International Journal of Bio-Science and Bio-Technology
Vol.12, No.1 (2020), pp.35-46
http://dx.doi.org/10.21742/ijbsbt.2020.12.1.05

Analyzing Brain Signals to Predict Seizure Events using Machine Learning Techniques

Jinan Fiaidhi[1*], Tejas Wadiwala[2] and Vikas Trikha[3]

Department of Computer Science, Lakehead University, Canada
[1]jfiaidhi@lakeheadu.ca, [2]twadiwal@lakeheadu.ca, [3]vtrikha@lakeheadu.ca

Abstract

This paper attempts to perform a comparative analysis on brain signals datasets to predict seizure events using various machine learning classifiers such as random forest, gradient boosting, support vector machine and extra trees classifier. The experimentation on these classifiers has been performed using the Rochester Institute of Technology EEG Dataset. The comparative analysis is measured based on the classifiers performance parameters such as accuracy, area under the ROC curve (AUC), specificity, recall, and precision. EEG signals are usually captivated to diagnose the problems related to the electrical activities of the brain as it tracks and records brain wave patterns to produce a definitive brain seizure activities. While exercising machine learning practices, various data preprocessing techniques were implemented to attain cleansed and organized data to predict better results and higher accuracy. Section II gives a comprehensive survey of existing work performed so far, while section III sheds light on the dataset used for this research.

***Keywor**ds: Electroencephalogram Signals, Brain activities, Machine learning, Support Vector Machine, Binary classification, Extra trees classifier*

1. Introduction

A seizure is a persistent disorder of the nervous system that happens in the brain due to unwanted electrical activities and affects nearly 50 million people worldwide and gets worse in developing countries. According to a European Union Survey, a population aged 65 or above is predicted to rise from 16.4% (2004) to 29.9% (2050) with epileptic seizures. It can be commonly referred to as epileptic seizures, which can be prompted due to numerous causes like brain injuries, brain tumors, low oxygen during birth, or any hereditary reasons. A typical seizure can lead to jerky movements, temporary confusion, loss of consciousness, or staring spell and it can last up to a few seconds to 5 minutes. For testing and diagnosing seizures, EEG comes in to play. EEG captures brain wave activities and illustrates the recordings in the structure of graphs, which are ordinarily recognized as EEG signals. It is recorded using Brain-Computer Interface (BCI), which can be invasive, semi-invasive, or non-invasive. A BCI is a computer-based system that receives brain signals, investigates them, and decodes them into commands that are transferred to an output device to communicate the aspired action. In my opinion, any brain signal could be utilized to establish a BCI system. BCI based seizure apprehension technology is beginning to exercise the influence of the mind to overcome the shortcomings of the body. The usefulness of brain seizure detection to

Article history:
Received (March 3, 2020), Review Result (April 11, 2020), Accepted May 26, 2020)

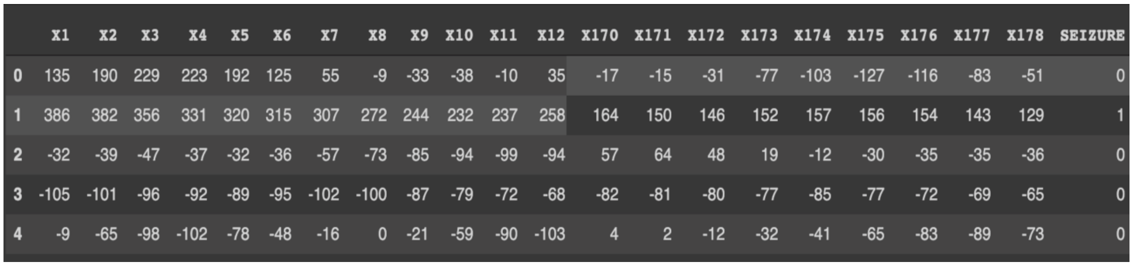
Print ISSN: 2233-7849, eISSN: 2208-9810 IJBSBT

investigate cognitive control in patients with neurological conditions produces new insight into the privileges for health care. In seizure detection systems, the objection is to recognize and engage the most potent signal processing an algorithm within many comparisons for the specific application. Computerized interpretation of EEG records in the investigation of epilepsy was started in the initial 1970s. Several algorithms for spike detection have been intended, including mimetic- and rule-based approaches [1], frequency-domain methods [2], ANNs [3], independent component analysis [4], data mining, template matching [5], and topographic classification [6].

2. Literature review

R.Vaitheeshwari et al. has proposed an artificial neural network approach on the epileptic seizure dataset. There has been preprocessing performed on the dataset, and then an artificial neural network has been proposed, which has six layers. The model is made up of an input layer with 100 units, followed by four fully connected of 100 nodes with the Rectified Linear Unit (ReLU) activation function. Similarly, another model has been performed, which has a Dropout Layer in it. In the feature extraction phase, R.Vaitheeshwari et al. have extracted mean and standard deviation from the dataset. There are three types of scaling that has been performed on the dataset: Standard Scaler, Min-Max Scaler, and Robust Scaler. Preprocessor analysis is then applied to the dataset to make the feature vectors nearly equal so that one feature does not dominate the other while examining the accuracy. There has been experimental analysis performed using different optimizers and comparisons made between them, of which the optimizer giving the highest accuracy is Stochastic Gradient Descent (SGD). Then a comparison between different learning rates has been made using SGD optimizer at 150 epochs, and it has been found that learning rate: 0.01 gives the highest accuracy, along-with different performance metrics that include accuracy, recall, f1 score, and precision [7].

Dr.R.Shantha Selva Kumari et al. have performed an analysis on the EEG epileptic dataset, which is provided by the University of Bonn, Germany. They have performed the analysis in three stages. Firstly, the discrete wavelet transform is used to decompose the EEG signal into a delta, theta, and gamma sub bands. Feature extraction is the second step where the statistical features are extracted from each sub band. In the final step, classification has been done on the EEG signal to predict if the person has a seizure or not. This classification has been done using Support Vector Machine (SVM); also, linear kernel function has been used for the same. The above methodology has been applied to two different groups of EEG signals: the first one is a healthy EEG dataset, the second one is the epileptic dataset during a seizure interval. The accuracy of performing the above steps is said to be quite reasonable by the authors [8].



	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X170	X171	X172	X173	X174	X175	X176	X177	X178	SEIZURE
0	135	190	229	223	192	125	55	-9	-33	-38	-10	35	-17	-15	-31	-77	-103	-127	-116	-83	-51	0
1	386	382	356	331	320	315	307	272	244	232	237	258	164	150	146	152	157	156	154	143	129	1
2	-32	-39	-47	-37	-32	-36	-57	-73	-85	-94	-99	-94	57	64	48	19	-12	-30	-35	-35	-36	0
3	-105	-101	-96	-92	-89	-95	-102	-100	-87	-79	-72	-68	-82	-81	-80	-77	-85	-77	-72	-69	-65	0
4	-9	-65	-98	-102	-78	-48	-16	0	-21	-59	-90	-103	4	2	-12	-32	-41	-65	-83	-89	-73	0

Figure 1. Overview of dataset

Alexandros T. Tzallas et al. have demonstrated the suitability of time-frequency (t-f) analysis to classify EEG segments for epileptic seizures, and they have compared several methods for the t-f analysis of EEGs. They have used a benchmark EEG dataset, and they have presented qualitative and quantitative results. They have utilized an approach based on t-f analysis and extraction of features reflecting the distribution of the signal's energy over the t-f plane. The analysis is performed in three stages: the first one is t-f analysis and calculation of the power spectrum density (PSD) of each EEG segment, the second one is feature extraction, measuring the signal segment fractional energy on specific t-f windows, the final part is the classification of the EEG segment (existence of epileptic seizure or not), using artificial neural networks [9].

Dattaprasad A. Torse et al., 2019, have used recurrence plots and machine learning techniques to classify epileptic seizures. Nonlinear techniques are used to examine the EEG signals because EEGs are nonlinear in nature. They have proposed a nonlinear technique of extracting features of EEG which is based on Recurrence Plots (RP), and Recurrence Quantification Analysis (RQA). The parameters derived from the RP have been used to categorize the EEG signal information as pre-ictal, ictal, and normal classes. RP is said to be an advanced technique of nonlinear data analysis, and the RQA parameters of RP compute the significant features of signals. These extracted features have been classified using Probabilistic Neural Network (PNN), Artificial Neural Network (ANN), and Support Vector Machine (SVM). The SVM system is found to have the highest prediction accuracy of 91.2% and is selected for the classification by the authors [10].

3. The dataset

The dataset on which we have performed brain signal analysis is Epileptic Seizure Recognition Dataset. It is a time-series dataset and has been collected from UCI Machine Learning Repository [11] which has been provided by Rochester Institute of Technology. This dataset is derived from a source dataset, this dataset has 5 different folders in the source dataset, having 100 files each, where each record represents a single person/subject. Each file represents the recording of brain activity for a total of 23.6 seconds. This dataset is sampled into 4097 data points. A data point represents the value of EEG recording at a different point in time. Therefore, in the source dataset, there are a total of 500 persons whose recording over 4097 points is taken over 23.5 seconds.

The dataset which we have used contains 178 data points for 1 second. These 178 data points are derived by shuffling the source dataset of 4097 data points into 23 pieces, where each piece contains 178 data points for 1 second. Therefore, there are a total of 23 x 500 = 11500 rows and 179 columns in the dataset.

The ground truth is represented in the last column of data set y= {1,2,3,4,5} where:

5 - eyes open, means when they were recording the EEG signal of the brain the patient had their eyes open

4 - eyes closed, means when they were recording the EEG signal the patient had their eyes closed

3 - they identified where the region of the tumor was in the brain and recorded the EEG activity from the healthy brain area

2 - they recorded the EEG from the area where the tumor was located

1 - recording of seizure activity [11]

Therefore, any subject that falls under the category of 2, 3, 4, and 5 do not have a seizure. Only subjects that fall under class 1 have an epileptic seizure. [Figure 1] represents an overview of dataset, since there are 178 columns that are to be displayed; we have not shown the columns between X12 and X170.

4. Dataset preprocessing and feature engineering

4.1. Creating a new column

We will be using the pandas library, which is an open-source data analysis and manipulation tool. First, the data is read into a data frame; then we insert a new column. This new column is depended on the y column which has values 1, 2, 3, 4, and 5; of which we put an if loop and keep only two values, i.e., 0 and 1 where 0 represents person does not have a seizure, and 1 represents person has a seizure.

4.2. Calculating prevalence

The percentage of the sample whose characteristic we are trying to predict is called prevalence. We have defined a function that helps in calculating the prevalence. After estimating the prevalence of positive class, it is seen to have a 20% prevalence, i.e., there is an imbalance of positive and negative class, which is represented in [Figure 2].

4.3. Maintaining authenticity and ambiguity

For maintaining authenticity and ambiguity, we perform a few checks on our dataset that include:

- Checking for duplicated columns.
- Checking the dimensions of the dataframe.
- Making sure that there is no order associated with our samples, we randomly shuffle the dataset and then provide them with a new index.

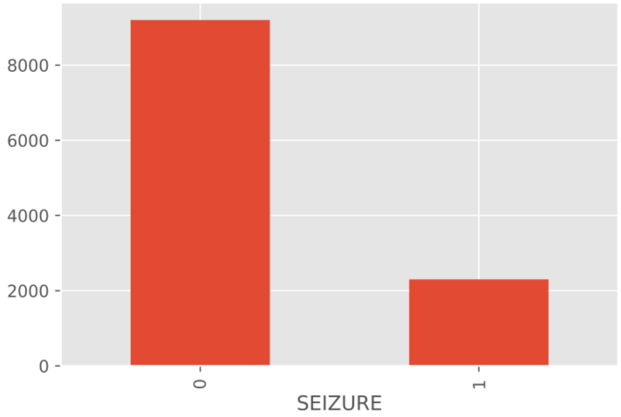


Figure 2. Proportion of imbalanced data

4.4. Data splitting

We split our dataset into training, testing and validation sets, where training consists of 70 percent of the dataset, and testing and validation consists of 15 percent each. After that, we again perform prevalence checks on the split data, and we find that there still is data imbalance, and the prevalence of positive class is 20 percent for each set, i.e., training, testing, and validation.

4.5. Balancing the dataset

To balance the dataset, we can either oversample the data set or sub-sample it. Oversampling means generating new samples of the underrepresented data. Sub-sampling means the down-sampling of data to balance the prevalence. On our dataset, we have performed sub-sampling by selecting the minimum number of values that the positive class has and then taking out a random sample from the negative class having the same number of values the positive class has. By doing this, we balance our dataset. [Figure 3] below is the balanced dataset. Once the dataset is preprocessed, and feature engineering is applied, we start performing experimental analysis on our data.

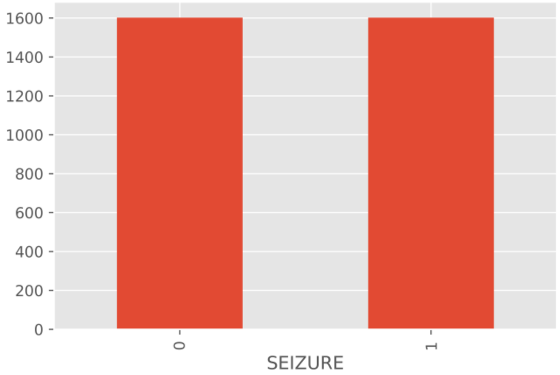


Figure 3. Proportion of balanced data

5. Proposed methodology

Since the data has been cleaned and organized, the data is ready to be deployed in the proposed model. The baseline model has been designed in a way such that it would provide meaningful analysis after performing mathematical computations on data employing machine learning classifiers and predict if a person is having a seizure or not. Since major part of machine learning is inclusive of classification and tells what class an observation belongs to. [Figure 4] shows the overview of the proposed methodology incorporated in our model using python libraries and classifiers.

The capacity to precisely distinguish observations is precious for numerous applications, which involves prediction, and it can be aligned with medical streams to extract the best results with the help of machine learning. Data science presents an overabundance of classification algorithms such as logistic regression, support vector machine, random forest, gradient boosting, and decision trees. But near the top of the classifier hierarchy is the extra trees classifier. Five basic classifiers have been employed for the calculation of predictive scores our model, which is explained below.

5.1. Random forest classifier

Random forest, alike its name refers, consists of a vast amount of unique decision trees that perform as an ensemble. Each tree in the random forest derives out a class prediction, and the class with the total votes enhances our model's prediction. A considerable amount of moderately uncorrelated models (trees) functioning as a committee will beat any of the individual constituent models. The coarse correlation among models is the core key. The training algorithm for random forests practices the conventional method of bootstrap aggregating, or bagging, to tree learners in which bagging repeatedly picks a random sample with replacement of the training set and furnishes trees to these samples.

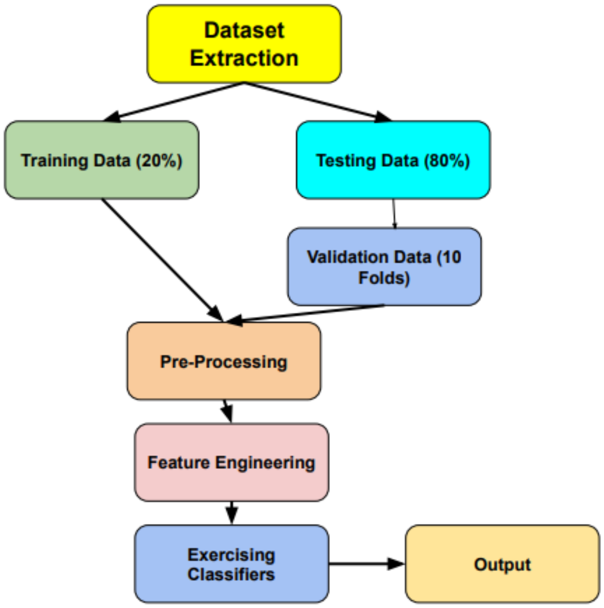


Figure 4. Structural architecture of proposed methodology

5.2. k-Nearest Neighbor (KNN) Classifier

The KNN algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be employed to determine both classification and regression problems. The KNN algorithm believes that related things endure in close proximity. Alternatively, it can be said that the same elements exist next to each other. The number of neighbors is the core deciding factor. K is generally an odd number if the number of classes is 2, which in our project has been initialized to 100. For finding closest neighbors, it obtains the distance among points using working distance measures such as Euclidean distance, Hamming distance, Manhattan distance, and Minkowski distance. After calculating the distance, the last step is to vote for labels, which in our research model are binary labels.

5.3. Gradient boosting classifier

Another classifier that is used majorly for classification problems is gradient boosting, which again works on the principle of decision trees and recommends an adjustment to gradient boosting method, which develops the quality of fit of each base learner. Boosting in gradient boosting refers to the technique of transforming weak learners to active learners where each new tree is a fit on a modified version of the original data set. Gradient Boosting trains many models in a gradual, additive, and sequential manner. After assessing the first tree, we improve the weights of those observations that are tough to classify and lower the weights for those that are easy to classify. The second tree is therefore grown on this weighted data. Here, the idea is to improve upon the predictions of the first tree.

5.4. Support Vector Machine (SVM)

As SVM classifiers are considered suitable for binary classification, and the dataset size is nearly equal to ten thousand, in this study, we are using the support vector classification (SVC) method of the SVM algorithm for building the classifier. For the classifier, the kernel is set to linear, and the regularization of 1.0 is applied. The SVM classifier performance is almost equal to the random forest classifier with a reasonable precision score. After applying feature engineering techniques, SVM classifies the coordinates by building an imaginary hyperplane and tries to maximize the margin of that hyperplane to build a clear boundary for binary classification and it also helps up in introducing non-linearity to the existing data and helps in avoiding the issue of overfitting.

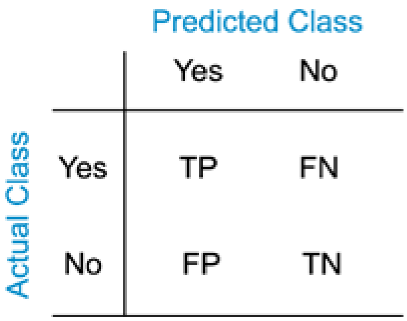
5.5. Extra trees classifier

Extra Trees classifier, also known as Extremely Randomized Trees classifier, is a type of ensemble learning technique that aggregates the results of various de-correlated decision trees solicited in a “forest” to output its classification result. In theory, it is quite comparable to a Random Forest Classifier and only varies from it in the manner of construction of the decision trees in the forest. Summarizing it, Extra trees classifier builds multiple trees with bootstrap equating it to false, which means it samples without replacement, and the other thing which plays a crucial role in it is nodes are broken based on random splits among a random subset of the features selected at every node. In Extra Trees, randomness doesn’t appear from bootstrapping of data; however, it instead develops from the random splits of all observations.

All the above-mentioned classifiers are applied in our proposed model using python libraries such as sklearn also known as Scikit-learn which is inclusive of various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, *k*-means and DBSCAN, and is intended to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Following performance parameters have been used to evaluate the performance of the proposed model which have been extracted using confusion matrix.

5.6. Confusion matrix

A confusion matrix is a type of specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one. Each column represents the instances of the actual class and each row of the matrix represents the instances of the predicted class (or vice versa). It is known as a special kind of contingency table having two dimensions actual and predicted. [Figure 5] represents Confusion Matrix.



Actual Class \ Predicted Class	Yes	No
Yes	TP	FN
No	FP	TN

Figure 5. Confusion matrix

The performance metrics which we have used for classification are as follows:

Specificity is the proportion of patients without epileptic seizure who test negative. The equation is represented below:

$$Specificity = \frac{TN}{(TN + FP)}$$

6. Experimental analysis

Following the application of various classifiers, the proposed methodology was evaluated using performance metrics. [Table 1] shows the test result comparisons of all performance parameters which have been deduced using 5 classifiers.

Table 1. Performance metrics of classifiers

Performance Parameters	KNN	Random Forest	Gradient Boosting	SVM	Extra Trees
Accuracy	84.6	96.5	95.4	96.8	96.8
Recall	23.9	91.6	93.4	94.5	94.5
Precision	97.6	91.1	85.0	89.9	89.9
Specificity	99.9	97.8	95.9	97.3	97.3

While performing comparative analysis, it was observed that Extra trees and SVM performed relatively better when compared with other classifiers as SVM and Extra trees achieved the accuracy of 96.8%. Furthermore, with the help of data visualization techniques we exercised to display the train, valid and test score of all classifiers in a bar graph for a better comparison which is displayed in [Figure 6] and [Figure 7] represents the Feature Importance score of different electrodes.

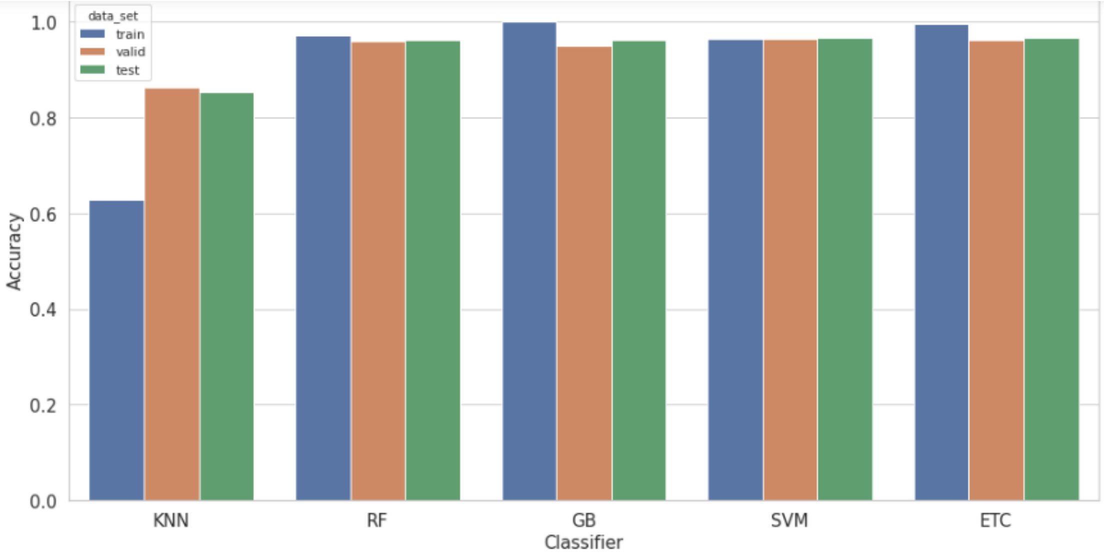


Figure 6. Accuracy of different classifiers

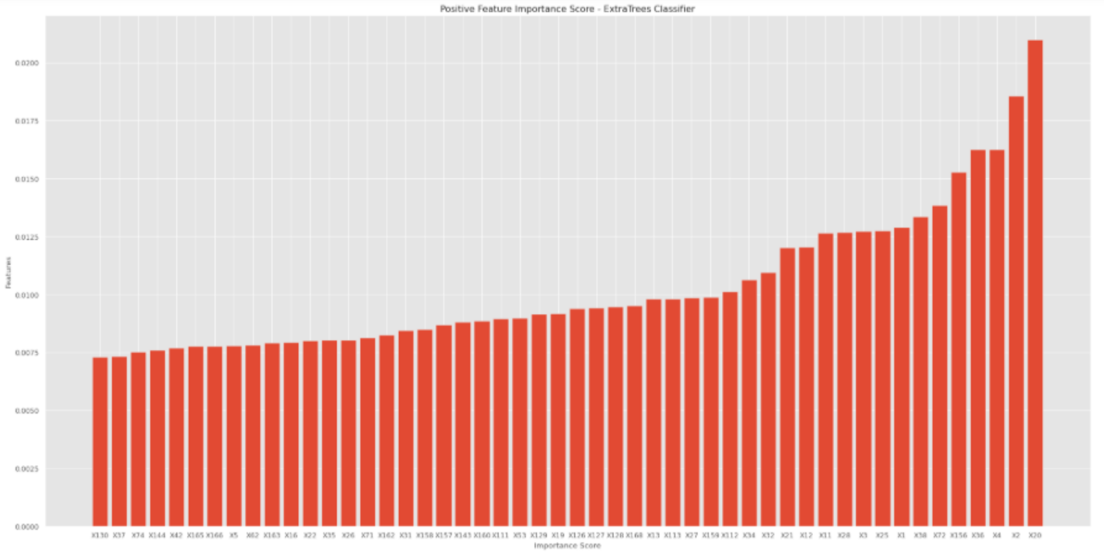


Figure 7. Positive features importance

For making it easier to understand which one of the electrodes have the highest impact on predicting a seizure, we have generated a bar graph using seaborn and matplotlib python's library.

7. Conclusions

In this research paper, we have implemented several classifiers that possess supervised learning capabilities on the brain signal seizure analysis dataset, which is provided by Rochester Institute of Technology, USA. The introduction and literature background describe previous and existing work with the same technology. The data pre-processing methods consist of data splitting, data prevalence calculation for data balancing, subsampling, and using python libraries for data cleaning. The classifiers applied on the dataset were SVM, Random Forest, Extra Trees, KNN, and Gradient Boosting. The results deduced from the comparative study of the classifiers conclude that Extra Trees classifier outclassed all the classifiers by achieving 96.8 % accuracy on the dataset provided to the algorithm. The analysis has been performed to display the impact of brain electrodes on a receptive area of the brain. The feature importance score graph discussed in the experimental analysis section states that the brain electrode 'X28' has the highest impact on the receptive area of the brain. The current research can be extended and expanded in the future by integrating it with real-time dynamic applications which can detect brain seizures when it's in the vulnerable stage. The proposed work can be improved by implementing the application in the fields of healthcare and medical diagnosis.

Acknowledgements

This research is part of the second and third authors through their research at the Research Methodology Graduate Course (COMP5112) supervised by Dr. Jinan Fiaidhi. The authors would like also to Rochester Institute of Technology for sharing their open dataset for domestic experimentation

References

[1] S.B. Wilson and R. Emerson, "Spike detection: A review and comparison of algorithms," Clin. Neurophysiol., vol.113, pp.1873-1881, **(2002)**

[2] J. Gotman and P. gloor, "Automatic recognition and quantification of interictal epileptic activity in the human scalp EEG".

[3] S.B. Wilson, C.A Turner, R.G. Emerson, and M.L. Scheuer, "Spike detection II: Automatic, perception-based detection and clustering," Clin. Neurophysiol., vol 110, pp 404-411, **(1999)**

[4] W.R. Webber, B. Litt, K. Wilson, and R. P. Lesser, "Practical detection of epoleptiform discharges (EDs) in the EEG using an artificial neural network: A comparison of raw and parameterized EEG data," Electroencephalogr, Clin. Neurophysiol., vol.91, pp.194–204, **(1994)**

[5] K. Kobayashi, I. Merlet, and J. Gotman, "Separation of spikes from Background by independent component analysis with dipole modeling and comparison to intracranial recording," Clin. Neurophysiol., vol.112, pp.405–413, **(2001)**

[6] N. Acir and C. Guzelis, "Automatic spike detection in EEG by a two stages procedure based on support vector machines," Comput. Biol. Med., vol.34, pp.561–575, **(2004)**

[7] R. Vaitheeshwari and V. SathieshKumar, "Performance analysis of epileptic seizure detection system using neural network approach," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, pp.1-5, **(2019)**

[8] R. Shantha Selva Kumari and J. Prabin Jose, "Seizure detection in EEG using time frequency analysis and SVM," 2011 International Conference on Emerging Trends in Electrical and Computer Technology, Nagercoil, pp.626-630, **(2011)**

[9] A. T. Tzallas, M. G. Tsipouras, and D. I. Fotiadis, "Epileptic seizure detection in EEGs using time–frequency analysis," in IEEE Transactions on Information Technology in Biomedicine, vol.13, no.5, pp.703-710, Sept. **(2009)**

[10] D. A. Torse, R. Khanai, and V. V. Desai, "Classification of epileptic seizures using recurrence plots and machine learning techniques," 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, pp.0611-0615, **(2019)**

[11] "Epileptic Seizure Recognition," uci.edu, https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition, **(2020)**

Authors

Dr. Jinan Fiaidhi
Dr. Fiaidhi is a full professor of Computer Science and Professional Software Engineer as well as the Grad Coordinator of the PhD program in Biotechnology at Lakehead University. She is an adjunct research professor at the Western University and the editor in chief of IGI Global International Journal of Extreme Automation and Connectivity in Healthcare. She is also the chair of Big Data for eHealth with the IEEE ComSoc. Contact him at sabah.mohammed@lakeheadu.ca.

Tejas Wadiwala
He is a grad student at Lakehead University, Computer Science working under the supervision of Dr. Fiaidhi.

Vikas Trikha
He is a grad student at Lakehead University, Computer Science working under the supervision of Dr. Fiaidhi.

This page is empty by intention.