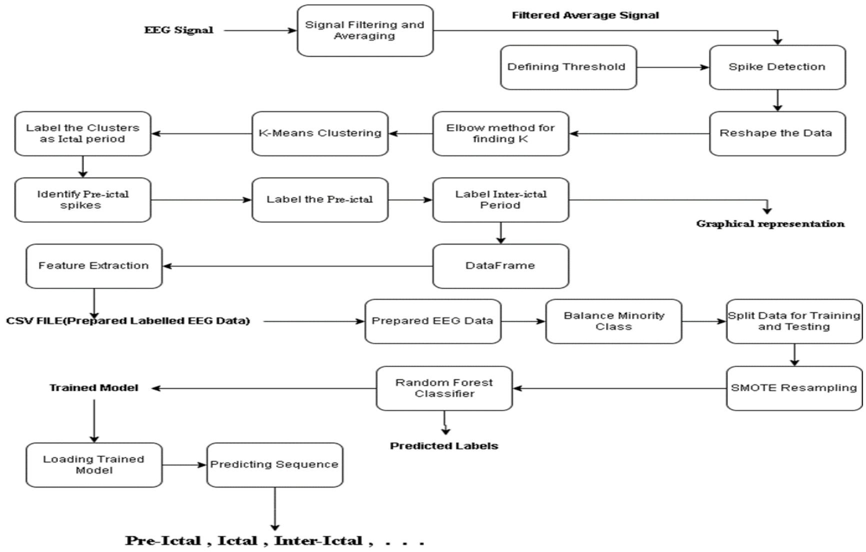


Figure 1. System Architecture

A. Steps Involved

Unsupervised algorithm
- Initially load the raw EEG data.
- Apply bandpass filter to remove noise.
- Get the data as a NumPy array.
- Calculate the average signal across all channels.
- Define thresholds for spike detection (find the highest amplitude and take half of the amplitude as threshold).
- Detect the spikes in the signal using the defined threshold values.
- for all the amplitude value in signal,

check for amplitude > positive_threshold or amplitude < negative_threshold.
- Reshape the data for clustering.
- Determine the optimal number of clusters using the elbow method.
- Apply K-means clustering to cluster the spikes.
- Group the clusters such that a group will contain spike epochs having time difference <=30 seconds and label them as ictal period in the signal.
- Identify pre-ictal spikes with the help of the pre-ictal threshold (half of the current threshold value) in between each ictal period and label as pre-ictal period in the signal.
- label the period between two ictal periods as inter-ictal period.
- save the following labelled EEG data to a CSV file, where label is as follows Ictal (0), Interictal (1), Preictal (2), None (3).
- Create features for each period and save it in the csv file ('Duration', 'Max_Amplitude', 'Min_Amplitude','Mean_Amplitude','Std_Amplitude','Spike_Count','Spike_Rate','Interictal_Duration','Preictal_Duration',’Change_in_Spike_Rate’), ('Label_Encoded'),
- Plot the average signal along with the labels.

Supervised Algorithm
- Load the EEG data from the CSV file produce with help of unsupervised algorithm into a pandas Data Frame.
- Filter out the instances where the label is "None" (class 3) to ensure that “None” class is not being classified.
- Select the relevant columns from the Data Frame as the input features and assign the encoded label column as the target output.

➢ Compute the size of the test set by calculating 20% of the dataset's total size.
➢ Split the dataset into training and test sets using the calculated test size.
➢ Identify the number of samples in the minority class.
➢ Apply SMOTE to the training data, ensuring the minority class is resampled to 50% of the training set size.
➢ Initialize a Random Forest Classifier and train it on the resampled data.
➢ Predict the labels for the test set using the trained Random Forest model.
➢ Measure the model's performance by calculating the accuracy and displaying a classification report.
➢ Save the trained model to a file for future predictions.
➢ Load the saved model and use it to predict labels on new data.
➢ Output the predicted labels with corresponding explanations: ictal = 0, interictal = 1, and preictal = 2.

B. Dataset

The Data set used is from American University of Beirut Medical Center. It has the EEG signal of six epileptic patients recorded by the Epilepsy monitoring unit of the American university of Beirut Medical Center, it was taken between January 2014 and July 2015. The data represents measurements from 21 scalp electrodes just like in Fig. 2, following the 10-20 electrode system, sampled at 500 Hz. The channels have been bandpass filtered between 1/1.6 Hz and 70Hz while filtering out the 50Hz. Some channels have been omitted from specific recordings due to artifacts constraints.
Data includes EDF (European Data Format) files containing above 10000 hours of Amplitude values of a patient. EDF file can be read using the MNE library in Python. The data set consist of 20 EDF files and each has a size around 1.5 – 2.5 GB. channel labels which show the electrode type and name. Fig. 3 shows the EDF files and Fig. 4 shows EEG signals of single EDF file.

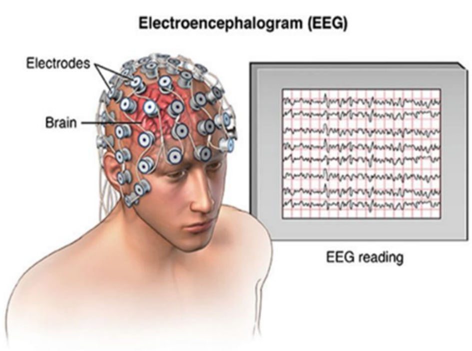
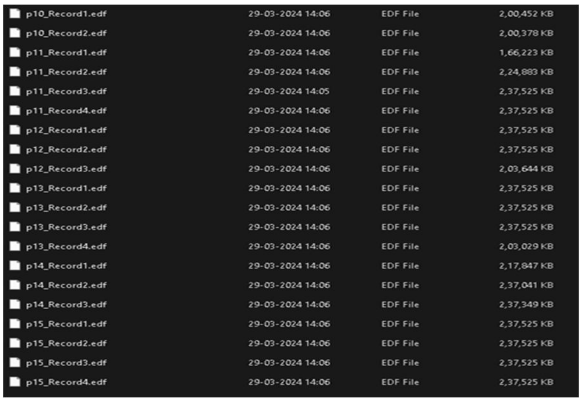


Figure 2. An illustration of EEG recording.
[pc: Siuly, S., Li, Y., Zhang, Y [13]]



p10_Record1.edf	29-03-2024 14:06	EDF File	2,00,452 KB
p10_Record2.edf	29-03-2024 14:06	EDF File	2,00,378 KB
p11_Record1.edf	29-03-2024 14:06	EDF File	1,66,223 KB
p11_Record2.edf	29-03-2024 14:06	EDF File	2,24,883 KB
p11_Record3.edf	29-03-2024 14:05	EDF File	2,37,525 KB
p11_Record4.edf	29-03-2024 14:06	EDF File	2,37,525 KB
p12_Record1.edf	29-03-2024 14:06	EDF File	2,37,525 KB
p12_Record2.edf	29-03-2024 14:06	EDF File	2,37,525 KB
p12_Record3.edf	29-03-2024 14:06	EDF File	2,03,644 KB
p13_Record1.edf	29-03-2024 14:06	EDF File	2,37,525 KB
p13_Record2.edf	29-03-2024 14:06	EDF File	2,37,525 KB
p13_Record3.edf	29-03-2024 14:06	EDF File	2,37,525 KB
p13_Record4.edf	29-03-2024 14:06	EDF File	2,03,029 KB
p14_Record1.edf	29-03-2024 14:06	EDF File	2,17,847 KB
p14_Record2.edf	29-03-2024 14:06	EDF File	2,37,041 KB
p14_Record3.edf	29-03-2024 14:06	EDF File	2,37,349 KB
p15_Record1.edf	29-03-2024 14:06	EDF File	2,37,525 KB
p15_Record2.edf	29-03-2024 14:06	EDF File	2,37,525 KB
p15_Record3.edf	29-03-2024 14:06	EDF File	2,37,525 KB
p15_Record4.edf	29-03-2024 14:06	EDF File	2,37,525 KB

Figure 3. Snapshot of EDF files

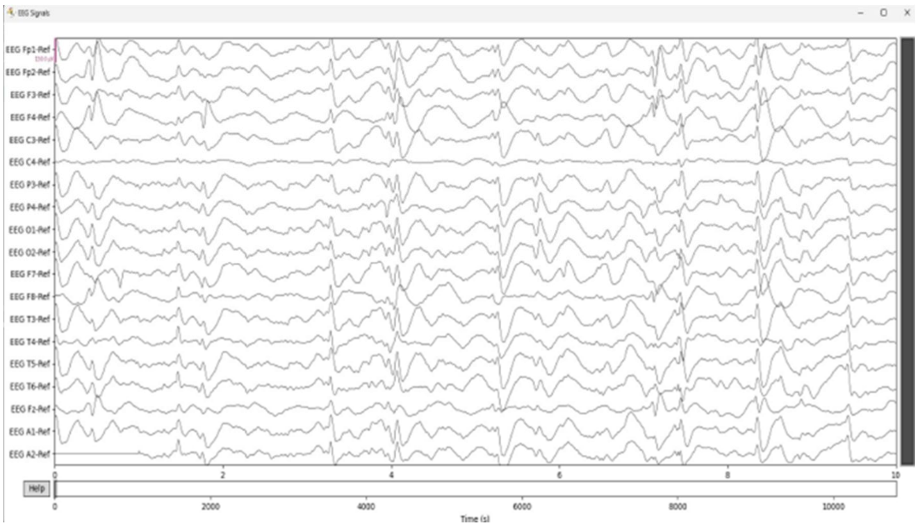


Figure 4. EEG signal waveform for one of the EDF file.

C. Classification Models

Spike Detection: After the preprocessing of the EEG signal next step is to detect the spikes for that a threshold value is set as positive threshold value and negative threshold value, which are being calculated by taking half of the highest amplitude in the EEG signal. The spike epochs are detected and undergoes K-Means Clustering. The pseudo code for spike detection is shown in Fig. 5.

K-Means clustering: The clustering is done and clusters are grouped with respect to a time variation i.e. for two spike epochs to be grouped together in a cluster it should have time difference less than 30 seconds (which is taken on the assumption that a normal ictal period may last 30 seconds to 5 minutes). The groups represent Ictal period and are labelled as Ictal 1, Ictal 2, and so on.

Now the pre-ictal thresholds (considered to be the half of previously identified threshold value) are used to predict the pre-ictal spike which are due to small variations seen in the evolution of the signal before reaching an ictal period. The first identified preictal spike in between two ictal periods mark the starting of the variation in the signal thus the period from this spike to the next ictal period is labelled as pre-ictal. Inter-ictal is the period between two ictal periods, so the period between two ictal period is labelled as inter-ictal. The pseudo code for clustering is shown in Fig. 6.

```
1.for i,amplitude in average_signal do
        a) if amplitude > positive_threshold or amplitude<negative_threshold
then
                a.1) if start_index == None then
                        a.1.1) start_index=i
        b) else if start_index != None then
                b.1) end_index=i-1
                b.2) spike_epochs.append((start_index,end_index))
                b.3) start_index=None
2.if spike_epochs == None then
        a) Display no epochs detected
```

Figure 5. Pseudo code for spike detection

```
1.set wcss = [ ]
2.for i from 1 to 11 do
        a) kmeans=KMeans(n_cluster=i,init='k-means++',max_iter=300,n_init=10,random_state=0)
3.kmeans.fit(X)
4.append(kmeans.inertia_) to wcss
5.difference=numpy.diff(wcss)
6.elbow_point=numpy.argmin(difference)+1
7.number_of_clusters=elbow_point
8. kmeans=KMeans(n_cluster=number_of_clusters,init='k-means++',max_iter=300,n_init=10,random_state=0)
9.pred_y=kmeans.fit_predict(X)
```

Figure 6. Pseudo code for K-Means clustering and Elbow method

Feature Extraction:After labelling is done the labelled data undergoes feature extraction and the following features are extracted.

'Duration': The difference between the end time and start time of each period, calculated as shown in (1).

$$Duration = EndTime - StartTime. \quad (1)$$

'Max_Amplitude' and 'Min_Amplitude': Here t1and t2 are indices of start and end times, shown in (2).and (3) respectively.

$$MaxAmplitude = \max(averageSignal[t1{:}\, t2]). \quad (2)$$

$$MinAmplitude = \min(averageSignal[t1{:}\, t2]). \quad (3)$$

'Mean_Amplitude': It is the average signal value over given period, as shown in (4).

$$MeanAmplitude = \frac{1}{t2-t1}\sum_{t=t1}^{t2} averageSignal(t). \quad (4)$$

'Std_Amplitude': It is the standard deviation of amplitude over given period, as shown in (5).

$$StdAmplitude = \sqrt{\frac{1}{t2-t1}\sum_{t=t1}^{t2}(averageSignal[t] - MeanAmplitude)^2}. \qquad (5)$$

'Spike_Count': Number of spikes occurring between start and end of a period shown in (6).

$$SpikeCount = \sum_{s \in S} 1(StartTime \leq s \leq EngTime). \qquad (6)$$

'Spike_Rate': Average number of spikes per second over the duration, as shown in (7).

$$SpikeRate = \frac{SpikeCount}{Duration}. \qquad (7)$$

'Change_in_Spike_Rate': In (7) for the first period its set as zero since there is no previous period.

$$ChangeInSpikeRate = SpikeRate_t - SpikeRate_{t-1}. \qquad (8)$$

After calculating all features, the prepared data is then saved into a CSV file for the model training.
Random Forest: Random forest is Supervised machine learning model that makes use of multiple decision trees while training to make predictions by combining the output of the decision trees. The trees are made by using the subsets of data and features. The classification is done based on majority vote of the trees. The work uses the random forest model since the model has the advantages such as high accuracy, does well in handling missing data which will be often seen in signal data, and the model gives more importance for features of the signal.The prepared CSV file is loaded and necessary features (X) and Labels (Y) are taken from the file. The test size is dynamically calculated by multiplying test_size_percentage with the number of samples in the data, then the data is split into training and testing set.
SMOTE (Synthetic Minority Over-sampling Technique) is used to balance the minority class. The k-neighbors are calculated dynamically by taking square root of the number of samples in the minority class and this k-neighbors is used by SMOTE to create synthetic samples.Random Forest classifier is initialized and then trained using the resampled training data.

IV. RESULT & ANALYSIS

The parameters used for filtering the raw data and the averaged signal of all the channels in a EDF file are shown in Fig. 7 and Fig. 8 respectively. Sample contents of the CSV file containing EEG data for model training that consist of 5 rows and 11 columns is depicted in Fig. 9. The graphical representation of EEG signal labelled as ictal, inter-ictal and pre-ictal stages is given in the Fig. 10.

```
FIR filter parameters
---------------------
Designing a one-pass, zero-phase, non-causal bandpass filter:
- Windowed time-domain design (firwin) method
- Hamming window with 0.0194 passband ripple and 53 dB stopband attenuation
- Lower passband edge: 1.00
- Lower transition bandwidth: 1.00 Hz (-6 dB cutoff frequency: 0.50 Hz)
- Upper passband edge: 70.00 Hz
- Upper transition bandwidth: 17.50 Hz (-6 dB cutoff frequency: 78.75 Hz)
- Filter length: 1651 samples (3.302 s)
```

Figure 7. Filtering parameters

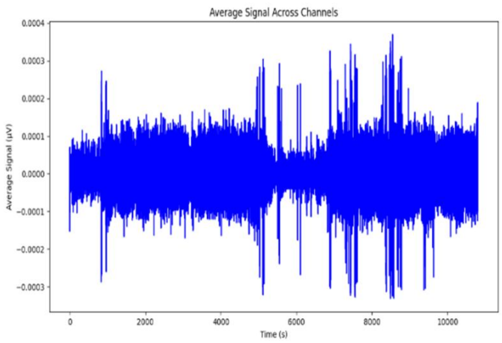
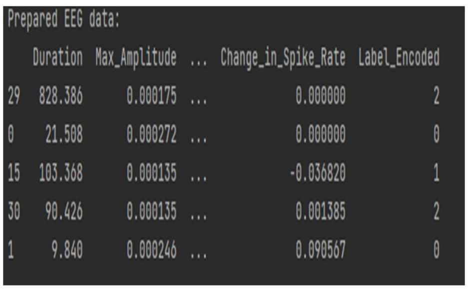


Figure 8. Average signal



Prepared EEG data:

	Duration	Max_Amplitude	...	Change_in_Spike_Rate	Label_Encoded
29	828.386	0.000175	...	0.000000	2
0	21.508	0.000272	...	0.000000	0
15	103.368	0.000135	...	-0.036820	1
30	90.426	0.000135	...	0.001385	2
1	9.840	0.000246	...	0.090567	0

Figure 9. Contents of CSV file

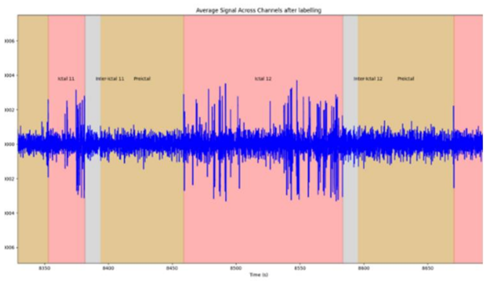


Figure 10. Graphical representation of labelled signal

The model gives an accuracy of 86.36 % in predicting the test set and the information about the "precision", "recall", "f1-score", and "support" is given below in Table. 1and the equations are shown in (8), (9), (10), and (11) respectively. After the prediction a sequence is displayed as shown in Fig. 11.

$$Precision = \frac{TruePositives(TP)}{TruePositives(TP)+FalsePositives(FP)}. \qquad (9)$$

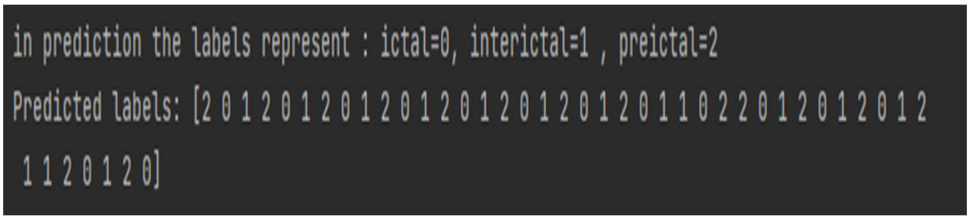
$$Recall = \frac{TruePositives(TP)}{TruePositives(TP)+FalseNegatives(FN)}. \qquad (10)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \qquad (11)$$

$$Support = NumberOfActualSamplesInTheClass. \qquad (12)$$

TABLE. I. MODEL ACCURACY

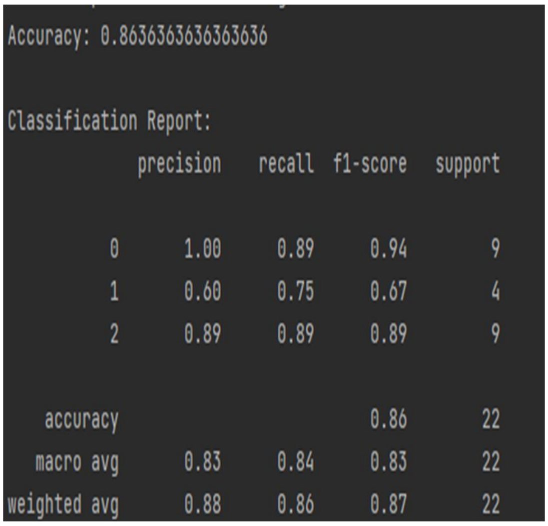
Stages	precision	recall	f1-score	support
0	1.00	0.89	0.94	9
1	0.60	0.75	0.67	4
2	0.89	0.67	0.89	9



```
in prediction the labels represent : ictal=0, interictal=1 , preictal=2
Predicted labels: [2 0 1 2 0 1 2 0 1 2 0 1 2 0 1 2 0 1 2 0 1 2 0 1 1 0 2 2 0 1 2 0 1 2 0 1 2
 1 1 2 0 1 2 0]
```

Figure 11. Predicted Sequence

The testing carried out and an accuracy of 86.36 % was obtained while 50 % of the data is being used for training and rest is for testing. A classification report was also produced for the model performance as shown in Fig 15. The confusion matrix given in Fig 16 shows that 8 of the ictal labels where correctly predicted and 1 was misinterpreted as inter-ictal, 3 of the inter-ictal labels where correctly predicted and 1 was misinterpreted as pre-ictal, and 8 of pre-ictal labels were predicted correctly and 1 was predicted as interictal.



```
Accuracy: 0.8636363636363636

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.89      0.94         9
           1       0.60      0.75      0.67         4
           2       0.89      0.89      0.89         9

    accuracy                           0.86        22
   macro avg       0.83      0.84      0.83        22
weighted avg       0.88      0.86      0.87        22
```

Figure 15. Classification Report

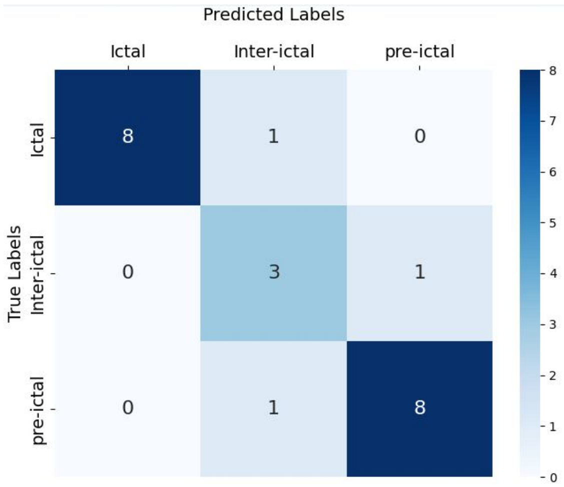


Figure 16. Confusion matrix

V. CONCLUSION

The work was done using K-Means clustering and Random Forest model. A unique way was implemented where we took unlabeled data and labelled them using unsupervised method which then was given as data for the random forest model. This approach was taken so that we can label the EEG data and train supervised model. The approach is inexpensive and was done using online Data set which was validated by ASTER MEDCITY, Kochi. The possibility in finding the on-set of a seizure is improved by the classification of the EEG signal, thus the future works can include identification of the on-set of a seizure in epileptic patients.

ACKNOWLEDGMENT

The authors would like to thank Dr. K M Mathew, Dr.Dhinla and Mr. M C V Arun, Aster Medicity, Kochi for their expert advice and validation of the results throughout the work.

REFERENCES

[1] L. D. Iasemidis, "Epileptic seizure prediction and control," in IEEE Transactions on Biomedical Engineering, vol. 50, no. 5, pp. 549-558, May 2003.

[2] L. Boubchir, B. Daachi and V. Pangracious, "A review of feature extraction for EEG epileptic seizure detection and classification," 2017 40th International Conference on Telecommunications and Signal Processing (TSP), Barcelona, Spain, 2017, pp. 456-460.

[3] A. T. Tzallas, V. P. Oikonomou and D. I. Fotiadis, "Epileptic Spike Detection Using a Kalman Filter Based Approach," 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, New York, NY, USA, 2006, pp. 501-504.

[4] J. Kaur and A. Kaur, "A review on analysis of EEG signals," 2015 International Conference on Advances in Computer Engineering and Applications, Ghaziabad, India, 2015, pp. 957-960.

[5] H. Khan, L. Marcuse, M. Fields, K. Swann and B. Yener, "Focal Onset Seizure Prediction Using Convolutional Networks," in IEEE Transactions on Biomedical Engineering, vol. 65, no. 9, pp. 2109-2118, Sept. 2018.

[6] F. E. Abd El-Samie, T. N. Alotaiby, M. I. Khalid, S. A. Alshebeili and S. A. Aldosari, "A Review of EEG and MEG Epileptic Spike Detection Algorithms," in IEEE Access, vol. 6, pp. 60673-60688, 2018.

[7] A. M. Abdelhameed and M. Bayoumi, "Semi-Supervised EEG Signals Classification System for Epileptic Seizure Detection," in IEEE Signal Processing Letters, vol. 26, no. 12, pp. 1922-1926, Dec. 2019.

[8] J. K. Jaiswal and R. Samikannu, "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression," 2017 World Congress on Computing and Communication Technologies (WCCCT), Tiruchirappalli, India, 2017, pp. 65-68.

[9] S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jian, China, 2010, pp. 63-67.

[10] F. Liu and Y. Deng, "Determine the Number of Unknown Targets in Open World Based on Elbow Method," in IEEE Transactions on Fuzzy Systems, vol. 29, no. 5, pp. 986-995, May 2021.

[11] Wei X, Zhou L, Chen Z, Zhang L, Zhou Y. Automatic seizure detection using three-dimensional CNN based on multi-channel EEG. BMC Med Inform Decis Mak. 2018 Dec 7;18(Suppl 5):111.

[12] Slimen IB, Boubchir L, Seddik H. Epileptic seizure prediction based on EEG spikes detection of ictal-preictal states. J Biomed Res. 2020 Feb 17;34(3):162-169. doi: 10.7555/JBR.34.20190097. PMID: 32561696; PMCID: PMC7324272.

[13] Siuly, S., Li, Y., Zhang, Y., 2016. Electroencephalogram (EEG) and Its Background. EEG Signal Analysis and Classification. pp. 3-21.