Epileptic seizure classification of electroencephalogram signals using extreme gradient boosting classifier

Millee Panigrahi[1], Dayal Kumar Behera[2], Krishna Chandra Patra[3]
[1]Department of Electronics, Sambalpur University Institute of Information Technology, Bhubaneswar, India
[2]Department of Computer Science and Engineering, Silicon Institute of Technology, Bhubaneswar, India
[3]Department of Electronics, Sambalpur University, Sambalpur, India

Article Info

ABSTRACT

Epilepsy causes repeated seizures in an individual's life, which causes transient irregularities in the brain's electrical activity. It results in different physical symptoms that are abnormal. Various antiepileptic drugs fail to minimize repeated patient seizures. The electroencephalogram (EEG) signal recordings provide us with time-series data set for epileptic seizure detection and analysis. These signals are highly nonlinear and inconsistent, and they are recorded over time. Predicting the ictal period (seizure period at the time of epilepsy) is thus a challenging task in the naked eye for the medical practitioners. Various machine learning techniques are applied to identify the seizure's occurrence and its classification in multiple domains. A classification model based on extreme gradient boosting (SCLXGB) is proposed here for the classification of the EEG signals. The SCLXGB model implements binary seizure classification on the benchmark dataset. Compared with K-nearest neighbor, linear regression, and Decision tree-based models, the proposed model achieves the best area under receiver operating curve (AUC) of 0.9462 and an accuracy of 96% which signifies accurate prediction of seizure and non seizure period. The proposed model SCLXGB was validated by taking different performance metrics to indicate the occurrence and non-occurrence of seizures in patients more appropriately.

Corresponding Author:

Dayal Kumar Behera
Department of Computer Science and Engineering, Silicon Institute of Technology
Bhubaneswar, Odisha, India
Email: dayalbehera@gmail.com

1. INTRODUCTION

Various recording systems are available nowadays to record the human brain signal in multiple forms. These signalling systems have many advantages and disadvantages [1]. For example, an electroencephalogram (EEG) is a signalling method that is more prone to processing steps such as enhancement, extraction of features, and classification for epileptic seizure prediction. Therefore, the EEG signals have higher temporal resolution and safety compared to other methods.

Electroencephalogram (EEG) recording is done through the standardized electrodes placed non-invasively on the brain's scalp. The clinical inspection for seizure occurrence is visualized by the cerebrum's electrical action, which contains much data about brain activity [2]. In this way, EEG signals have an incredible significance in brain diagnosis in establishing pre-estimated epileptic seizure detection [3]. The EEG varies from average to spike, sharp, spike-and-moderate complex wave, complex pointed-and-medium wave, and another epileptic form of signals [4]. Thus, EEG recordings help the researchers with

greater efficiency and less time consumption for automated epilepsy diagnosis than visual inspection of seizures.

Long EEG recordings can recognize the seizure period rather than a short duration of time. Therefore, the long-term EEG signals can significantly categorize the interictal and ictal periods. The EEG signal is a long-term recording of the brain signal, segregated into five different frequency bands to properly analyze seizure and non-seizure patterns [5]. This non-uniform and non-stationary signal is time-dependent and characterized by the repetitive high amplitude and combination of spike and slow waves [6].

The EEG signals are highly non-stationary. These signals are analyzed with frequency, time, and time-frequency decomposition methods involving discrete wavelet transform, particularly for feature extraction [7]. The visual finding of seizures is time-consuming and tedious to identify the exact duration of the physicians' ictal episodes. Therefore, the researchers implemented different automated seizure prediction algorithms to enhance the precise detection of epileptic seizures from electroencephalogram recordings. Li *et al.* [8] proposed other soft computing techniques like genetic algorithm (GA), fuzzy logic approaches are applied to classify the epileptic and non-epileptic EEG segments. Due to the long EEG recordings, the size reduction with optimum information about the ictal activity has been challenging for the researchers. This problem statement resulted in selecting and extracting features using complex dual-tree transformation with variations in granularity level [9]. The recorded signals are featured using the wavelet decomposition method and classified using Naïve Bayes and K-nearest neighbor (KNN) classifiers to generate fourteen combinations of 2-class epilepsy [10]. During the ictal and preictal states, the amount of information inflow and outflow between the brain's cortical regions is considered a parameter for seizure classification [11]. The research implemented feature extraction and feature selection methods through correlation and random forest (RF) for seizure and non-seizure classification [12], [13]. The EEG signals are also helpful for detecting other brain disorders like autism, as illustrated in [14], [15].

A machine learning (ML) approach for predicting the onset of seizures is proposed in this present work. The extreme gradient boosting algorithm (SCLXGB) is implemented to estimate the descriptive samples from a trained predictive model. SCLXGB is a seizure classification method using the XGBoost model. This method can considerably approach the parallel and distributed calculation to verify the model's attributes. XGBoost library [16] is a productive and disseminated application of gradient boosting [17]. It provides a faster interpretation of the model with a decreased overfitting than other boosting algorithms [18]. The significant contribution of the paper is given below:

- Design of an efficient classification model for epileptic seizure identification from EEG signal.
- Extreme gradient boosting is a variant of a gradient boosting model with efficient computational flexibility, can eliminate the probability of missing the values.
- The existing extreme gradient boosting model is modified to estimate the early stopping of epochs for training data by computing the root mean square error (RMSE) values for each epoch from the log loss graph. Accordingly, the parameters are optimized.
- The proposed model is compared with the existing classifiers and found to be more accurate withless computation time for predicting the seizures.

2. MATERIALS AND METHOD

The epileptic seizure data set is taken from the University of California at Irvine (UCI) machine learning repository [19] which consists of five states of patients data sets, each with 100 files. Each of the 100 files represents a single subject's brain activity record for 23.6 seconds. The data is a time-series signal and is sampled in 4097 data points. Each signal is then processed by dividing and shuffling each set of 4097 data points into 23 pieces [19]. This time series dataset has 178(4097/23) data points for 1 second in every 23 pieces. The data point represents the EEG recording value on separate time instances. Thus a total of 23x500=11500 pieces of information containing 178-dimensional input vectors represented for 1 second by X1, X2, X3 ...... X178. There are five patient statuses represented by y{1, 2, 3, 4, 5} as indicated [19]:

- EEG Recording of instances of seizures (ES).
- EEG signal from brain tumour site (ETB).
- EEG Record from a healthy area of the brain (EHB).
- EEG signal of the healthy subject with eyes closed (EYEC).
- EEG signal of the healthy subject with eyes open (EYEO).

The subjects in categories 2, 3, 4, 5 have no epilepsy, and the subject in category 1 has epilepsy. The data processing is done such that it is easier for binary classification against the rest of the classes. Figure 1 represents the five states of the subjects.
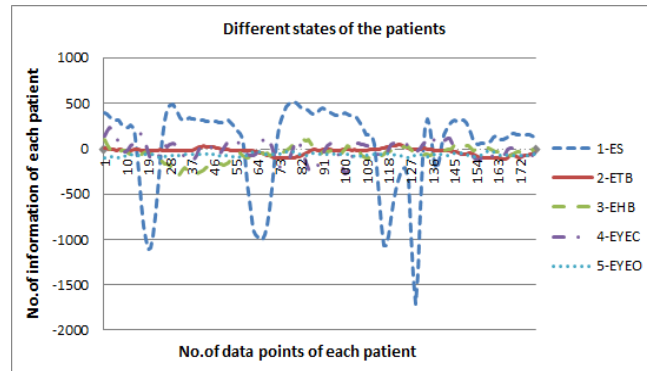
Figure 1. An illustration of EEG recordings

3. METHOD

Figure 2 represents the flow diagram of the proposed method. Firstly the benchmark dataset is considered for classification and then data preprocessing is done using filtering technique. Then training and testing dataset is considered performing train test split with 70% train data and 30% test data. After that the classification is done using SCLXGB classifier for seizure prediction. The detailed explanation is illustrated in Figure 2.
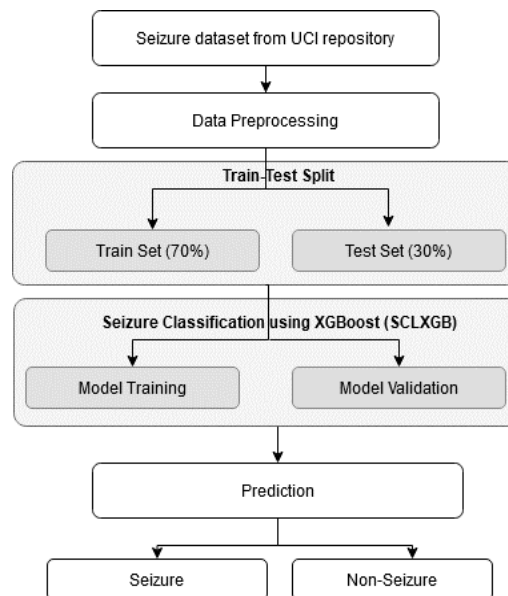

Figure 2. The flow diagram of the proposed model SCLXGB

3.1. Data preprocessing

The dataset used in our study for epileptic seizure detection is publicly available and easily downloadable [19]. The EEG recordings are multivariate time series signals sampled at 173.61 Hz having a spectral bandwidth of 0.5 Hz to 85 Hz. The preprocessing is done by applying a low pass filter at a frequency of 40 Hz; the data is downsampling is carried out to tune with the classifiers. Figure 1 represents the different dataset levels represented in terms of voltage levels concerning time. The dataset having the seizure class is marked as 1, and other datasets are marked as 2, 3, 4, 5, respectively, representing different states of the subjects. We only concentrate on the dataset having a seizure and the time duration of the occurrence of the seizure.

3.2. Extreme gradient boosting (SCLXGB)

The XGBoost algorithm [16] incorporates a more regularized model to manage the overfitting of the attributes by implementing a parallel processing technique with a varying number of terminal nodes. They

comprise an array of split points and the nodes represented as input variables. Here the last node is the leaf of the trees that gives the specific values of output variables. The leaf weights having less information about the occurrence of seizures are shrieked to reduce model complexity. XGBoost algorithm has two significant upgradations than other boosting algorithms as:

− It is robust and speeds up the construction of the tree.
− Proposing a new distributed algorithm for tree searching and extra randomization parameter is applied to decrease the correlation between trees.
− XGB provides a more regularized structure to manage overfit and allows parallel processing better than the gradient boosting model.
− The novel contribution of this paper is that the parameters are optimized to obtain higher accuracy with reduced complexity of the model.

The extreme gradient boosting (XGBoost) algorithm [16], [20] was implemented using the Scikit-learn [21] Python modules.

3.3. Model description

The EEG signal classification has become an arduous task for medical practitioners to detect an epileptic patient's seizure and non-seizure activity [22]. The main idea behind this model is to carry out a more automated binary classification of the EEG signal using a state-of-the-art XGBoost learning model (SCLXGB) that outperforms other seizure class-determination models. Extreme gradient boosting (XGBoost) [23] is a breakthrough among ensemble learning models that incorporate separate variables without overfitting directly and can handle non-linear signals such as EEG. After data preprocessing, the overall steps carried out are as follows.

− The training dataset is 70%, and the testing dataset is 30%. Out of 11500 data points, 8050 data points are considered for the training data set, and 3450 data are taken for testing.
− Test dataset is used to validate the SCLXGB model.
− Validated model is used for prediction

Algorithm 1: proposed model SCLXGB
```
Input: UCI machine learning repository
Output: performance metrics and area under the curve (AUC)
1.   Data preprocessing and generating the feature set.
2.   Normalizing the data using StandardScaler.
3.   Modeling as a classification problem.
4.   Train-test split.
5.   Train and tuning the XGBoost Model using all the features.
6.   Calculating the classification metrics.
7.   Plot the train and test confusion matrix and ROC.
```

Algorithm 1 represents the steps followed in the proposed model. The train and test error of the SCLXGB model is trained with different epochs and the corresponding RMSE value is represented in Figure 3. The stochastic nature of the algorithm is estimated by plotting the RMSE values with 70% train data and 30% test data sets for each epoch. It is observed that after round 40 epochs the iterations should be stopped to avoid overfitting of the training model.
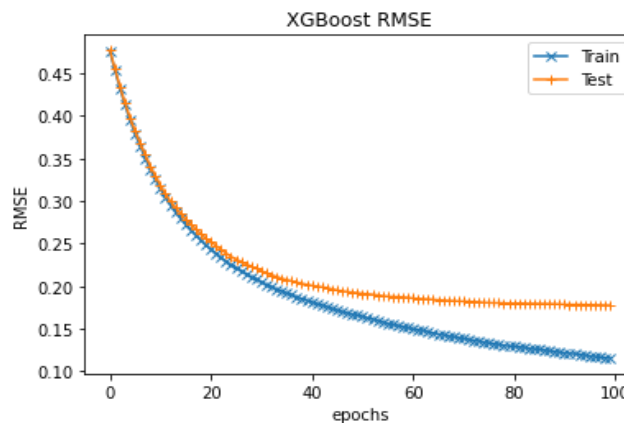

Figure 3. The learning curve (logloss) of the proposed model SCLXGB

4. RESULTS AND DISCUSSION

The UCI machine learning repository dataset [19] for epileptic seizure detection is considered for testing the proposed model. The results obtained are compared with the state of the art models like KNN, logistic regression and decision tree and Gaussian Naïve Bayes classifiers [24]-[26]. Gradient boosted decision trees (GBDT) is an ensemble learning methodology that combines many decision trees in series. XGBoost was initially introduced by Chen and Guestrin [16] and was then further employed by researchers due to its efficacy in decreasing processing time and efficiently utilising memory resources. Table 1 shows the parameters that are considered for the classification of the models.

Table 1. Model parameters

Classifiers	Parameters
KNN	Neighbors(k)=5, weights=uniform, leaf_size=30, p=2, metric=minkowski
Decision tree	criterion='gini', splitter='best', min_samples_split=2, min_samples_leaf=1
SCLXGB	base_score=0.5, gamma=0, learning_rate=0.1, max_delta_step=0, max_depth=3, n_estimators=100, objective='binary: logistic', reg_alpha=0, reg_lambda=1

Table 1 elaborates the parameters set for optimizing the performance of the classifier. Our research focused on the model complexity and adjusted the following three proposed model parameters SCLXGB. The model works efficiently with less computation time and increased accuracy. The "max_depth" is a parameter that emphasizes the model complexity. The higher the value, the more complex the model becomes [27]. The "min_child_weight" always takes positive integer values. It further evaluates the splitting of nodes if the sum of weights is greater than the peak value. This parameter makes the algorithm more customary as the value increases. The third parameter is the "learning_rate", which is tuned to prevent the system from being less overfitting and more robust. It shrinks the step size so that it provides more area for further enhancement. The classifier's performance is evaluated by the metrics such as accuracy, precision, recall, f1-measure and AUC are calculated. The performance metrics of various models are illustrated in Table 2. A higher value indicates a better result.

From the above table it shows that KNN exhibits better precision of 98.3% than any other model for this benchmark dataset. It can be seen that Logistic regression and decision tree model does not have a remarkable performance in this context. LR has an accuracy of 73.6% which is very low and DT has an accuracy of about 89.9 which is lower as compared to the proposed model. The proposed model SCLXGB outperforms other classifiers in recall, F1 measure, accuracy and AUC. It has achieved an accuracy of 96% with less computation time. The performance of the system can be evaluated in terms of computation time comparing with KNN and DT [28].

From the graph shown in Figure 4 the computational time or the elapsed training time for the proposed model SCLXGB is 2.66 ms and that of KNN is 7.87 ms and DT is 3.15 ms. The confusion, precision and recall matrix of K nearest neighbor shown in Figure 5(a), linear regression in Figure 5(b), decision tree in Figure 5(c), Naïve Bayes and proposed model SCLXGB classifier in Figure 5(d).

Table 2. Performance metrics of test data

Model	Accuracy	Precision	Recall	F1 Score	AUC
KNN	0.931	0.983	0.678	0.803	0.8378
LR	0.736	0.377	0.438	0.406	0.6259
DT	0.898	0.708	0.860	0.777	0.8844
GNB	0.958	0.907	0.894	0.899	0.9349
SCLXGB (Proposed)	0.960	0.892	0.916	0.904	0.9462

Figure 4. Computational time of the classifiers

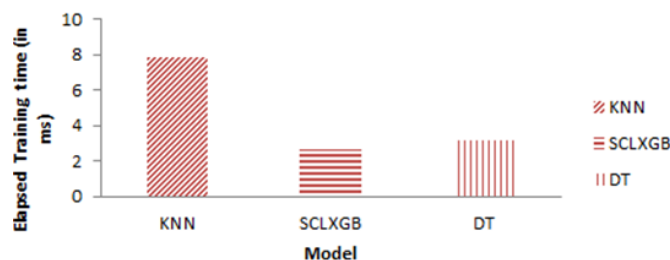(a)   (b)

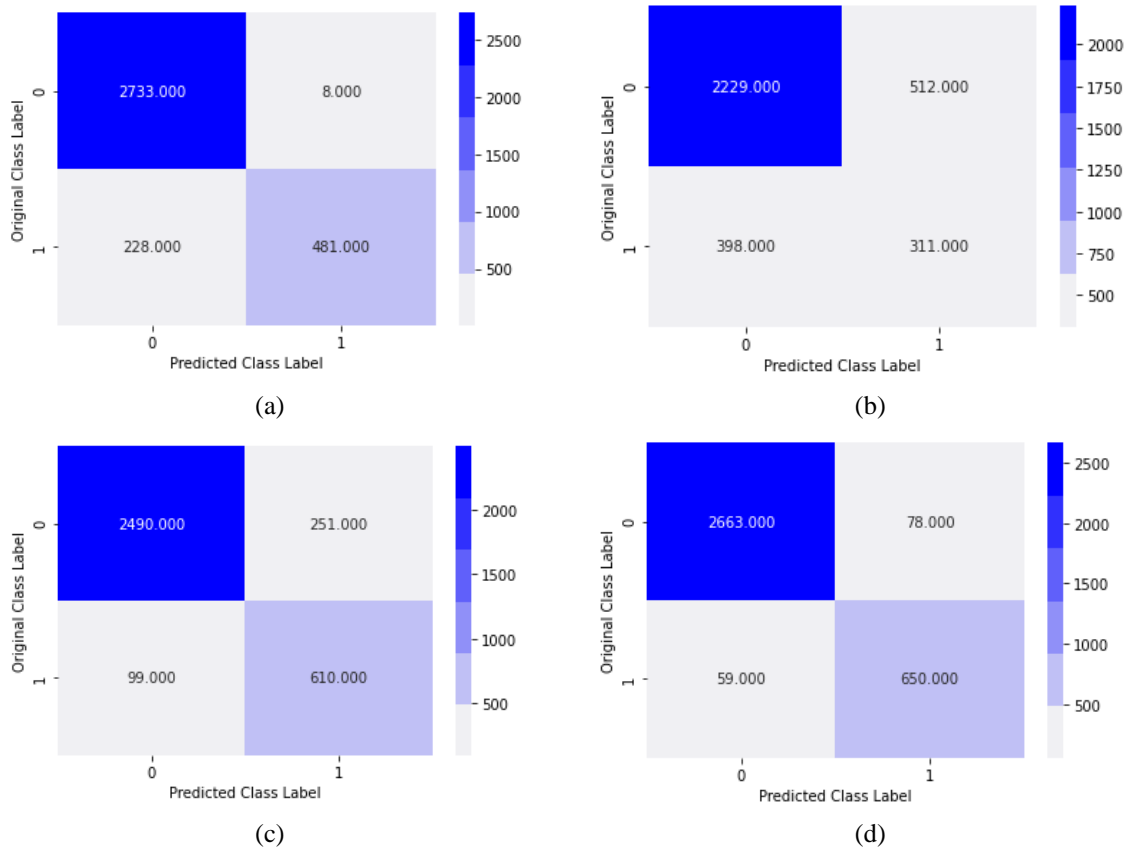(c)   (d)

Figure 5. Confusion matrix: (a) KNN, (b) LR, (c) DT, and (d) SCLXGB

The AUC of the various classification models is depicted in Figure 6. Higher AUC value reflects a better model. It is seen that the proposed model exhibits better performance as compared to other models.
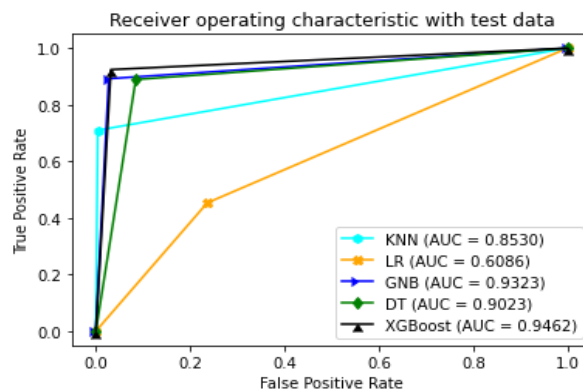

Figure 6. AUC of classification models on test data

5.   CONCLUSION

In this present study, different ML classifiers are compared to obtain the ictal period of seizures in epileptic patients. In the form of the above investigation, it is found that the binary classification of seizures outperforms other models in terms of regularization and minimization of computational time. Seizure classification using the SCLXGB model is a novel approach and provides the best AUC of 0.9462 and test accuracy of 96% compared to other machine learning models such as KNN, linear regression, decision tree

and Naïve Bayes models. After pre-processing, the data train test splitting is carried out, and 70% of the train and 30% of the test data are considered for validation of the model. The area under the ROC curve was 0.9462 for the proposed model. Compared to the KNN, LR decision tree and Naïve Bayes models, SCLXGB is the most acceptable classification model. The proposed model provides for a suitable regularization that prevents it from being over-fitted. SCLXGB integrates data sparsity with a split-finding algorithm to manage different types of data sparsity patterns. The data is organized in memory cells called blocks to be reused in repeated iterations instead of re-calculated. This helps to reduce the computation time of 2.66 ms and makes the model more robust for classification. In SCLXGB, discontinuous memory access is required to obtain gradient information by row indicator for optimum hardware utilization. This is done by assigning internal buffers to each thread where gradient statistics can be stored. This feature enables optimized use of available disc space with a large dataset handling capability.

REFERENCES

[1] G. Brihadiswaran, D. Haputhanthri, S. Gunathilaka, D. Meedeniya, and S. Jayarathna, "A Review of EEG-based Classification for Autism Spectrum Disorder," *Journal of Computer Science (JCS)*, vol. 15, no. 8, pp. 1161-1183, 2019, doi: 10.3844/jcssp.2019.1161.1183.

[2] M. Tanveer and A. A. Salman, "Epileptic seizure classification using gradient tree boosting classifier," In *Proceedings of the 2019 9th International Conference on Biomedical Engineering and Technology*, 2019, pp. 71-77, doi: 10.1145/3326172.3326182.

[3] A. Subasi, "Epileptic seizure detection using dynamic wavelet network," *Expert Systems with Applications*, vol. 29, no. 2, pp. 343-355, 2005, doi: 10.1016/j.eswa.2005.04.007.

[4] F. L. Da Silva, W. Blanes, S. N. Kalitzin, J. Parra, P. Suffczynski, and D. N. Velis, "Epilepsies as dynamical diseases of brain systems: basic models of the transition between normal and epileptic activity," *Epilepsia*, vol. 44, pp. 72-83, 2003.

[5] M. D. Alessandro, R. Esteller, G. Vachtsevanos, A. Hinson, J. Echauz, and B. Litt, "Epileptic seizure prediction using hybrid feature selection over multiple intracranial EEG electrode contacts: a report of four patients," *IEEE transactions on biomedical engineering*, vol. 50, no. 5, pp. 603-615, 2003, doi: 10.1109/TBME.2003.810706.

[6] C. Maswanganyi, C. Tu, P. Owolawi, and S. Du, "Factors influencing low intension detection rate in a non-invasive EEG-based brain computer interface system," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 20, pp. 167-175, 2020, doi: 10.11591/ijeecs.v20.i1.pp167-175.

[7] T. Y. Wen and S. A. M. Aris, "Electroencephalogram (EEG) stress analysis on alpha/beta ratio and theta/beta ratio," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 17, no. 175, pp. 10-11591, 2020, doi: 0.11591/ijeecs.v17.i1.pp175-182.

[8] F. Li *et al.*, "Brain variability in dynamic resting-state networks identified by fuzzy entropy: a scalp EEG study," *Journal of Neural Engineering*, vol. 18, no. 4, 2021, doi: 10.1088/1741-2552/ac0d41.

[9] M. Chakraborty and D. Mitra, "Automated detection of epileptic seizures using multiscale and refined composite multiscale dispersion entropy," *Chaos, Solitons & Fractals*, vol. 146, 2021, doi: 10.1016/j.chaos.2021.110939.

[10] D. K. Behera, M. Das, S. Swetanisha, and P. K. Sethy, "Hybrid model for movie recommendation system using content K-nearest neighbors and restricted boltzmann machine," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 23, no. 1, 2021, doi: 10.11591/ijeecs.v23.i1.pp445-452.

[11] G. Wang, Z. Sun, R. Tao, K. Li, G. Bao, and X. Yan, "Epileptic seizure detection based on partial directed coherence analysis," *IEEE journal of biomedical and health informatics*, vol. 20, no. 3, pp. 873-879, 2015, doi: 10.1109/JBHI.2015.2424074.

[12] M. Mursalin, Y. Zhang, Y. Chen, and N. V. Chawla, "Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier," *Neurocomputing*, vol. 241, 2017, doi: 10.1016/j.neucom.2017.02.053.

[13] S. Raghavendra and J. S. Kumar, "Performance evaluation of random forest with feature selection methods in prediction of diabetes," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 353, 2020, doi: 10.11591/ijece.v10i1.pp353-359.

[14] D. Haputhanthri *et al.,* "An EEG based Channel Optimized Classification Approach for Autism Spectrum Disorder," In *2019 Moratuwa Engineering Research Conference (MERCon)*, 2019, pp. 123-128, doi: 10.1109/MERCon.2019.8818814.

[15] D. Haputhanthri *et al.,* "Integration of Facial Thermography in EEG-based Classification of ASD," *International Journal of Automation and Computing*, vol. 17, no. 6, pp. 837-854, 2020, doi: 10.1007/s11633-020-1231-6.

[16] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[17] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367-378, 2002, doi: 10.1016/S0167-9473(01)00065-2.

[18] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *The Journal of Machine Learning Research*, vol. 11, pp. 2079-2107, 2010, doi: 10.5555/1756006.1859921.

[19] Q. Wu and E. Fokoue, Epileptic Seizure Recognition, UCI Machine Learning Repositor. Accessed: Jun 15, 2021. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition

[20] X. Wang, L. Wang, S. Wang, J. Chen, and C. Wu, "An XGBoost-enhanced fast constructive algorithm for food delivery route planning problem," *Computers & Industrial Engineering*, vol. 152, 2021, doi: 10.1016/j.cie.2020.107029.

[21] S. Raschka, Python machine learning. Packt publishing ltd, 2015. [Online]. Avaliable: https://raw.githubusercontent.com/rasbt/python-machine-learning-book/master/docs/equations/pymle-equations.pdf

[22] M. Panigrahi, D. K. Behera, and K. C. Patra, "Classification of EEG Recording in an Epileptic Patient Using DWT and Extreme Gradient Boosting Classifier," *International Journal of Advanced Science and Technology*, vol. 29, no. 5, pp. 4163-4180, 2020.

[23] D. K. Behera, M. Das, S. Swetanisha, J. Nayak, S. Vimal, and B. Naik, "Follower Link Prediction Using the XGBoost Classification Model with Multiple Graph Features," *Wireless Personal Communications*, pp. 1-20, Apr. 7, 2021, doi: 10.1007/s11277-021-08399-y.

[24] M. Panigrahi, D. K. Behera, and K. C. Patra, "A Hybrid Model for Epileptic Seizure Classification Using Wavelet Packet Decomposition and SVM," *Advances in Intelligent Computing and Communication*, pp. 211-219, 2021, doi: 10.1007/978-981-16-0695-3_21.

[25]   R. G. Andrzejak, K. Lehnertz, C. Rieke, F. Mormann, P. David, and 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," *Physical Review E*, vol. 64, no. 6, 2001, doi: 10.1103/PhysRevE.64.061907.
[26]   S. Kuila, N. Dhanda, and S. Joardar, "Feature extraction of electrocardiogram signal using machine learning classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 6, pp. 6598-6605, 2020, doi: 10.11591/ijece.v10i6.pp6598-6605.
[27]   A. Haider and M. Z. Al-Faiz, "Online multiclass EEG feature extraction and recognition using modified convolutional neural network method," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 5, pp. 4016-4026, 2021, doi: 10.11591/ijece.v11i5.pp4016-4026.
[28]   H. R. Esmaeel, "Analysis of classification learning algorithms," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 17, no. 2, pp. 1029-1039, 2020, doi: 10.11591/ijeecs.v17.i2.pp1029-1039.

BIOGRAPHIES OF AUTHORS

Millee Panigrahi ⒾⒼ ⓈⒸ Ⓟ has completed her bachelor in electronics and telecommunication enggineering and master in signal processing and telematics from BPUT, Odisha, India. Presently she is working as assistant professor in dept. of electronics and telecomm engg. at Trident Academy of Technology, Bhubaneswar, Odisha, India. She can be contacted at email: millee.panigrahi82@gmail.com.

Dr. Dayal Kumar Behera ⒾⓇ ⓈⒸ Ⓟ has completed Ph. D. from KIIT University, Odisha. He has been working as an asst. professor in the department of CSE at Silicon Institute of Technology, Bhubaneswar. He is having fourteen years of teaching experience and carrying out research in various aspects of recommender system, machine learning, data science, IoT and natural language processing. He is having 20 publications in various journals and conferences and guided numerous B. Tech. and M. Tech. projects in his area of interests. He is a life time member of professional societies like ISTE and IAENG. He can be contacted at email: dayalbehera@gmail.com.

Dr. Krishna Chandra Patra ⒾⓇ ⓈⒸ Ⓟ has completed his Ph. D. in electronics from Delhi University. Presently he is working in department of electronics in SUIIT, Sambalpur University, Odisha, India. He has published more than 16 papers in journal and conference. More than 5 candidates have completed their M. Phil. under his supervision in the area of electronics communication system and related disciplines. His area of research is advanced communication techniques and on photonics. He can be contacted at email: krishnach.patra@gmail.com.