An Explainable AI approach towards Epileptic Seizure Detection

Neeta Chapatwala
E & C Engineering
SCET
Surat, India
ORCID: 0000-0002-1401-0029

Chirag N. Paunwala
E & C Engineering
SCET
Surat, India
ORCID: 0000-0002-0201-7370

Poojan Dalal
E & C Engineering
SCET
Surat, India
ORCID:0000-0001-6338-1371

2022 IEEE 19th India Council International Conference (INDICON) | 978-1-6654-7350-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/INDICON56171.2022.10039982

***Abstract*— Epilepsy is a neurological condition resulting in abnormal behavior and recurring seizures. Epilepsy diagnosis requires analysis of large amount of EEG data bya trained specialist. Many potential methods for epilepsy diagnosis using EEG signals utilizing machine learning have been proposed in the literature, but understanding why a model predicts a certain output becomes important when applied to tasks in medical domain. Although the term "accuracy" may sound promising while applying on datasets, justifying a predicted output is important when it comes to diagnosis in real world. In response, we present an explainable AI approach for epilepsy diagnosis which explains the output features of a model using SHAP (Shapley Explanations) - a unified framework developed from game theory. The explanations generated from Shapleyvalues prove efficient for feature explanation for a model's output in case of epilepsy diagnosis. Explanations show the impact of each feature driving the output towards particularprediction. Moreover, the paper shows experimentation on different machine learning models and their SHAP explanations on the University of Bonn Dataset. The feature explanation approach demystifies the black-box algorithms and can be used to interpret a model's prediction even in- case of false diagnosis.**

Keywords—Electroencephalogram (EEG), epilepsy, Explainable AI (XAI), SHAP, Machine Learning, ethical AI

I. INTRODUCTION

An important task in the future prospects of artificial intelligence when it comes to biomedical diagnosis is howtrue its predictions are. When it comes to the diagnosis of a disease, the robustness of an algorithm towards true as well as false results has to be justifiable. The ethicalaspects of the future impact of AI on diagnosis have to beaddressed from this perspective. Despite some spectacular deep learning success stories, researchers were astonishedby the lack of validity of algorithms that were thought as efficient. Image classification have been found to malfunction when subjected to adversarial changes [1]. As a result, there remains a requirement for fair decisions and discussions about why an algorithm reacheda particular conclusion. This explainable issue can be solved as an attribution problem, i.e., distributing of a model's prediction for a specific input to its base features. Tracing the feature values (or even their derivatives) leading to a prediction isquite challenging due to complex network structures, attribute interactions and possible binary/non-binary feature values. And one of the leading approaches to the attribution problem is using the Shapley values [2] developed from game theory. As a result, applying this attribution to a model prediction in the context of disease diagnosis will tell us how important a feature was inmaking the diagnosis. We implemented such approach to demonstrate its application by combining the attribution method, with a machine learning model to find predictors for epilepsy diagnosis from EEG signals

EEG is a test which detects abnormalities in brain signals due to electrical activity. Functional Magnetic Resonance Imaging (fMRI), Electroencephalography (EEG), Magneto encephalography (MEG) are the various brain-computer interface methods for measuring neuronal signals. EEG is the most commonly used by reason of compensations such as high temporal resolution and low price for an efficient epilepsy diagnosis. Therefore, it is an important tool for diagnosis of epilepsy, which is a neurological disorder characterized by recurrent unprovoked seizures [2]. From scalp-recorded EEG data, scalp topography features are extracted to detect features related to abnormal activity of electrodes. This EEG signal must be examined in order to give the physician useful information [3]. Recent advancements in Artificial Intelligence and Machine Learning imply that computer-assisted procedures can more efficiently detect diversified brain problems.

The paper proposes explanation of machine learning model's output for a particular epilepsy diagnosis from EEG signals. The paper's organization is- Related Work is mentioned in Section II, Section III covers the proposed method together with feature extraction explanation and Shapley values, dataset description, experiments and results explaining the relevant features in Section IV, conclusion in Section V and references.

II. RELATED WORK

Various researchers made many efforts to diagnose epilepsy using modern machine learning techniques. Multiple researches showed several machine learning methods applied in medical diagnosis [8-10]. H. Gabaniet. al. [11] proposed an SVM (Support Vector Machine) based automatic epilepsy seizure classification system using ApEn (Approximation Entropy) and SVM (Support Vector Machine). Their purpose of using ApEn was to act as feature extraction, thus reducing the dataset size without losing patient data to classify the epileptic seizure easily. ApEn [12] is a statistical parameter, and using the prior amplitude value as a baseline, it calculates the current amplitude values of an EEG signal. T. Siddharth et al. [13] proposes a new method for classifying focal and non-focal EEC signal types. The EEG signal was divided into

reconstruction components (RC) using sliding mode-singular spectral analysis (SM- SSA). A number of five RCs were obtained from individual EEG signal using SM-SSA. Subsequently, a classification of the Sparse-Auto Encoder (SAE) hidden layer and the radial basic neural network function (RBFN) was proposed.

The main goal of the study [14] was to determine the cerebral characteristics of candidates with mesial Temporal Lobe Epilepsy (mTLE) along with data on late realization. To do this, SVM and XGBoost are used to diagnose the extracted parts left or right. A. Antoniades et al. used two CNN networks, CNN1 consisting of one convolution layer, and CNN2 consistingof two convolution layers. This makes the EEG background more different from the spikes. This improved differentiation between background movement and epileptic spikes allows the move from superficial to deep learning in spike detection. The conventional method based on SVM was demonstrated Lan-Lan Chen et al. [15], which used a decomposition technique based onwavelets.

With the increasing focus on explainable AI, various metrics and methods are developed for interpreting a machine learning model's result. Rich Caruana et al. [16] used different and representative samples, to summarize the data set or explain the test sample. A non-negative mean weight is also determined for individual of the selected samples. Hind et al. [17] propose a technique for developing a predictive model based on user-provided descriptions in addition to input-output labels. A label and explanation are retrieved for an unseen test instance.

The different methods demonstrated in the literature for the diagnosis of epilepsy do not include model construal or explanation about the particular outputof the model. The work demonstrated in this paper is inspired by concepts of game theory and Shapley values demonstrated as a feature explanation by Lundberg, S. M.and Lee [4]. To overcome the explainability issues of machine learning models, we propose the usage Shapley based explanations for interpreting model outputs for epileptic seizure diagnosis.

III. PROPOSED METHOD

A block diagram of proposed method is shown in Fig.1. The method takes the EEG Signal as input, followed by wavelet decomposition and feature extraction. All the 16 features are then normalizedand then machine learning model like SVM is trained for classification. The feature explainer SHAP is then appliedon the test results of model

A. Wavelet Decomposition

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Feature Extraction

Feature extraction is very important as it boosts overall

accuracy due to its uniqueness. Thus, the accurate choice of characteristics can facilitate straightforward classification. Following features were used for the feature extraction step:

- Approximate entropy – main signal

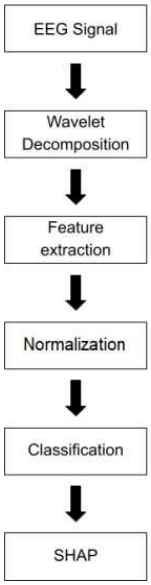

Fig. 1. Block Diagram of proposed method

- Approx. entropy – d3 to d7 sub band
- Sample entropy – main signal
- Recurrence Quantification Analysis
- Sub band energy of d3 to d7 sub bands

Thus, total of 16 features were extracted during this step. Each of these feature's impact is then interpreted using Shapley Explanations.

1) Approximate entropy – main signal

Approximate entropy (ApEn), a statistical technique, isused to assess the degree of regularity and predictability offluctuations in time-series data.

Approximate Entropy is given in the (1) as:

$$ApEn = \phi^m(r) - \phi^{m+1}(r) \quad (1)$$

where,

$$\phi^m(r) = (N-m+1)^{-1} \cdot \sum_{i=1}^{N-m+1} \ln C_i^m(r) \quad (2)$$

and

$$C_i^m(r) = \frac{(number\ of\ x(j)\ such\ that\ d[x(i),x(j)] \leq r)}{(N-m+1)} \quad (3)$$

where, x(i) and x(j) are vectors:

$$x(i) = [u(1), \ldots, u(i = m-1)] \quad (4)$$

$$x(j) = [u(1), \ldots, u(j = m-1)] \quad (5)$$

from time series data u(1), u(2),…,u(N).

2) Sample Entropy

Entropy, as a concept that a value may be reasonably defined from a series in an ordered system, can be stated as a type of index of regularity or the degree of unpredictability. If the number of sequences in a series is more intricate or non-ordered, the entropy will be greater,

and vice versa. A variant of approximate entropy, sample entropy reduces the bias brought on by self-matching. Sample Entropy is given in (6).

$$SampEn = \lim_{N\to\infty}[\ln(\phi^m(r) - \phi^{m+1}(r))] \quad (6)$$

3) Recurrence Quantification Analysis (RQA)

A nonlinear data analysis technique for assessing dynamical systems is recurrence quantification analysis (RQA). It determines how many and how long a dynamical system's recurrences are as represented by its phase space trajectory.

The four parameters extracted are:

a) Recurrence rate (%REC):

In a recurrence plot, it refers to the density of recurrence dots. The likelihood that a specific circumstance will return is known as the recurrence rate. The percentage of REC is determined as stated in (7).

$$\%REC = \frac{1}{n^2}\sum_{i,j=1}^{N} R(i,j) \quad (7)$$

b) Determinism (%DET):

It is the proportion of diagonal recurrence dots tototal recurrence dots. It is shown in (8).

$$\%DET = \frac{\sum_{l=2}^{N} lP(l)}{\sum_{i,j=1}^{N} R(i,j)} \quad (8)$$

Where, P(l) is the frequency distribution of the length of the diagonal structures in the Recurrence Plot.

c) Shannon Entropy (ENTR):

The Shannon Entropy of the frequency distribution on diagonal line lengths is referred to as ENTR, and it is ameasure of the complication of the recurrence pattern. ENTR is given is eq9.

$$ENTR = -\sum_{l=2}^{N} p(l)\ln p(l) \quad (9)$$

where,

$$p(l) = \frac{P(l)}{\sum_{l=2}^{N} P(l)} \quad (10)$$

d) Length (LEN):

The fourth feature from RQA is LEN, which is the average diagonal length of recurrence plot.

4) Sub-band Energies

The sub-band signals d3,d4,d5,d6,d7 are not used directly as a feature vector, but their sub-band energy is calculated. The sub-band energies is given in the (11).

$$\begin{bmatrix} X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \end{bmatrix} = \begin{bmatrix} log(\sum_n |d_3(n)|) \\ log(\sum_n |d_4(n)|) \\ log(\sum_n |d_5(n)|) \\ log(\sum_n |d_6(n)|) \\ log(\sum_n |d_7(n)|) \end{bmatrix} \quad (11)$$

C. Classification

Due to its built-in mechanisms to ensure proper generalization leading to accurate prediction, the use of fundamental kernel functions to model nonlinear distributions, the ability to train relatively quickly on large data sets, and the utilization of new mathematical optimization techniques, SVM was selected as one of the classifiers. As the primary objective of this research is to showcase the feature explanation, a simple Radial Basis Function (RBF) SVM was used for the classification stage. In contrast to a linear kernel, this kernel tackles the circumstance where the relationship between class labels and characteristics is not linear. It does this by mapping patterns non-linearly to higher dimensional space. Other classifiers such as Random Forest and XGBoost werealso tested.

D. SHAP (SHapley Additive exPlanations)

In many use situations, understanding the reasoning behind a model's predictions is just as crucial as knowing whether or not they were accurate. There is a conflict between accuracy and interpretability because the maximum accuracy for huge datasets is typically achieved by complicated machine learning and deep learning models that even experts have troubleunderstanding.

In the same approach, although multiple highest accurate methods are available in literature, those complex models are not yet ready to be applied in real- world, because in case of a false diagnosis, the model's output has to be explained in order to enhance the performance and avoid such false results in further diagnosis. We have employed SHAP (SHapley Additive exPlanations [4]), a game theory-inspired framework for understanding predictions, to overcome this issue. Foreach prediction, SHAP rates the relevance of each feature.

KernelSHAP was used for feature interpretation. Kernel SHAP is a method that allows computing SHAP values with a very small number of alignment samples. The kernel is based on weighted linear regression where the solution parameters are Shapley values. To build a linear weighted model, n samples are taken, for eachsample a prediction is obtained and the weights are calculated using the kernel size. Finally, the weighted linear model is fitted and the resulting coefficients are Shapely values.

For an instance x, KernelSHAP calculates the involvement of each feature to the estimate made by the model. KernelSHAP consists of the following five steps:

i. Sample the coalitions $Z_k' \in \{0,1\}^M$, $k \in \{1,\ldots,K\}$ (Feature present in coalition is represented by '1', feature absent by '0').
ii. Get estimation for every Z_k' by first translating z' to the original feature space and then applying

model f': $f'(h_x(Z_k'))$

iii. Find the weight for each z' with the SHAP kernel.

iv. Fitting weighted linear model.

v. Return Shapley values ϕ_k, the coefficients from the linear model.

TABLE I. ACCURACY COMPARISION OF DIFFERENT METHOD

Model Type	SVM	Random Forest	XGBoost
Accuracy	100%	95%	90%

IV. EXPERIMENTS AND RESULTS

A. Dataset description

The university of Bonn database [18] is used in this study, there are 100 single-channel EEG recordings of 23.6 seconds. The spectral bandwidth is from 0.5 Hz to 85 Hz, with a sampling rate of 173.61 Hz. The recordings come from a system with 128 channels for acquisition. Five patients' EEG data were separated from a multi-channel EEG recording. The surface EEG of healthy patients was recorded in sets A and B, out of the five sets A, B, C, D, and E, with their eyes closed and open,respectively. Epileptic patients' intracranial EEGs captured in sets C and D during seizures were free from the seizure causing area and outside the seizure generating area respectively. An epileptic patient's intracranial EEG recorded during an epileptic episode is known as Set E. Input data is filter out using a band-pass filter with cut off frequencies 0.53 Hz to 40 Hz. For the proposed method, data from set C and E are used for binary classification. A split of 80:20 was used for training and testing.

B. Training Details

As the task is to assign if the input EEG signal is epileptic positive or negative, the dataset was randomly sampled, followed by standard scalar as a standardization step before model training. Three different machine learning models were trained for the purpose. Support Vector Machine was trained with the following parameters:

- Regularization parameter c - 1.0
- Kernel – Radical Basis Function (RBF)
- Degree = 3
- Gamma – Scale

The Random Forest Classifier was trained using thefollowing parameters:

- Maximum Depth – 2
- No. of estimators – 100
- Criterion – Gini

C. Performance Metrics

The accuracy performance metric was picked to evaluate the models. A ratio of correctly identified samples to all samples can be used to define accuracy. Let TP, TN, FP, and FN stand for True Positive, False Positive, and False Negative, respectively. Accuracy may be denoted as in equation (12)

$$Accuracy = \frac{TP+FP}{TP+FP+TN+FN} \quad (12)$$

The proposed method with 16 extracted features showed the accuracy as mentioned in table I on thetest data. Comparison of proposed method with other authors are listed in table II.

TABLE II. COMPARATIVE ANALYSIS

Methods	Dataset	Accuracy
T. Siddhart et al. [13]	Bern- barcelona	99.7%
L.Chen et. Al [15]	Bonn	92.6%
J. Antonio et. al [20]	Bonn	94.88 %
R. Bairagi et. al [21]	Bonn	99.33 %
Proposed work	Bonn	100 %

D. SHAP Analysis

This part demonstrates the explainable approach to the outputs of the proposed model in classifying input EEG signal. SHAP is a feature explainer for a model. The importance of Shapley score arises when a question aboutwhy a machine learning model resulted in a particular solution is asked. SHAP interprets a particular test case and shows how much each feature contributed for the model to attain a particular output.

In Fig. 2, the average impact of individual features across all test cases on model output is visualized. We canobserve that SVM, which provides the maximum accuracytakes into account each feature, with approximate entropyof d4 sub-band, approximate entropy of d3 sub-band and approximate entropy of d5 sub-band being the feature thatcause maximum variation in the model's output. This explains impact of individual how feature the final output of the machine learning model.

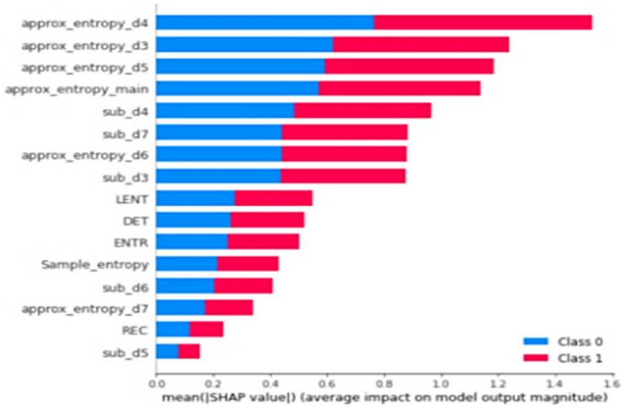

Fig. 2. Mean SHAP value plots of all testcases for SVM

In this scenario, we have used KernelSHAP for interpretation forthe selected features. The blue and red parts represent individual feature's impact on predicting the output class to 0 (epilepsy negative) and 1 (epilepsy positive) respectively.

One more observation that we ascertain is, removing the features which have low Shapley score caused decrease in accuracy and model's performance. Shapley score does not prove directly for feature selection [19]. Instead, one addition observation we make is in the case of false diagnosis. In Fig. 3, the plot in (a) and (b) showsthe Shapley scores for a test case that has been classified correctly by the model, classified as 0 and 1 respectively. Some major features causing this output can be seen as approx. entropy d4, approx. entropy main, approx. entropy d5.

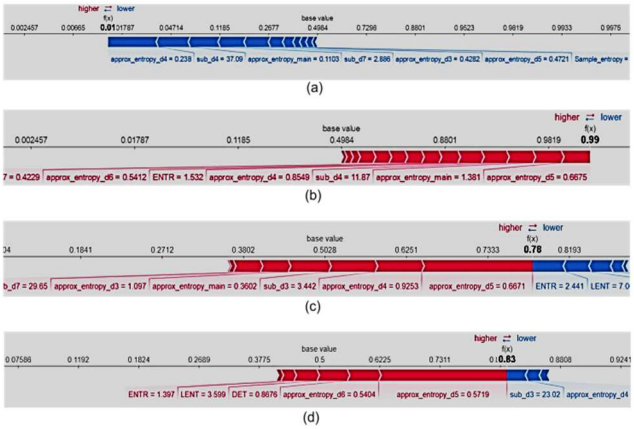

Fig. 3. (a) Shapley plot for correct test case with output 0 (b) Shapley plot for correct test case with output 1 (c & d) Shapley plot for wrong model output 1 (supposed to be 0)

But, in case of false diagnosis as shown in Fig. 3(c)and (d), one observation found in all false test results as compared to all true classifications is the Shapley score of features changes its regular direction and model gets confused to provide a discrete output in such cases. The reason we ascertain for the same is, either problem persisting with the input data, at the data gathering step, or there still might be more chances of model improvement.

V. CONCLUSION

In this study, we propose a method for epilepsy diagnosis from EEG signals. Steps such as pre-processing with standard scaling, and feature extraction are applied. Different machine learning models are trained and tested on the Bonn University EEG dataset, with SVM providing the maximum 100% accuracy in testing. A major addition was to interpret a model's results as opposed to focusing on accuracy. Looking at the major features contributing towards epileptic positive, additional comments such as presence of unusual alpha rhythm in EEG can be justified if a major contribution of alpha band is observed for the classification. In the similar way, if the Shapley score direction shows major variation than common true samples, then in-depth analysis may be required as it can be a case of false diagnosis. Hence, SHAP adds a layer of interpretation over metrics such as accuracy, and explains the reason and effect of particular feature for each test case towards the model's output. Enhancing this work's direction can be adding more methods for further explanation and interpretation at model level, as compared to feature explanation

REFERENCES

[1] Kurakin et. al, *"Adversarial examples in the physical world"*, in Artificial Intelligence Safety and Security, pages 99–112. Chapman and Hall/CRC, 2018.

[2] M. Le Van Quyen et al. "Anticipation of epileptic seizures from standard EEG recordings,'' Lancet, vol. 357, no. 9251, pp. 183–188, Jan. 2001.

[3] J. Paul and T. S. Sivarani, ''Computer aided diagnosis of brain tumor using novel classification techniques,'' J. Ambient Intelligence and Humanized Computing., pp. 1–11, Jul. 2020.

[4] Scott M. Lundberg and Su-In Lee, 2017, "*A unified approach to interpreting model predictions*", in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (*NIPS'17*). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[5] O. A. Ramwala, H. Mulchandani, P. Dalal, M. C. Paunwala and C.N. Paunwala *"COVID-19 Diagnosis from Chest Radiography Images using Deep Residual Network,"* 2020 11th International Conference on Computing, Communication and Networking Technologies(ICCCNT),2020,pp.1-5, doi:10.1109/ICCCNT49239.2020.9225521

[6] Dalal P. et al. (2021), *" Throat Inflammation Based Mass Screening of Covid-19 on Embedded Platform"*. In: Patel K.K., Garg D., Patel A., Lingras P. (eds) Soft Computing and its Engineering Applications. icSoftComp 2020. Communications in Computer and Information Science, vol 1374. Springer, Singapore. https://doi.org/10.1007/978-981-16-0708-0_23

[7] Ojas A. Ramwala et al. (2021), *"Novel Multi-Modal Throat Inflammation and Chest Radiography based Early-Diagnosis and Mass-Screening of COVID-19"*, In The Open Biomedical Engineering Journal, Vol. 15, Issue 1, pages 226-234, doi: http://dx.doi.org/10.2174/1874120702115010226

[8] A. Vora, C. N. Paunwala and M. Paunwala, *"Statistical analysis of various kernel parameters on SVM based multimodal fusion,"* in 2014 Annual IEEE India Conference (INDICON), 2014, pp. 1-5, doi: 10.1109/INDICON.2014.7030414

[9] S. K. Parmar, O. A. Ramwala and C. N. Paunwala, *"Performance Evaluation of SVM with Non-Linear Kernels for EEG-based Dyslexia Detection,"* in 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC), 2021, pp. 1-6, doi: 10.1109/R10- HTC53172.2021.96

[10] P. Dalal et al., *"Statistical feature rich Deep Learning based Epileptic Seizure detection"*, accepted at 2022 IEEE Region-10 Symposium (TENSYMP).

[11] H. Gabani et. al, *"ApEn-Based Epileptic EEG Classification Using Support Vector Machine"*, in Patel Z., Gupta S. (eds) FutureInternet Technologies and Trends, pp. 75-85, 2017

[12] H. Gabani, et. al, *"EEG Signal Classification for Epileptogenic Zone and Seizure Zone"*, in Proceedings of the International Conference on Intelligent Systems and Signal Processing, vol 671 pp. 45-52, 2018.

[13] T. Siddharth et. al, *"Discrimination of Focal and Non-Focal Seizures From EEG Signals Using Sliding Mode Singular Spectrum Analysis,"* in IEEE Sensors Journal, vol. 19, no. 24, pp. 12286-12296, 15 Dec.15, 2019, doi: 10.1109/JSEN.2019.293990

[14] Roger, L. Torlay, J. Gardette, C. Mosca, S. Banjac, L. Minotti,P. Kahane, and M. Baciu, *''A machine learning approach to explore cognitive signatures in patients with temporo-mesial epilepsy,''* in Neuropsychology, vol. 142, May 2020, Art. no. 107455.

[15] Lan-Lan Chen et. al , *"A framework on wavelet-based nonlinear features and extreme learning machine for epileptic seizure detection,"* in Biomedical Signal Processing and Control, Volume 10, 2014, pp. 1- 10, ISSN 1746-8094

[16] Rich Caruana et. al, *"Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission"*, in*Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD '15*), Association for Computing Machinery, New York, NY, USA, 1721–1730. DOI:https://doi.org/10.1145/2783258.2788613

[17] Hind, M.; Wei, D.; Campbell, M.; Codella, N. C. F.; Dhurandhar, A.; Mojsilovic, A.; Ramamurthy, K. N.; and Varshney, K. R.2019, *"TED: Teaching AI to Explain its Decisions"* in AAAI/ACM Conference on AI, Ethics, and Society, 123–129.

[18] G. Andrzejak et. al,*"Indications of nonlinear deterministic and finite-dimensional structures intime series of brain electrical activity: Dependence on recording region and brain state"*, Physical Review E 64 (6) (2001) 061907

[19] Fryer et. al, (2021), *"Shapley values for feature selection: The good, the bad, and the axioms."* in *IEEE Access*, vol. 9, pp. 144352-144360, 2021, doi: 10.1109/ACCESS.2021.3119110

[20] J. A. de la O Serna, M. R. A. Paternina, A. Zamora-Méndez, R. K. Tripathy and R. B. Pachori, "EEG-Rhythm Specific Taylor–Fourier Filter Bank Implemented With O-Splines for the Detection of Epilepsy Using EEG Signals," in IEEE Sensors Journal, vol. 20, no. 12, pp. 6542-6551, 15 June15, 2020, doi:10.1109/JSEN.2020.2976519.

[21] R. N. Bairagi and M. Maniruzzaman, "Identification of Epileptic Seizure in EEG Signals Using DWT and ANN," 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 142-145, doi: 10.1109/TENSYMP50017.2020.9230746