Hand-Engineered Features for Detection of Epilepsy

Ishita Sharma
S23097, SCEE (IIT Mandi)
Himachal Pradesh, India
s23097@students.iitmandi.ac.in

Jeet Bandhu Lahiri
D23146, SCEE (IIT Mandi)
Himachal Pradesh, India
d23146@students.iitmandi.ac.in

***Abstract*—Electroencephalography (EEG) is essential for assessing epilepsy, a neurological disorder marked by recurring seizures. Understanding the statistical properties of EEG signals can give insights for the diagnosis of several neurological disorders. In our research, we develop a method to statistically analyze EEG signals to build a classifier distinguishing between epileptic and healthy subjects.**

The experiment utilized electroencephalogram (EEG) signals collected from participants with epileptic and healthy individuals. The EEG setup adhered to the international 10-20 system, capturing brain activity through 19 channels. Statistically hand-engineered features from these data serve as computational biomarkers for distinguishing epileptic from non-epileptic brain activity.

Our detector analyzes EEG recordings to differentiate epileptic and healthy subjects. EEG data is recorded using the 10-20 electrode system, available in a referential montage. Signals last 3-5 minutes, sampled at 125 Hz. Finally, we constructed the ROC curve of the detector, calculating an AUC for evaluation using various combinations of the dataset. We achieved an AUC of 0.88, the best result for all of the various combinations on the dataset that we used.

***Index Terms*—epilepsy, detector, hand-engineered features, ROC, AUC**

I. Introduction

Epilepsy affects approximately 50 million people globally[1], making it a significant public health concern. Recent research has explored using classifiers to analyze brain activity (EEG) for epilepsy detection. Existing studies exploiting complex machine learning architectures often show very high accuracy but might have several limitations because the models are often seen to perform poorly in a holdout test dataset. There are concerns that the existing benchmark dataset may not be suitable for developing reliable classifiers for epilepsy diagnosis[2].

In his work [4], Zhihao Guo demonstrated that seizures primarily spread within, rather than across, intrinsic networks. Severe neurological disorders like epilepsy can cause a loss of synchronization between the signals from different EEG channels [5].

We propose a unique hand-engineered statistical feature specifically designed to identify epilepsy patterns in scalp EEG recordings. This feature was tested on a dataset of 86 individuals (both epileptic and healthy).

Our classifier achieved a maximum area under the curve (AUC) of **0.88**, indicating good performance in distinguishing between epileptic and healthy EEG patterns. This method surpasses existing models and fulfills the goal of creating a binary classifier for epilepsy diagnosis. This approach offers a promising path for improving epilepsy detection techniques and brings an aspect of explainability to detection systems.

II. Objective

We aim to model a classifier system that can accurately distinguish between epileptic and healthy patients based on features extracted from their EEG recordings. Ideally, this classifier would perfectly separate the patterns of brain activity (represented by probability density functions) between the two groups. To measure the effectiveness of our classifier, we will track a value called 'area under the curve' (AUC) using a tool called 'receiver operating characteristic curve' (ROC curve). A higher AUC indicates a better ability to differentiate between epileptic and healthy EEG patterns.

III. Methodology

The data was collected from the standard 10-20 system of scalp EEG, placing 19 electrodes across various regions of the brain. The signal is sampled at 125 Hz. For each sliding window, we calculated an 8x8 covariance matrix for each hemisphere taking normalized values and following a particular sequence of 8 electrodes for both the hemispheres. After calculating the 8x1 dominant eigenvector for the covariance matrix, we took the difference of them for each sliding window. This difference captured the asymmetry in electrical activity between the two hemispheres. We then transformed this obtained 8x1 vector into a spherical coordinate system. The angles, representing the distribution of electrical activity across the brain, became our features for further analysis. We repeated this process for all sliding windows in the EEG data from epilepsy and healthy subjects. Finally, we used Kernel Density Estimation (KDE), a non-parametric estimation system, to visualize the probability distributions of each of these seven angles for all subjects, allowing us to compare the electrical activity patterns in epilepsy and healthy brains.

A. Data and Classification

The 10-20 system, a widely accepted international EEG standard, guides scalp electrodes' placement during electroencephalogram recordings, ensuring consistency and enabling data comparison across studies and medical facilities. This system employs a numbered grid to pinpoint scalp locations

corresponding to specific brain regions, ensuring uniform electrode placement regardless of individual head size or shape. Our dataset comprises 86 .fif files, each containing EEG data from all 19 electrodes, gathered from 36 healthy individuals and 50 epileptic patients. To construct the covariance matrix, we utilized a specific electrode sequence given below, excluding the midline frontal (FZ), central (CZ), and parietal (PZ) electrodes, as our focus is on discerning left-right brain activity differences, with midline electrodes having minimal impact on this contrast.

The sequence of electrodes on the left: Fp1, F3, F7, C3, T3, P3, T5, O1.

The sequence of electrodes on the right: Fp2, F4, F8, C4, T4, P4, T6, O2.

The following figure illustrates the same.

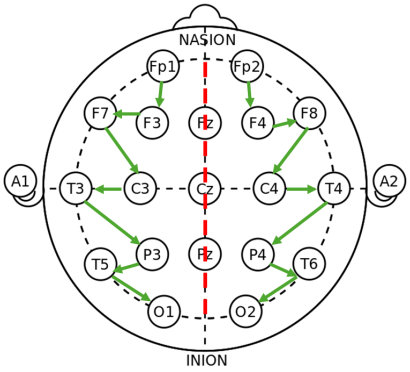

Fig. 1: Channel selection paradigm

B. Framework

1) ***Feature Extraction****:* To extract features we first took a sliding window of 1 second, with a stride of 1 sample, and for each window, formed an 8x8 covariance matrix of the sequence of electrodes for each hemisphere of the brain.

Given an 8x8 covariance matrix C, we can find its eigenvalues λ_i and corresponding eigenvectors $\mathbf{v}_i$ by solving the equation:

$$C\mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{1}$$

where $i = 1, 2, \ldots, 8$.

We then take the dominant eigenvector (the eigenvector with the maximum absolute eigenvalue) for both hemispheres under the same sliding window and take the difference of them to obtain an 8x1 vector.

$$\Delta\mathbf{v} = \mathbf{v}_1 - \mathbf{v}_2 \tag{2}$$

The difference captured the asymmetry between the two sides which we hope to provide insights into the electrical activity of epileptic subjects. We then transformed this new vector into an 8-dimensional spherical coordinate system, denoted as $(r, \theta_1, \theta_2, ..., \theta_7)$, where r is the norm and $\theta_1, \theta_2, ..., \theta_7$ represent the set of 7 angles. The probability distribution of these angles across all the sliding windows in a sample, mapped by KDE, a non-parametric estimation technique, became our features for further analysis.

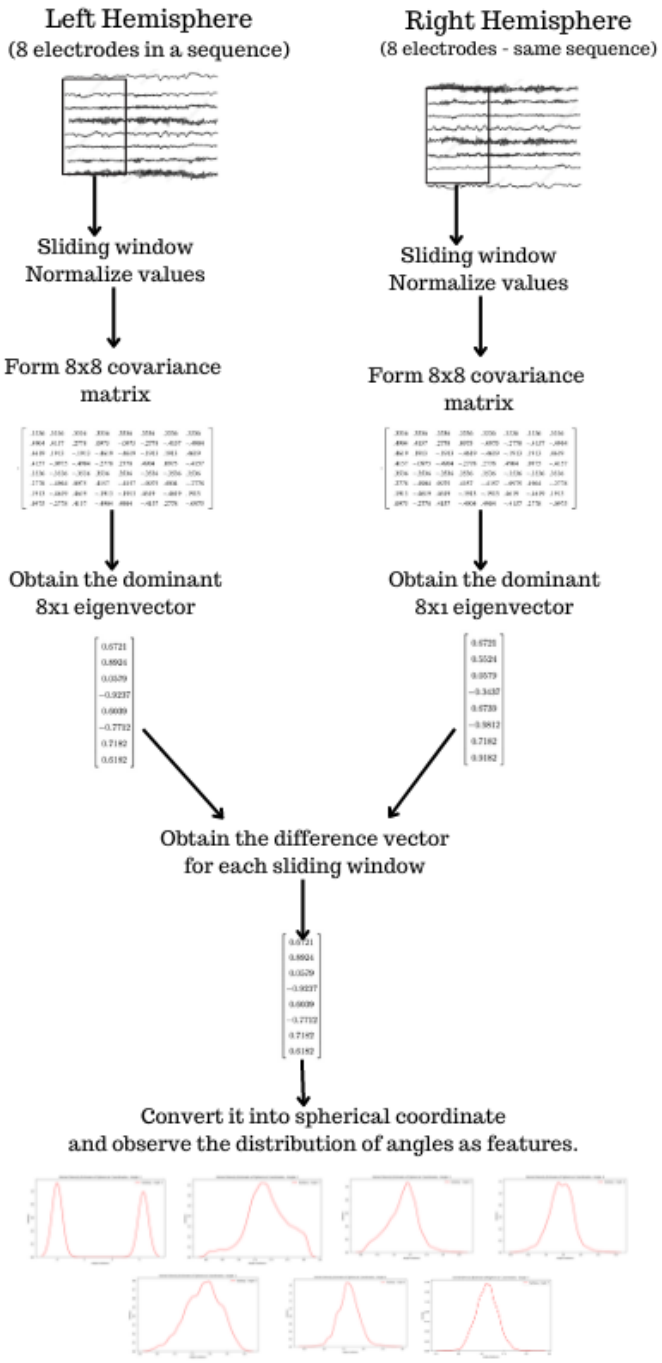

Fig. 2: Methodology

2) ***Setting up the LRT****:* Once we have the probability distributions for both epileptic and healthy patients we shall use the *Neyman-Pearson* approach for identification of EEG signals of epileptic patients using statistical hypothesis testing.

H_0 is defined as the null hypothesis under the assumption that the EEG signal is of a non-epileptic patient.

H_1 is defined as the alternate hypothesis under the assumption that the EEG signal is of an epileptic patient.

From our observations the pdfs follow a distribution that could be related with a Gaussian distribution which is denoted as $\mathcal{N}(\mu, \sigma^2)$. Distribution function for which is given as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

We have 86 files, or we can say that we have 86 pdfs (50 for epileptic and 36 for healthy patients.) All the samples are Independent and Identically distributed because they belong to different individuals.

Now we shall find the Likelihood Ratio Test or the LRT.

The various cases of setting up the LRT for our experiments are:
Case 1: Using any one random sample each of epileptic and non-epileptic patient.
Case 2: Using joint distribution of any 10 random samples each of epileptic and non-epileptic
Case 3: Using joint distribution of any 20 random samples each of epileptic and non-epileptic
Case 4: Using joint distribution of any 30 random samples each of epileptic and non-epileptic

The **LRT** can be set according to the following:

Given hypotheses H_0 and H_1, let $f(x; H_0)$ be the probability density function (PDF) under H_0 and $f(x; H_1)$ be the PDF under H_1. The likelihood ratio $\Lambda(x)$ for a single sample x is:

$$\Lambda(x) = \frac{f(x; H_1)}{f(x; H_0)}$$

The likelihood ratio test (LRT) decides H_1 if $\Lambda(x)$ exceeds a certain threshold η:

$$\Lambda(x) \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

Assuming all the samples to be independently and identically distributed (i.i.d.) $x_1, x_2, \ldots, x_i$, where i can be 10, 20 or 30, the likelihood ratio say for 10 samples can be given as: $\Lambda(x_1, x_2, \ldots, x_{10})$ where it can be written in terms of the ratio of joint pdfs which in this case is the product of individual pdfs of the i.i.d. samples.

$$\Lambda(x_1, x_2, \ldots, x_{10}) = \frac{f(x_1; H_1) f(x_2; H_1) \cdots f(x_{10}; H_1))}{f(x_1; H_0) f(x_2; H_0) \cdots f(x_{10}; H_0))}$$

Or, equivalently,

$$\Lambda(x_1, x_2, \ldots, x_{10}) = \frac{\prod_{i=1}^{10} f(x_i; H_1)}{\prod_{i=1}^{10} f(x_i; H_0)}$$

The LRT decides H_1 if $\Lambda(x_1, x_2, \ldots, x_{10})$ exceeds a certain threshold η:

$$\Lambda(x_1, x_2, \ldots, x_{10}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

In logarithmic form, the LRT becomes:

$$\sum_{i=1}^{10} \log\left(\frac{f(x_i; H_1)}{f(x_i; H_0)}\right) \underset{H_0}{\overset{H_1}{\gtrless}} \log \eta$$

The value of η is unknown so we have different thresholds for different cases.

*3) **ROC**:* We then computed Receiver Operating Characteristic (ROC) curves to evaluate the performance of our extracted features in distinguishing between different conditions. Using the False Positive Rate (FPR) and True Positive Rate (TPR) calculated from the labels and the feature values, we plotted the ROC curve. The area under the ROC curve (AUC) was calculated to quantify the effectiveness of our features in classification tasks. A higher AUC indicates better discrimination ability, with a value closer to 1 representing perfect classification.

C. Figures and Tables

We iterated the experiments and found the PDFs for combinations of patient samples, including 1, 10, 20, and 30 patients. Corresponding to each combination, different AUC values were obtained for various values of θ. The three best θ values, which yield the highest AUC values, are presented in the table below:

TABLE I: Angles of interest

Number of Samples	**AUC for values of θ**		
	θ_2	θ_3	θ_4
1	0.68	0.76	0.88
10	0.65	0.59	0.56
20	0.51	0.55	0.52
30	0.58	0.57	0.50

The following figures showcase the results derived from the experimentation process. To streamline the presentation, we have exclusively included plots corresponding only to the optimal AUC values attained for a particular number of samples.

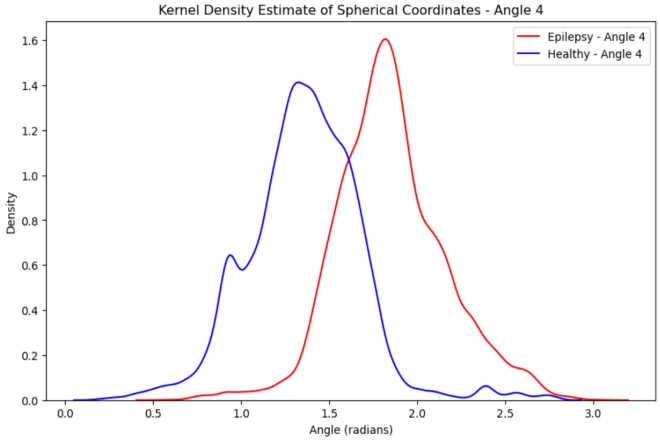

Fig. 3: KDE of θ_4 for number of samples = 1

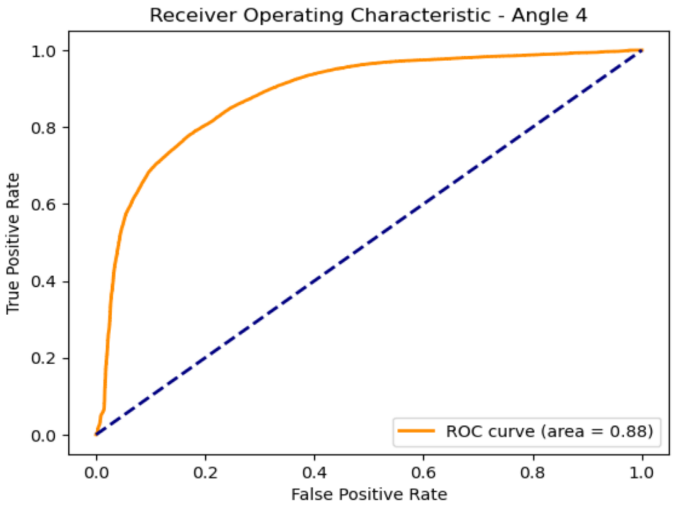

Fig. 4: ROC curve provided by θ_4 for a sample size of 1

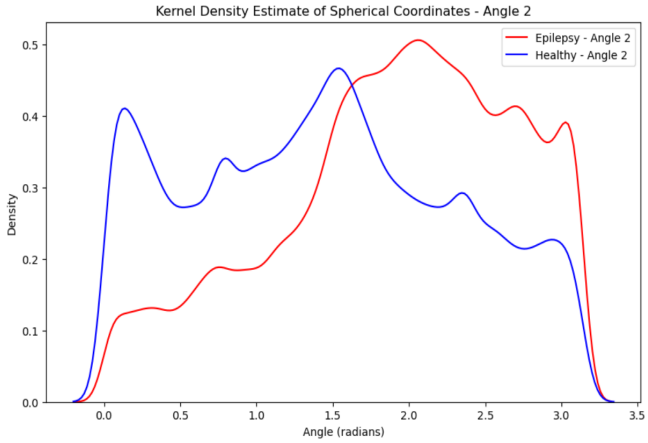

Fig. 5: KDE of θ_2 for number of samples = 10

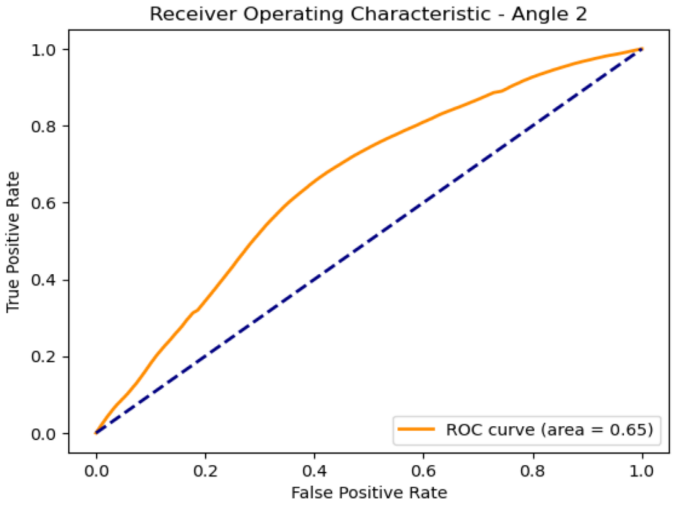

Fig. 6: ROC curve provided by θ_2 for sample size of 10

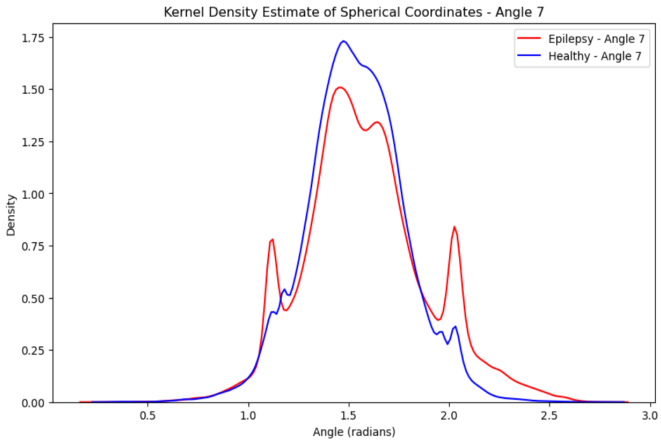

Fig. 7: KDE of θ_7 for number of samples = 20

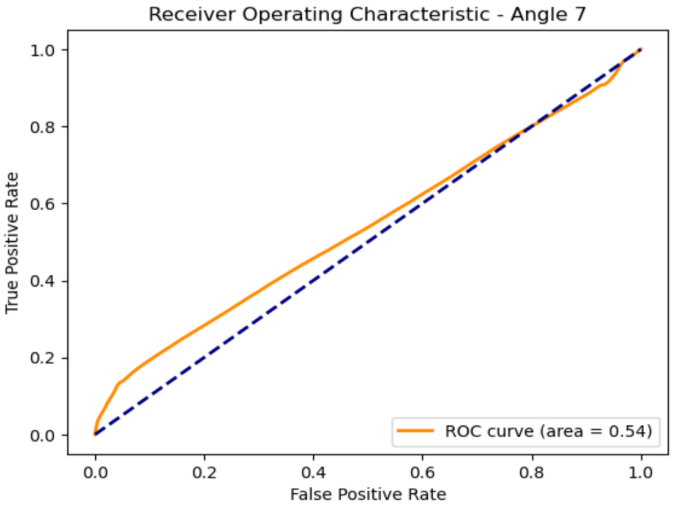

Fig. 8: ROC curve provided by θ_7 for a sample size of 20

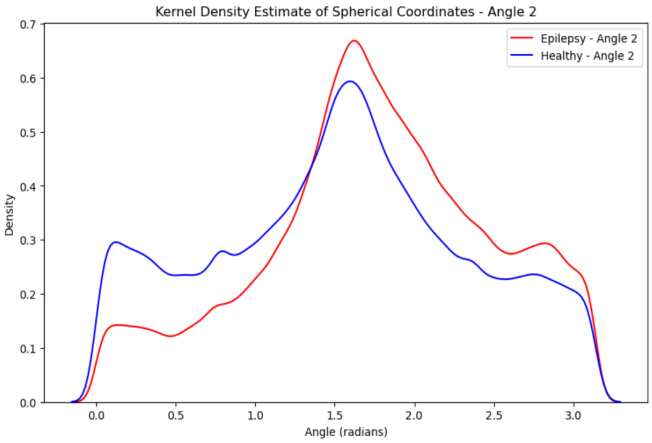

Fig. 9: KDE of θ_2 for number of samples = 30

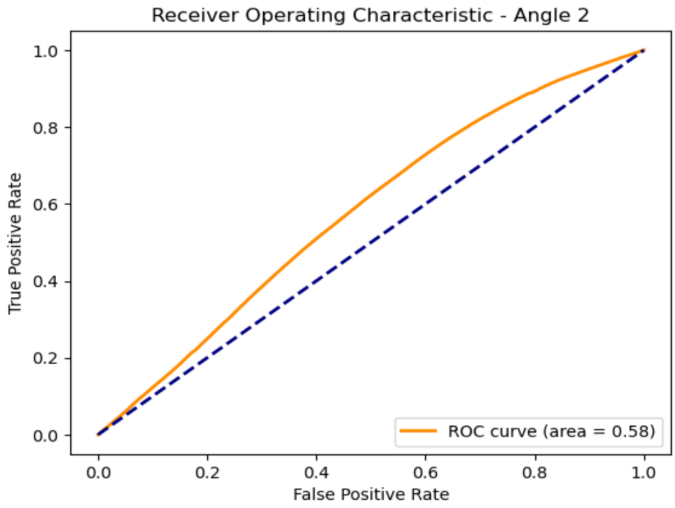

Fig. 10: ROC curve provided by θ_2 for a sample size of 30

IV. Results and Discussions

We observed that the highest AUC, 0.88, was achieved with a single pair of samples randomly chosen. As the number of randomly selected samples increased to 10 for each class, the AUC dropped to an optimal value of 0.65. For 20 samples each, the optimal AUC further decreased to 0.55, and for 30 samples each, it slightly improved to 0.58. This pattern suggests that increasing the number of samples leads to a gradual decrease in the AUC value. This trend may be due to the accumulation of skewness in the joint probability density function (pdf). As more samples are included, the added skewness causes the pdfs to become less distinctly separated.

V. Conclusion

We developed a hand-engineered feature that classifies patients as either epileptic or healthy based on their EEG data. While many existing models are computationally intensive, we introduced a straightforward classifier which is derived from the fundamentals of the Neyman-Pearson Theorem for statistical hypothesis testing. Our handcrafted feature provided a basic level of classification. In the future, this fundamental approach could enhance the performance of more complex classifiers, potentially increasing their efficacy.

Code

The code is made publicly available for research and reproduction purposes at the following link: https://github.com/jeetblahiri/EEG-Hand-Engineered-Features/tree/main

References

[1] Huo Q, Luo X, Xu ZC, Yang XY. Machine learning applied to epilepsy: bibliometric and visual analysis from 2004 to 2023. Front Neurol. 2024;15. doi: 10.3389/fneur.2024.1374443.

[2] Panwar S, Joshi SD, Gupta A, Agarwal P. Automated Epilepsy Diagnosis Using EEG With Test Set Evaluation. IEEE Trans Neural Syst Rehabil Eng. 2019 Jun;27(6):1106-1116. doi: 10.1109/TNSRE.2019.2914603. Epub 2019 May 3. PMID: 31059452.

[3] 124.'Own work,' Public Domain. Available: https://commons.wikimedia.org/w/index.php?curid=10489987.

[4] Z. Guo, J. Zhang, W. Hu, X. Wang, B. Zhao, K. Zhang, and C. Zhang, "Does seizure propagate within or across intrinsic brain networks? An intracranial EEG study," *Neurobiology of Disease*, vol. 184, 2023, Art. no. 106220. doi: https://doi.org/10.1016/j.nbd.2023.106220.

[5] Sobayo, T., Farahmand, S., Mogul, D.J. (2023). Determining the Role of Synchrony Dynamics in Epileptic Brain Networks. In: Thakor, N.V. (eds) Handbook of Neuroengineering. Springer, Singapore. https://doi.org/10.1007/978-981-16-5540-1_71