Explainability Using SHAP for Epileptic Seizure Recognition

Simone A. Ludwig
North Dakota State University
Fargo, ND, USA
simone.ludwig@ndsu.edu

***Abstract*—With the help of an electroencephalogram (EEG) the electrical activity of the brain is measured, and this can help identify chronic neurological disorders such as epilepsy. Epileptic episodes are detected by monitoring patients in order to provide preventive measures. Current research studies are using a combination of time and frequency features to recognize epileptic seizures automatically. In order to automatically detect epileptic seizures, different machine learning approaches have been used. Gradient boosting decision tree (GBDT) is a machine learning technique that is known for its efficiency, accuracy, and interpretability. In terms of performance of GBDT, many machine learning tasks such as multi-class classification, learning to rank, etc. have reported competitive performance. In this paper, epileptic seizure recognition data is investigated and split into a binary and multi-class data set for which the GBDT method is applied. In addition, the SHAP (Shapley Additive Explanations) method is used as an explanation tool to interpret the machine learning models that are produced via training for both the binary and the multi-class data set.**

***Index Terms*—Epilepsy data set, Gradient boosting decision tree algorithm, multi-class classification, Shapley Additive Explanations (SHAP) method.**

I. Introduction

Abnormal electrical activity is measured in the brain for an epileptic episode. An episode usually occurs suddenly and thus an automated way to monitor and detect this has to be devised so that that the patient and neurologist can be warned in advance [1]. One difficulty though is to detect the uncertain time frequencies of epileptic episodes. These are unfortunately not easy to detect.

Many different ways were looked into in order to directly measure 'epilepsy signals' from brain signals. The single photon emission computed tomography (SPECT), positron emission tomography (PET), functional magnetic resonance imaging (fMRI), and magnetic resonance imaging (MRI) [2] are examples; with many research studies making use of video-electroencephalograms (V-EEGs) [3], [4]. V-EEGs is currently seen as the best approach to study epilepsy. The reason being that the physiological processes of a seizure are typically dynamic, non-stationary, and nonlinear, and thus the differentiation of rhythmic changes from non-stationary processes provides challenges to the analysis of the signals in a EEG.

For the automation to detect EEG signals two tasks are involved. The first is that the features need to be extracted and the second is classification task. There are four categories that the extracted features can be divided into, namely entropy features, fractal dimension features, time-frequency domain features, and statistical features. Most research studies have used a combination of time and frequency features for the epileptic seizure recognition.

In this paper, the epileptic seizure recognition data set is investigated as a 2-class and 5-class problem. In particular, a gradient boosting decision tree algorithm is used for the classification part followed by the SHAP (Shapley Additive Explanations) method to explain the output of the gradient boosting decision tree models for both the binary and the multi-class data set.

II. Related Work

Most of the related research has focused on the Bonn data sets [5]. In the field of brain science, a neural network (NN) classification technique was applied [6]. Another approach used support vector machines (SVMs) to identify epilepsy patients. This work resulted in good recognition performance as shown in [7], [8], [4]. Furtermore, another related approach is the least squares support vector machine (LS-SVM) [9]. LS-SVM classifies two-class seizure and non-seizure EEG signals. An accuracy ranging between 98.0-99.5% was achieved using radial basis function (RBF) kernel, with 99.5-100% achieved with the Morlet kernel function.

An Ada-Boost classifier was applied in [10]. The approach achieved good accuracy for epileptic seizures detection. Given the no-free-lunch theorem [11], several different kinds of classification algorithms were applied to seizure detection. Examples of classification algorithms include K-nearest neighbors (KNN) [12], Bayesian neural networks, and random forests (RF) [13] with accuracy results ranging from 93% to 99.66%. However, these approaches only used binary classification and are time consuming and thus not practical for certain clinical applications.

Given the data set used in this paper, directly related research use deep learning methods to predict epileptic seizures in [14]. The researchers used deep learning to distinguish the signals 'before' and 'after' a seizure using held-out data from all patients. A comparison is done with a random predictor using a modified system to adjust for each patient's feature set. The prediction system could either choose 'sensitivity' or 'time in warning' for each patient and thus provide time and functional seizure prediction.

A trained deep neural networks with EEG data for predicting the seizure is presented in [15]. Spectral, temporal and spatial information was recorded for the analysis and the study focused on cross-patients. The finding was that the deep learning model generalizes very well base on the different patient data provided.

The author of this paper has also investigated this particular data set before by conducting a performance analysis of different ensemble configurations [16] as well as a comparison of a deep neural network ensemble method with the Choquet Fuzzy Integral Fusion method [17].

In [18], traditional machine learning algorithms, such as KNN (K Nearest Neighbors), Logistic Regression, and Linear SVM were applied to predict seizures. In addition, CNN (Convolutional Neural Network), RNN (Recurrent Neural Networks), and LSTM (Long Short-Term Memory) were used.

III. Proposed Approach

In this section, the methods applied are described. First, gradient boosting decision tree (GBDT) algorithm is described followed by the SHAP method, which is applied to the resulting GBDT model run on the epileptic seizure recognition data set.

A. Gradient Boosting Decision Tree - LightGBM

In this paper, a gradient boosting decision tree (GBDT) algorithm is used as the classifier. GBDT represents an ensemble model of decision trees whereby the most time-consuming portion is the learning of the best split points for each feature. There are two major algorithms, one is the pre-sorted algorithm [19], [20], which enumerates all possible split points on the pre-sorted feature values. The second algorithm is the histogram-based algorithm N10,N11 which finds the split points on the sorted feature values using histogram-based buckets putting continuous feature values into bins, and then uses these bins to construct feature histograms during training. However, both algorithms have a high memory consumption as well as long training times especially when big data sets with a large number of features are involved.

LightGBM [21] is the implementation that addresses the scalability and efficiency with two improvements. The first improvement uses a gradient-based one-side sampling (GOSS) and the second is the exclusive feature bundling (EFB). GOSS works as follows making use of the information gain measure. Instances with larger gradients will contribute more to the information gain. Thus, for downsampling the data instances, the accuracy of the information gain estimation should be retained and thus those instances with larger gradients should be retained and the one with smaller gradients should be discarded. The EFB method makes use of the sparsity of real-world data. Thus, the design of a lossless approach to reduce the number of effective features can be done. In particular, in a sparse feature space many features are exclusive, which means that they rarely take nonzero values simultaneously. Thus, such exclusive features can be bundled using an optimization algorithm similar to the graph coloring problem solving the optimization using a greedy algorithm with a constant approximation ratio.

B. Shapley Additive Explanations - SHAP

The correct interpretation of a prediction model's output is a very important issue. The interpretation provides insight into how a model may be improved, and facilitates the understanding of the application/process being modeled. For some application, simple models are often preferred since their interpretation is easy to understand and to follow even though the model might be less accurate than a more complex one. However, with the growing need of big data processing, the benefits of using complex models are needed but this results in the trade-off between accuracy and interpretability of a model's output. Many different methods have been proposed to address this issue [22], [23], [24], [25], [26], [27]. However, how these different methods relate and which method to use under which circumstances is still an open question. Thus, Shapley Additive Explanations (SHAP) was introduced.

The SHAP is an extension of the Shapley value. In particular, it was inspired by several methods on model interpretability, the SHAP value is used as a united approach to explain the output of any machine learning model. SHAP exhibits three benefits:

- Global interpretability: the collective SHAP values show the contribution each predictor (either positively or negatively) makes to the output variable. The variable importance plot shows the positive or negative relationship of each variable with the output.
- Local interpretability: each observation has its own set of SHAP values, which increases its transparency. Thus, why a case receives its prediction and the contributions of the predictors can be explained. Traditional variable importance algorithms have only been able to show the results across the entire population but not for each individual case.
- Versatility: the SHAP values can be calculated for any tree-based model, while other methods need to use linear regression or logistic regression.

More information on the SHAP framework can be found in [28].

IV. Experiments and Results

A. Description of Data Set

The Epileptic Seizure Recognition data set [29] containing 4,097 data points was selected. Each data point is collected from a EEG recording. Five-hundred individuals were recorded to obtain the data. More details on the data set can be found in [16].

This data set has been mostly used as a binary data set where class 1 was classified as 'epileptic seizure', and classes 2, 3, 4 and 5 were categorized as 'no epileptic seizure'. In this paper, both the binary and also the 5-class data sets are used to conduct experiments on.

The class distribution is as provided in Table IV-A totaling in 8,627 samples/rows.

Class	Value
1	1,735
2	1,732
3	1,693
4	1,744
5	1,726

Parameter	Value
max bin	512
learning rate	0.05
boosting type	gbdt
objective	binary
metric	binary logloss
num leaves	15
min data	100
boost from average	True

B. Results: 2-Class Model Classification

Figure 3 shows the decision tree model using LightGBM with the parameters as provided in Table IV-B. The model tree has a depth of 11, which represent how the classification is being made for unseen data. The tree model is quite natural and can be easily understood.

Figure 1 and 2 show the ROC curve and the precision-recall curve, respectively. The accuracy that was achieved by the model was 85.02%.

Table I shows the precision, recall, f1-score and support whereas Figure 5 shows the confusion matrix. As can be seen in the table, the precision for the Yes class (having a seizure) is 0.86, whereas a for the No class (not having a seizure) it is 0.84. The corresponding values for recall are 0.83 and 0.87, respectively. The confusion matrix shows that 1,870 samples were correctly classified for the Yes class whereas 2,042 samples were correctly classified for the No class. The other two entries are false positives and false negatives, respectively.

The SHAP values reporting on the most important variables are show in Figure 6. Please note that Class 0 represents the Yes class and Class 1 represents the No class. The color represents the SHAP value for each feature.

The figure shows the density scatter plot of SHAP values for each feature highlighting the impact each feature has on the model output based on the test set. The sum of the SHAP value magnitudes are sorted across all samples in order to identify the features with the biggest impact. For example, we can see that Feature 167 has the biggest effect and thus contributes most to the overall classification. As can be seen, Feature 167 has a higher total model impact than Feature 14, but for those samples where Feature 14 matters it has more impact than Feature 28. Thus, Feature 14 strongly influences a few predictions, while Feature 28 influences all predictions by a lesser amount.

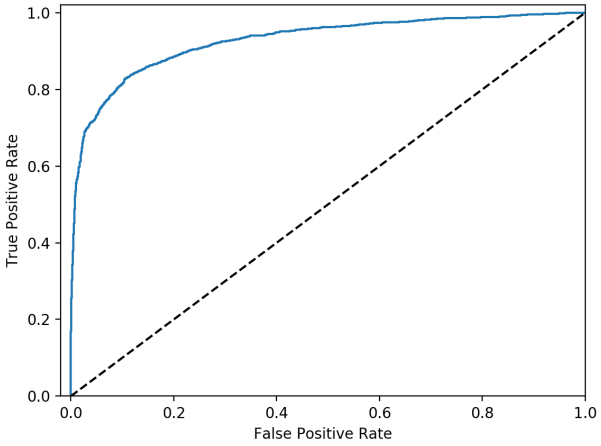

Fig. 1. ROC Curve of 2-Class Model

TABLE I
PRECISION, RECALL, F1-SCORE, AND SUPPORT FOR 2-CLASS MODEL

	Precision	Recall	F1-score	Support
Yes	0.86	0.83	0.84	2249
No	0.84	0.87	0.86	2352
Accuracy			0.85	4601
Macro avg	0.85	0.85	0.85	4601
Weighted avg	0.85	0.85	0.85	4601

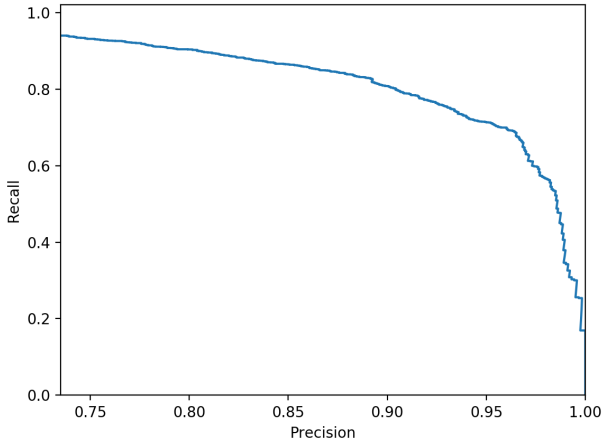

Fig. 2. Precision Recall Curve of 2-Class Model

Fig. 3. Decision Tree of 2-Class Model

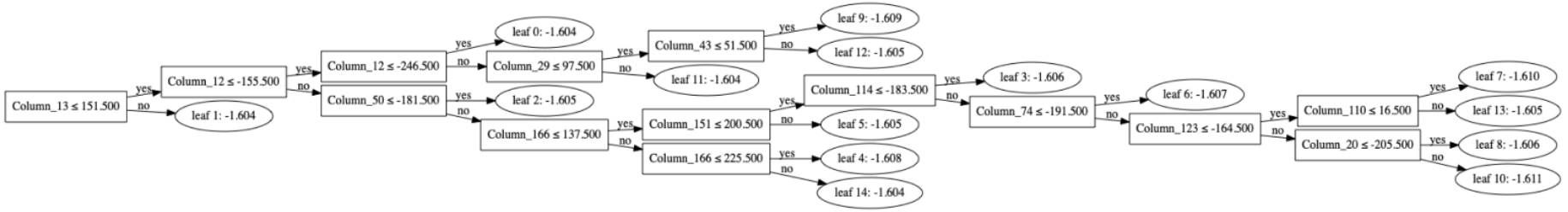

Fig. 4. Decision Tree of 5-Class Model

Parameter	Value
max bin	512
learning rate	0.002296
boosting type	gbdt
objective	multiclass
metric	multi logloss
num leaves	15
max depth	10
feature fraction	0.4
bagging fraction	0.6
bagging freq	15

C. Results: 5-Class Model Classification

LightGBM was run with the parameters as provided in Table IV-C.

Figure 4 shows the decision tree model for the 5-Class data set. The tree has a depth of 9 and is compared to the 2-Class model wider, i.e., branching factor is 3.

The confusion matrix of the 5-Class model is shown in Figure 7. We can see the correct predictions in the diagonal of the matrix. The other entries represent the missclassifications based on the class label.

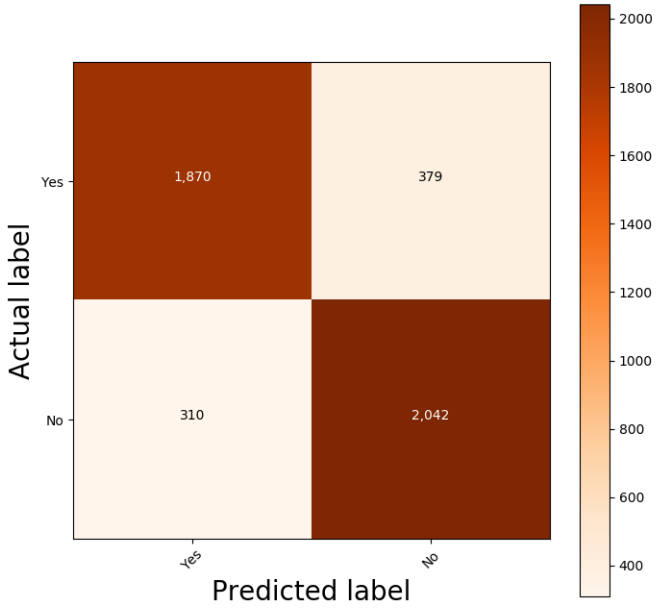

Fig. 5. Confusion Matrix of 2-Class Model

Table II shows the precision, recall, f1-score, and support for the 5-Class model. The accuracy of the resulting model was 71.76%, which is relatively good for a 5-Class model outcome.

Figure 8 shows the SHAP values for the 5-Class model. As also seen in Figure 6, each feature in the plot shows the importance and also their range of effects over the data set. The color in the plot shows how changes in the value of a feature effect the change the different classes of the seizure data set. As can be seen, the SHAP values explain the margin output of the model, which is the change in the log odds for a Cox proportional hazards model. We can see that Feature 2 contributes most to the 5-Class model followed by Feature 14 and Feature 0.

TABLE II
PRECISION, RECALL, F1-SCORE, AND SUPPORT FOR 5-CLASS MODEL

	Precision	Recall	F1-score	Support
1	0.95	0.93	0.94	575
2	0.62	0.55	0.58	575
3	0.60	0.63	0.62	575
4	0.76	0.75	0.76	575
5	0.65	0.72	0.69	575
Accuracy			0.72	2875
Macro avg	0.72	0.72	0.72	2875
Weighted avg	0.72	0.72	0.72	2875

V. CONCLUSION

This paper investigated the epileptic seizure recognition data set, which was split into a binary and a multi-class data set for which the Gradient boosting decision tree (GBDT) method was applied. The analysis included accuracy, precision, f1-score, support, and confusion matrix. Moreover, the decision tree models that were produced from both data sets were provided. The SHAP (Shapley Additive Explanations) method was used in order to explain the output of the machine learning models.

The models of the GBDT are easily interpretable since they are in the form of a decision tree. The accuracy of the 2-Class model was 85.02% and for the 5-Class model 71.76% was achieved. However, for a researcher to identify which of the features contains the most importance on the built model, the SHAP summary plots were analyzed. The plots nicely show the effect or contribution which feature makes on each class. This allows for a better explanation ability of the built model and also helps to guide decision makers in explaining the machine learning model that was produced.

REFERENCES

[1] F. Leijten. Multimodal seizure detection: A review. Epilepsia. 59:42-47, 2018. doi: 10.1111/epi.14047.

[2] S.S. Spencer. MRI, SPECT, and PET imaging in epilepsy: Their relative contributions. Epilepsia. 35:S72-S89, 1994. doi: 10.1111/j.1528-1157.1994.tb05990.x.

[3] C.Á. Szabó, L.C. Morgan, K.M. Karkar, L.D. Leary, O.V. Lie, M. Girouard, J.E. Cavazos. Electromyography-based seizure detector: Preliminary results comparing a generalized tonic-clonic seizure detection algorithm to video-EEG recordings. Epilepsia. 56:1432-1437, 2015. doi: 10.1111/epi.13083.

[4] Y. Gu, E. Cleeren, J. Dan, K. Claes, W.V. Paesschen, S.V. Huffel, B. Hunyadi. Comparison between Scalp EEG and Behind-the-Ear EEG for Development of a Wearable Seizure Detection System for Patients with Focal Epilepsy. Sensors. 18:29, 2018. doi: 10.3390/s18010029.

[5] Y. Wang, Z. Li, L. Feng, H. Bai, C. Wang. Hardware design of multiclass SVM classification for epilepsy and epileptic seizure detection. IET Circuits Devices Syst. 12:108-115, 2018. doi: 10.1049/iet-cds.2017.0216.

[6] Q. He, B. Wu, H. Wang, L. Zhu. VEP Feature Extraction and Classification for Brain-Computer Interface; Proceedings of the 8th International Conference on Signal Processing; Guilin, China. 16-20 November 2006.

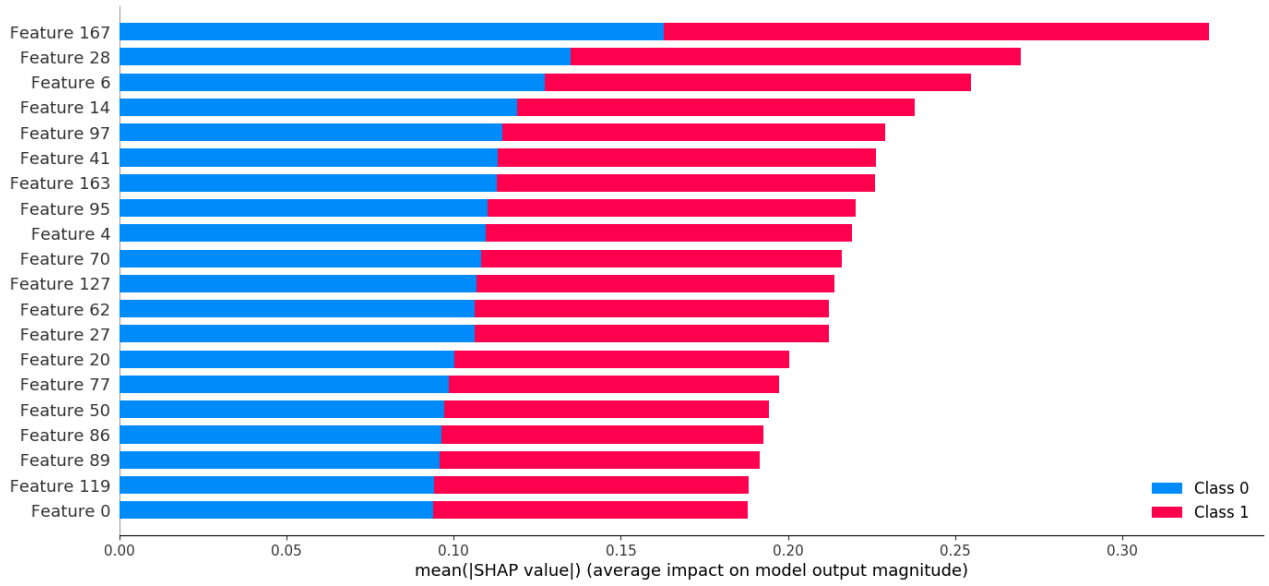

Fig. 6. SHAP Plot of 2-Class Model

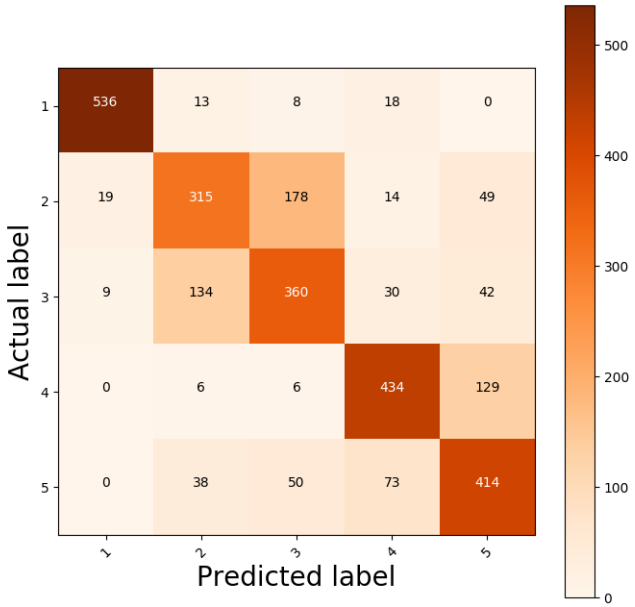

Fig. 7. Confusion Matrix of 5-Class Model

[7] B.E. Boser, I.M. Guyon, V.N. Vapnik. A Training Algorithm for Optimal Margin Classifiers; Proceedings of the Annual Workshop on Computational Learning Theory; Pittsburgh, PA, USA, pp. 144-152. 27-29 July 1992.

[8] K. Fu, J. Qu, Y. Chai, T. Zou. Hilbert marginal spectrum analysis for automatic seizure detection in EEG signals. Biomed. Signal Process. Control. 18:179-185, 2015. doi: 10.1016/j.bspc.2015.01.002.

[9] K.D. Brabanter, P. Karsmakers, F. Ojeda, C. Alzate, J.D. Brabanter, K. Pelckmans, B.D. Moor, J. Vandewalle, J.A.K. Suykens. LS-SVMlab Toolbox User's Guide: Version 1.7. Ku Leuven Leuven; Leuven, Belgium: 2010.

[10] Y.C. Liu, C.C.K. Lin, T. Jing-Jane, Y.N. Sun. Model-Based Spike Detection of Epileptic EEG Data. Sensors. 13:12536-12547, 2013. doi: 10.3390/s130912536.

[11] D.H. Wolpert, W.G. Macready, No Free Lunch Theorems for Optimization, IEEE Transactions on Evolutionary Computation 1, 67, 1997.

[12] R.M. Isa, I. Pasya, M.N. Taib, A.H. Jahidin, W.R.W. Omar, N. Fuad, H. Norhazman. EEG brainwave behaviour due to RF Exposure using kNN classification; Proceedings of the IEEE International Conference on System Engineering and Technology; Shah Alam, Malaysia. 19-20 August, pp. 385-388, 2013.

[13] R. Chai, Y. Tran, G.R. Naik, T.N. Nguyen, S.H. Ling, A. Craig, H.T. Nguyen. Classification of EEG based-mental fatigue using principal component analysis and Bayesian neural network; Proceedings of the Engineering in Medicine and Biology Society; p. 4654, Orlando, FL, USA, 16-20 August 2016.

[14] I. Kiral-Kornek, S. Roy, E. Nurse, B. Mashford, P. Karoly, T. Carroll, D. Payne, S. Saha, S. Baldassano, T. O'Brien, D. Grayden. Epileptic seizure prediction using big data and deep learning: toward a mobile system. EBioMedicine, 27, pp.103-111. Vancouver, 2018.

[15] P. Thodoroff, J. Pineau, A. Lim. Learning robust features using deep learning for automatic seizure detection. In Machine learning for healthcare conference (pp. 178-190), 2016.

[16] S.A. Ludwig, Performance Analysis of Different Ensemble and Choquet Configurations Applied to the Multi-label Classification for Epileptic Seizure Recognition, Journal of Artificial Intelligence and Soft Computing Research, vol. 12, no. 1, pp. 5-17, 2022.

[17] S. A. Ludwig, Epileptic Seizure Recognition: Deep Neural Network Ensemble versus Choquet Fuzzy Integral Fusion, 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, Australia, December 2020.

[18] H. Liu, L. Xi, Y. Zhao, Z. Li. Using Deep Learning and Machine Learning to Detect Epileptic Seizure with Electroencephalography (EEG) Data, arXiv, eprint:1910.02544, 2019.

[19] M. Mehta, R. Agrawal, and J. Rissanen. Sliq: A fast scalable classifier for data mining. In International Conference on Extending Database Technology, pages 18-32. Springer, 1996.

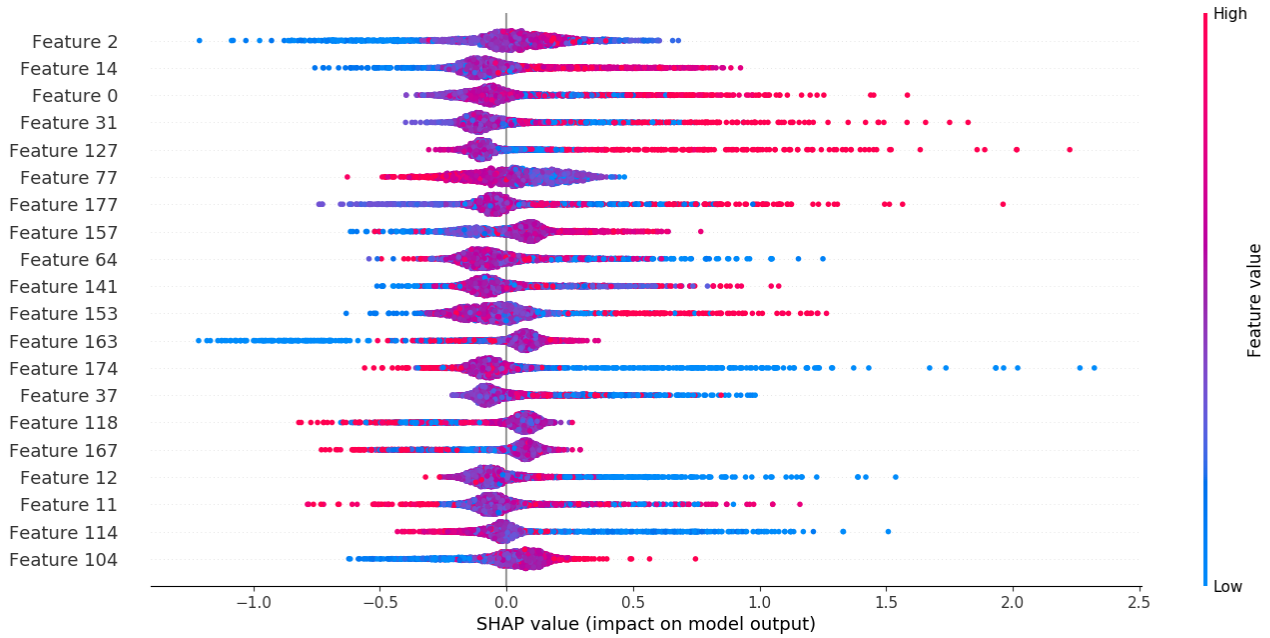

Fig. 8. SHAP Plot of 5-Class Model

[20] J. Shafer, R. Agrawal, and M. Mehta. Sprint: A scalable parallel classi er for data mining. In Proc. 1996 Int. Conf. Very Large Data Bases, pages 544-555. Citeseer, 1996.

[21] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: a highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 3149-3157, 2017.

[22] M.T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. pp. 1135-1144, 2016.

[23] A. Shrikumar et al. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. In: arXiv preprint arXiv:1605.01713, 2016.

[24] E. Štrumbelj and I. Kononenko. Explaining prediction models and individual predictions with feature contributions. In: Knowledge and information systems 41.3, pp. 647-665, 2014.

[25] A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In: Security and Privacy (SP), IEEE Symposium on. IEEE. pp. 598-617, 2016.

[26] S. Lipovetsky and M. Conklin. Analysis of regression in game theory approach. In: Applied Stochastic Models in Business and Industry 17.4, pp. 319-330, 2001.

[27] S. Bach et al. On pixel-wise explanations for non-linear classifier decisions by layerwise relevance propagation. In: PloS One 10.7, e0130140, 2015.

[28] S.M. Lundberg and S.-I. Lee. 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, 2017.

[29] R.G. Andrzejak, K. Lehnertz, C. Rieke, F. Mormann, P. David, C.E. Elger. Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state, Phys. Rev. E, 64, 061907, 2001.