Improved Epilepsy Identification with XGBoost vs. Logistic Regression

Gadamsetty Hemanth* and N. Navaprakash†
Department of Electronics and Communication Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, 602105, India

Keywords: Novel XGBoost Algorithm, Logistic Regression Algorithm, Epilepsy, Detection, Analysis, Identification, Disease.

Abstract: This study aims to analyse and detect epilepsy with enhanced accuracy by implementing the XGBoost algorithm, comparing its performance against the Logistic Regression algorithm. Methodology: The research involves preprocessing and analysing a dataset containing instances of epileptic seizures. The dataset, sourced from CHB-MIT, is divided into two subsets, each consisting of 15 samples. The first subset applies the XGBoost algorithm, while the second employs the Logistic Regression algorithm for epilepsy disease detection. G power (80%) and alpha (0.005) values are determined for the study. Findings: The XGBoost algorithm achieves an impressive accuracy rate of 89%, surpassing the accuracy of the Logistic Regression algorithm at 74%. The observed difference is statistically significant, as confirmed by an independent sample t-test resulting in a p-value of 0.000 ($p < 0.05$). Conclusion: The study concludes that the innovative XGBoost algorithm excels with an accuracy of 89% compared to the Logistic Regression algorithm, establishing its effectiveness in epilepsy disease analysis and detection.

1 INTRODUCTION

Epilepsy, a neurological disorder characterized by recurrent seizures, affects individuals of all ages. It arises from neurobiological processes leading to epileptogenesis within the brain, where abnormal neuronal firing occurs in the cerebral cortex. The World Health Organization (WHO) reports a prevalence of approximately 70 million individuals affected by epilepsy. These seizures result from electrical discharges in the brain and can have a significant impact on people's lives (Muhammad Usman, Khalid, and Aslam 2020).

The research focuses on assessing the accuracy of epilepsy identification and analysing the effectiveness of the identification process. Machine learning algorithms offer the potential to predict the disease at its early stages, leading to timely intervention and improved patient outcomes. This early detection is particularly advantageous for neurologists, as treatment is more effective in the initial stages of the disease (Siddiqui et al. 2020).

Electroencephalography (EEG) plays a pivotal role in diagnosing epilepsy, as it monitors and records brain activities. However, analysing EEG data requires expertise, is time-intensive, and susceptible to errors (Nanmaran, R et al. 2022). This underscores the importance of machine learning algorithms in detecting epilepsy, as they offer an efficient and accurate means of diagnosis (Zhou et al. 2018) (Vickram, A. S et al. 2020).

The real-time applications of epilepsy disease detection are significant. Neurologists often face challenges in analysing EEG data, a process that consumes substantial time. By leveraging machine learning algorithms for epilepsy detection, diseases can be identified in their early stages, enabling more accurate diagnosis and treatment by medical professionals (Worley 2016).

Over the past five years, there has been a significant body of research focused on epilepsy detection, with a notable number of papers published. Notably, around 30 research articles can be accessed on IEEE Xplore, while Google Scholar hosts approximately 100 related articles. One notable

[*] *Research Scholar*
[†] *Research Guide, Corresponding Author*

Hemanth, G. and Navaprakash, N.
Improved Epilepsy Identification with XGBoost vs. Logistic Regression.
DOI: 10.5220/0012561800003739

In *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT 2023)*, pages 83-90
ISBN: 978-989-758-661-3

contribution involved the development of a novel technique aimed at distinguishing healthy individuals through the utilization of language networks identified via functional magnetic resonance imaging (fMRI). This accomplishment was realized by implementing a machine learning (ML) strategy (Torlay et al. 2017) (G. Sajiv et 2022).

Another study introduced a supervised machine learning classifier designed to accurately predict the outcomes of antiepileptic drug (AED) treatment for individuals recently diagnosed with epilepsy (Yao et al. 2019). Noteworthy outcomes have been proposed, including metrics such as accuracy, precision, recall, F1-score, and proper implementation of the XGBoost algorithm for identifying epileptic seizures at an early stage (Rahman et al. 2021).

In addition, an EPI-AI approach was presented, employing XGBoost for automated and impartial seizure detection using single-channel EEG data across various rodent epilepsy models (Murugesan 2022). A precise method for identifying epileptic seizures was introduced using an integrated approach, CEEMD-XGBoost, which combines complementary ensemble empirical mode decomposition (CEEMD) with extreme gradient boosting (XGBoost) (Wu, Zhou, and Li 2020).

Among these, the work by Torlay et al. (2017) stands out as a particularly influential article in the field, providing valuable techniques for the detection of epilepsy.

Previous studies in the field of epilepsy detection have encountered limitations in achieving accurate results. The primary objective of this research is to address this issue by enhancing the early-stage identification of epilepsy and improving the overall accuracy of the process. This is achieved through the innovative implementation of the Novel XGBoost algorithm, which is compared with the traditional Logistic Regression algorithm.

2 MATERIALS AND METHODS

The research study was conducted at the Department of Electronics and Communication Engineering in Saveetha School of Engineering. The analysis involved the formation of two distinct groups, each containing 15 samples. Group 1 was subjected to the utilization of the XGBoost algorithm with 15 samples, while Group 2 employed the Logistic Regression algorithm with the same sample size. The determination of the sample size for each group was guided by G power calculations, which took into account an 80% pretest power, an alpha error of 0.05, a threshold of 0.95, and a confidence level of 95%.

Table 1: XGBoost algorithm and Logistic regression algorithm are compared with a set of 15 samples from each algorithm and a comparison of the two algorithms respective efficiency percentages.

SL.no	Test	ACCURACY RATE	
		XG Boost Algorithm	Logistic Regression Algorithm
1	Test1	98	81
2	Test2	97	80
3	Test3	96	76
4	Test4	94	78
5	Test5	95	77
6	Test6	90	79
7	Test7	91	75
8	Test8	89	71
9	Test9	88	73
10	Test10	85	72
11	Test11	86	74
12	Test12	84	70
13	Test13	83	69
14	Test14	80	68
15	Test15	79	67
Mean Average (In Percentage)		89.00	74.00

For the assessment of epilepsy presence in individuals, an epilepsy dataset was employed. This dataset was sourced from CHB-MIT and comprises 178 attributes and 11,500 instances, classified into 5 groups. Both groups in the study consisted of 15 samples each. The dataset was divided into training and testing data, with 30 samples allocated for training purposes. Following the partitioning of the data, an algorithm was applied to both the training and testing sets to predict and evaluate accuracy levels (Jones 2018).

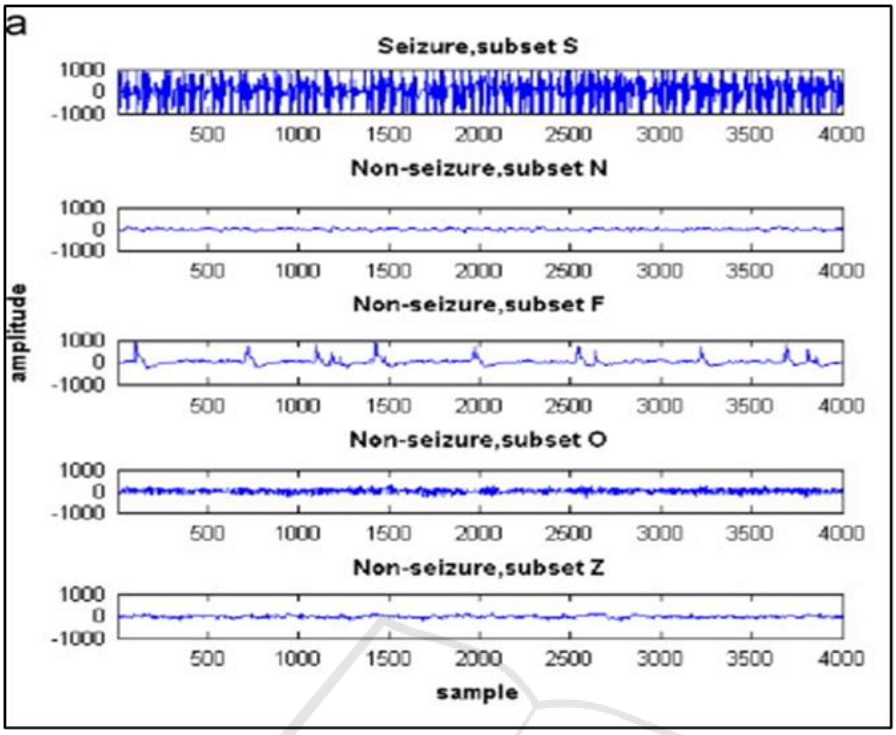


Figure 1: The different subsets, subset S is seizure; the amplitude of the brain waves is high at all the time intervals having more than the 6HZ of amplitude said to be a seizure and the remaining subsets the amplitude is less so it is said to be non-seizure.

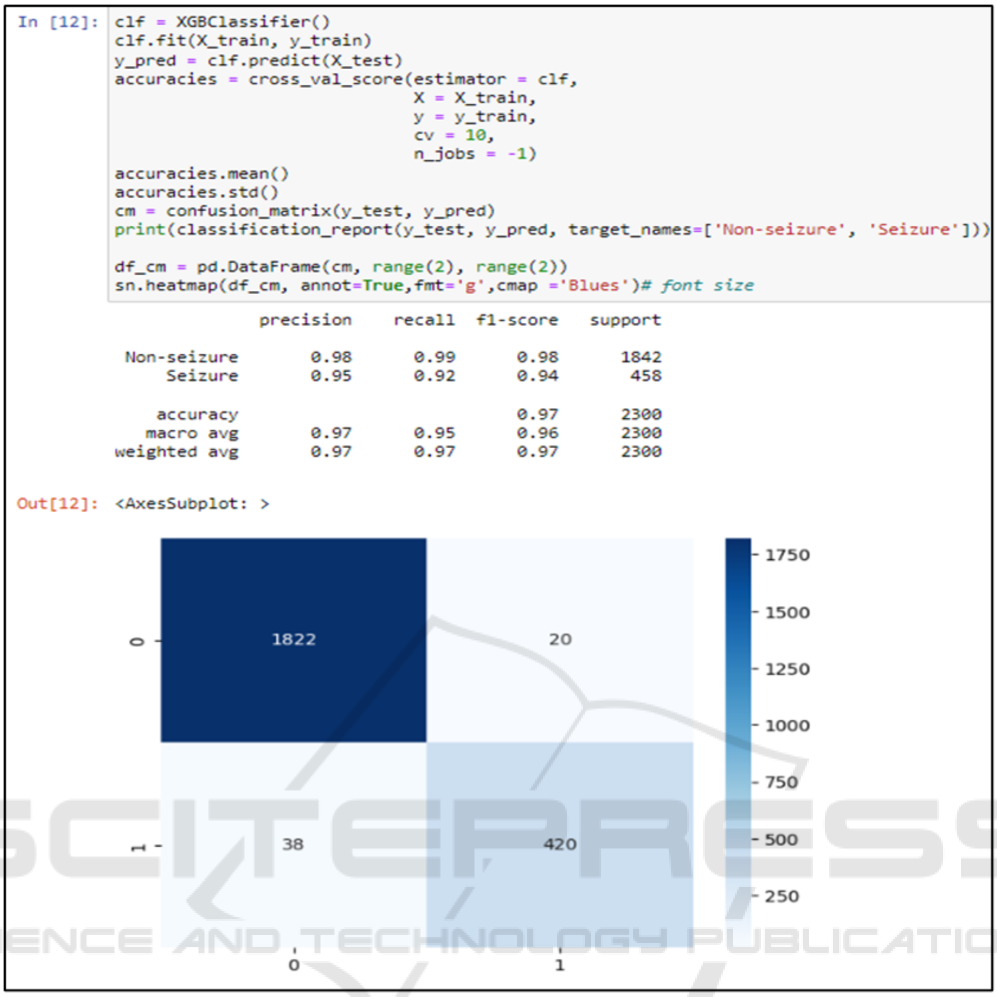
2.1 XGBoost Algorithm

XGBoost, a widely embraced and potent open-source gradient-boosted tree technique implementation, stands out as an ensemble machine learning approach rooted in decision trees and operating within a gradient boosting framework. It demonstrates notable effectiveness in prediction tasks that involve unstructured data types like images and text. One of its key strengths lies in its utilization of parallel processing, tree-pruning techniques, handling of missing values, and incorporation of regularization methods to counteract tendencies of overfitting.

Algorithm for Sample 1 Preparation:

1. Collect and preprocess the data for the epilepsy disease dataset. This may involve feature engineering, normalizing the data, and handling missing values.
2. Divide the dataset into a training set and a validation set to prevent overfitting.
3. Select the optimal XGBoost model for the epilepsy disease dataset. Experiment with hyperparameters such as learning rate, maximum tree depth, and number of estimators to find the best configuration.
4. Train the XGBoost model on the training set. The algorithm will iteratively adjust the decision tree weights to improve performance.
5. Evaluate the performance of the XGBoost model using the validation set. Use various metrics like accuracy, precision, recall, and F1 score to assess the model's effectiveness.
6. Test the selected XGBoost model on a separate test set to measure its performance on new and unseen data.
7. Interpret the model's predictions to understand how it is making decisions. Utilize methods like feature importance analysis, partial dependence plots, and SHAP values to gain insights into the model's behaviour.

Once satisfied with the model's performance, deploy it in a real-world setting. This may involve integrating it with other systems or databases, such as electronic health records or medical equipment, for practical use.



```
In [12]: clf = XGBClassifier()
         clf.fit(X_train, y_train)
         y_pred = clf.predict(X_test)
         accuracies = cross_val_score(estimator = clf,
                                      X = X_train,
                                      y = y_train,
                                      cv = 10,
                                      n_jobs = -1)
         accuracies.mean()
         accuracies.std()
         cm = confusion_matrix(y_test, y_pred)
         print(classification_report(y_test, y_pred, target_names=['Non-seizure', 'Seizure']))

         df_cm = pd.DataFrame(cm, range(2), range(2))
         sn.heatmap(df_cm, annot=True,fmt='g',cmap ='Blues')# font size
```

```
              precision    recall  f1-score   support

 Non-seizure       0.98      0.99      0.98      1842
     Seizure       0.95      0.92      0.94       458

    accuracy                           0.97      2300
   macro avg       0.97      0.95      0.96      2300
weighted avg       0.97      0.97      0.97      2300
```

```
Out[12]: <AxesSubplot: >
```

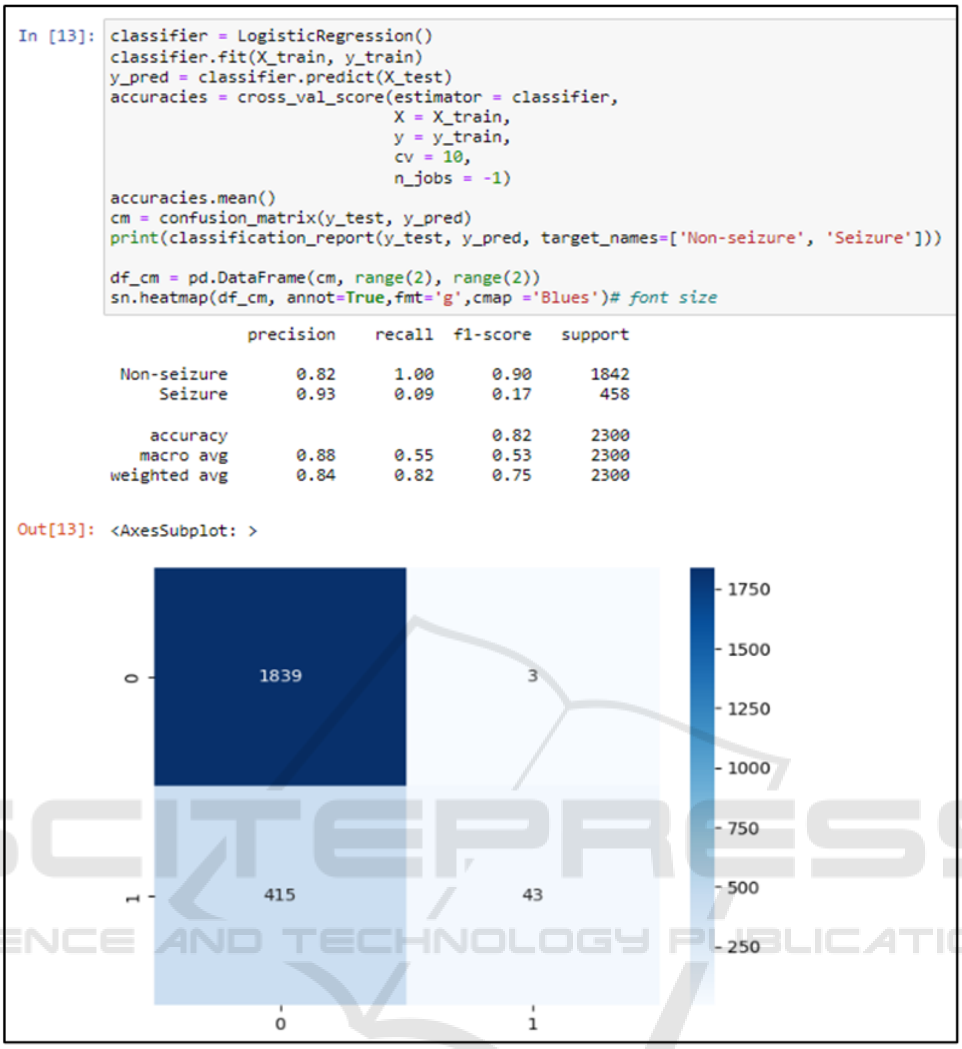
Figure 2: Above confusion matrix shows the performance and classification of the XGBoost algorithm. It predicts all the actual values of the algorithm.

2.2 Logistic Regression Algorithm

Logistic regression, a highly efficient classification technique in supervised learning, is employed to predict categorical outcomes based on independent variables. This method establishes a relationship between these independent variables and the dependent variable, allowing accurate predictions for categorization.

1. Algorithm for sample 2 preparation
2. Import the required libraries.
3. Load the dataset
4. Perform preprocessing of the dataset, which includes data cleaning and data transformation.
5. Perform the feature Engineering and then Exploratory data analysis.
6. Partitioned the dataset into training and testing sets.
7. Validation of the data.
8. Use the Logistic regression algorithm to train and test the data.
9. Show accuracy and confusion matrix.

The computer system configuration comprises an Intel i5 processor, 8GB of RAM, and Windows 11th generation with a 64-bit operating system. The implementation makes use of the Python programming language, along with Jupyter from the Anaconda software suite. In the epilepsy detection model, the independent variables consist of the EEG dataset and frequency signals, while the dependent variables are characterized by improved accuracy metrics. The manipulation of the independent variables directly impacts the associated dependent outcomes.



```
In [13]: classifier = LogisticRegression()
         classifier.fit(X_train, y_train)
         y_pred = classifier.predict(X_test)
         accuracies = cross_val_score(estimator = classifier,
                                      X = X_train,
                                      y = y_train,
                                      cv = 10,
                                      n_jobs = -1)
         accuracies.mean()
         cm = confusion_matrix(y_test, y_pred)
         print(classification_report(y_test, y_pred, target_names=['Non-seizure', 'Seizure']))

         df_cm = pd.DataFrame(cm, range(2), range(2))
         sn.heatmap(df_cm, annot=True,fmt='g',cmap ='Blues')# font size
```

```
              precision    recall  f1-score   support

 Non-seizure       0.82      1.00      0.90      1842
     Seizure       0.93      0.09      0.17       458

    accuracy                           0.82      2300
   macro avg       0.88      0.55      0.53      2300
weighted avg       0.84      0.82      0.75      2300
```

```
Out[13]: <AxesSubplot: >
```

Figure 3: The performance and classification of the Logistic regression algorithm. It predicts all the actual values of the algorithm.

2.3 Statistical Analysis

Using the IBM SPSS statistical tool, the XGBoost algorithm and Logistic regression algorithm were examined through an independent sample T test. This analysis encompassed both dependent and independent variables. The results showed that the XGBoost algorithm achieved a higher mean accuracy of 89% in contrast to the Logistic regression algorithm's accuracy of 74%. The significance level, indicated as p = 0.000 ($p<0.05$), highlighted the statistical significance of this comparison. Notably, the accuracy serves as the dependent variable, while the independent variables encompass features extracted from the dataset, such as mean and variance (Gaur and Gaur 2009).

3 RESULTS

The analysis and identification of epilepsy using the Novel XGBoost algorithm demonstrates superior performance compared to the Logistic regression approach in terms of accuracy. The CHB-MIT dataset includes non-invasive extracranial scalp EEG data collected from 24 patients, featuring 9 to 42 recordings per patient. Each recording was taken at a sampling rate of 256 Hz, covering a duration of around 1 hour. In addition to EEG data, demographic information such as age and gender was also gathered from the patients.

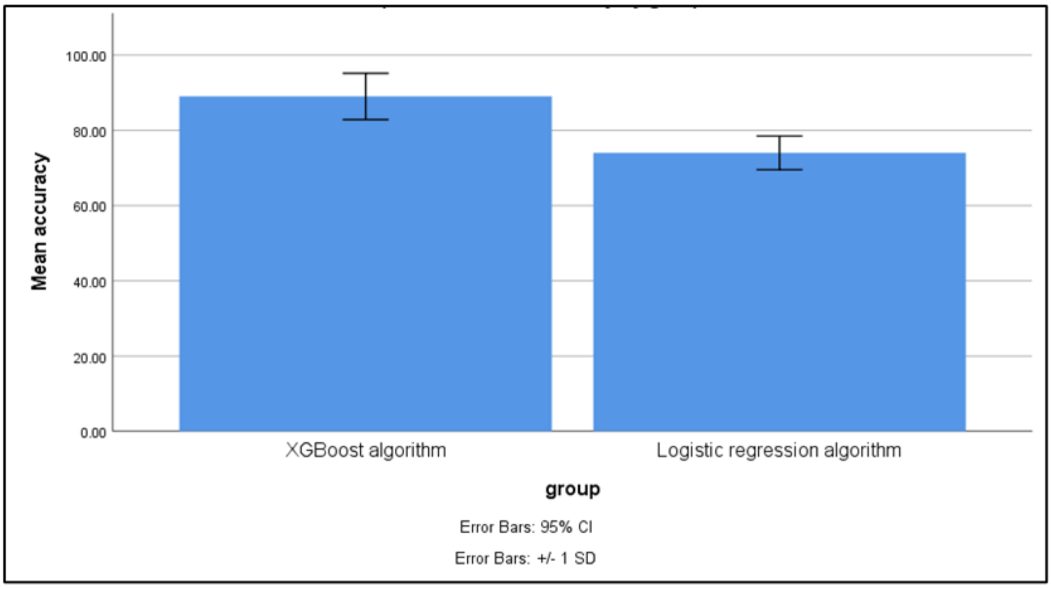


Figure 4: This research evaluates the mean accuracy of the XGBoost and Logistic regression algorithms. The XGBoost algorithm demonstrates a superior mean accuracy of 89%, outperforming the Logistic regression algorithm which achieves a mean accuracy of 74%. The outcomes emphasize the XGBoost algorithm's superiority. The X-axis delineates the algorithm comparison, and the Y-axis signifies the mean accuracy within a range of +1 SD or -1 SD.

Figure 1 illustrates the distinct subsets within the dataset, where subset S represents seizures. In subset S, brain wave amplitudes exhibit high values across all time intervals, surpassing 6 HZ in amplitude, indicative of a seizure occurrence. Conversely, in the remaining subsets, the amplitudes are lower, leading to their classification as non-seizure instances.

Figure 2 provides insights into the performance of the XGBoost algorithm through a confusion matrix. This matrix effectively predicts all the algorithm's actual values, showcasing its classification capabilities.

Figure 3 presents a similar depiction for the Logistic regression algorithm. It showcases the performance of the confusion matrix and the accuracy of predicting all the algorithm's actual values.

In Figure 4, a comparative analysis is presented, highlighting the mean accuracy contrast between the XGBoost algorithm and the Logistic regression algorithm. Notably, the XGBoost algorithm attains a superior mean accuracy of 89%, surpassing the 74% accuracy achieved by the Logistic regression algorithm.

Table 1 showcases the accuracy rates attained by the XGBoost and Logistic regression algorithms. Specifically, the XGBoost algorithm demonstrates a commendable mean accuracy of 89%, while the Logistic regression algorithm achieves a lower accuracy of 74%.

In Table 2, a more detailed insight into the algorithms' performance is provided through mean and standard error values. The XGBoost algorithm exhibits a mean accuracy of 89, accompanied by a standard error of 1.58. In contrast, the Logistic regression algorithm records a mean accuracy of 74, accompanied by a standard error of 1.15. Notably, the independent sample tests highlight a statistically significant distinction in accuracy between the two algorithms. This observation underscores the higher accuracy of the XGBoost algorithm, which is further emphasized by its standard error value of 1.458, distinct from that of the Logistic regression algorithm.

In Table 3, a comprehensive comparison of group distribution and accuracy scores for the XGBoost and Logistic regression algorithms is presented. A statistically significant discrepancy in accuracy, specifically in terms of Point Increment Accuracy, is evident between the two algorithms. This substantial difference is validated through a two-tailed significance test, yielding a probability of p = 0.000 ($p<0.05$). Notably, the XGBoost algorithm emerges as the superior performer with the highest accuracy score of 89%, whereas the Logistic regression algorithm lags behind with an accuracy score of 74%.

Table 2: The mean and standard error are calculated for both the XGBoost algorithm and the Logistic regression algorithm. A statistically significant variance in accuracy between these two algorithms is demonstrated through independent sample tests. Notably, the XGBoost algorithm exhibits a standard error of 1.58565 when compared to the Logistic regression algorithm.

Group Statistics

	Group	N	Mean	Std.Deviation	Std.Error Mean
Accuracy	XGBoost algorithm	15	89.0000	6.14120	1.58565
Accuracy	Logistic regression algorithm	15	74.0000	4.47214	1.15470

Table 3: Contrasting the groups and accuracy of the XGBoost algorithm and Logistic regression algorithm, a notable statistical distinction in point increment accuracy between the two methods is evident. This significance is established through a two-tailed independent sample test with a significance probability of p=0.000 (p<0.05).

		Levene's Test for Equality of variances		T-test for Equality of Means						
		F	Sig	t	df	sig(2-tailed)	Mean Difference	Std. Error Difference	95% confidence interval of the Difference	
									Lower	Upper
Accuracy	Equal variances assumed	1.746	.197	7.647	28	.000	15.000	1.96153	10.9819	19.0180
	Equal variances, not assumed			7.647	25.589	.000	15.000	1.96153	10.9818	19.0180

4 DISCUSSION

The results obtained from the research indicate that the XGBoost algorithm has demonstrated superior accuracy in epilepsy detection, achieving a rate of 89% compared to the Logistic regression algorithm. This conclusion is substantiated through an independent sample t-test. The specific application of the XGBoost algorithm was focused on precise point rate identification, and its performance outperformed that of the Logistic regression algorithm.

The analysis underscores that the innovative XGBoost algorithm excels over the Logistic regression approach, effectively addressing the challenges inherent in epilepsy detection. Supporting this study, various articles have highlighted the strengths of the XGBoost algorithm. For instance, Balachandra et al. (2020) provide supportive insights. Yao et al. (2019) present a machine learning classifier based on XGBoost that accurately forecasts potential antiepileptic drug (AED) treatment outcomes with a remarkable 91% accuracy. In another study, Long et al. (2018) successfully utilized MFCC-based features and XGBoost to achieve an impressive accuracy rate of 97.5% for epilepsy identification. Similarly, Amin et al. (2022) achieved a 96% accuracy with the XGBoost algorithm in the context of epilepsy identification, with no contradictory findings reported.

However, a limitation of this research arises from the unavailability of accessible datasets, constraining the ability to present a deep learning-based Computer-Aided Diagnostic System (CADS) for epilepsy disease. As neuroimaging modality is utilized for diagnosing epileptic seizures due to the lack of datasets, achieving optimal performance in detection becomes challenging. Looking ahead, the future scope of this project involves leveraging relevant datasets to enhance the accuracy of identifying disease symptoms at an early stage. This would enable doctors to promptly detect and effectively treat epilepsy in patients.

5 CONCLUSION

This study undertakes a comprehensive analysis and comparison of two distinct methodologies: the innovative Novel XGBoost algorithm and the conventional Logistic Regression algorithm, both aimed at detecting epilepsy disease. The study's outcomes exhibit a notable disparity in accuracy rates between the two approaches. Specifically, the XGBoost algorithm showcases an impressive accuracy rate of 89%, in contrast to the Logistic Regression algorithm's accuracy of 74%. This discernible variance substantiates the conclusion that the Novel XGBoost algorithm distinctly outperforms the Logistic Regression algorithm in the realm of epilepsy disease detection.

REFERENCES

Amin, Moein, Christopher Newey, Vineet Punia, Stephen Hantus, and Aziz Nazha. (2022). “Personalized Model to Predict Seizures Based on Dynamic and Static Continuous EEG Monitoring Data.” Epilepsy & Behavior: E&B 135 (October). https://doi.org/10.1016/j.yebeh.2022.108906.

Balachandra, Akshara R., Erik Kaestner, Naeim Bahrami, Anny Reyes, Sanam Lalani, Anna Christina Macari, Brianna M. Paul, Leonardo Bonilha, and Carrie R. McDonald. (2020). “Clinical Utility of Structural Connectomics in Predicting Memory in Temporal Lobe Epilepsy.” Neurology 94 (23): e2424–35.

“Developing Window Behavior Models for Residential Buildings Using XGBoost Algorithm.” (2019). Energy and Buildings 205 (December): 109564.

G. Sajiv and G. Ramkumar, (2022) "Multiple Class Breast Cancer Detection Method Based on Deep Learning and MIRRCNN Model," International Conference on Inventive Computation Technologies (ICICT), Nepal, 2022, pp. 981-987, doi: 10.1109/ICICT54344.2022.9850707.

Gaur, Ajai S., and Sanjaya S. Gaur. (2009). Statistical Methods for Practice and Research: A Guide to Data Analysis Using SPSS. SAGE Publications India.

Jones, Nigel C. (2018). “Disease-Modification in Epilepsy by Nonpharmacological Methods.” Epilepsy Currents. https://doi.org/10.5698/1535-7597.18.1.45.

Long, Jie-Min, Zhang-Fa Yan, Yu-Lin Shen, Wei-Jun Liu, and Qing-Yang Wei. (2018). “Detection of Epilepsy Using MFCC-Based Feature and XGBoost.” In 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE. https://doi.org/10.1109/cisp-bmei.2018.8633051.

Muhammad Usman, Syed, Shehzad Khalid, and Muhammad Haseeb Aslam. (2020). “Epileptic Seizures Prediction Using Deep Learning Techniques.” IEEE Access: Practical Innovations, Open Solutions 8: 39998–7.

Murugesan, Balamurugan. (2022). “DETECTION OF EPILEPSY USING MACHINE LEARNING.” California State University, San Bernardino. https://scholarworks.lib.csusb.edu/etd/1384.

Nanmaran, R., Srimathi, S., Yamuna, G., Thanigaivel, S., Vickram, A. S., Priya, A. K., ... & Muhibbullah, M. (2022). Investigating the role of image fusion in brain tumor classification models based on machine learning algorithm for personalized medicine. Computational and Mathematical Methods in Medicine, 2022.

Ramalakshmi, M., & Vidhyalakshmi, S. (2021). GRS bridge abutments under cyclic lateral push. Materials Today: Proceedings, 43, 1089-1092.

Rahman, Ahnaf Akif, Fahim Faisal, Mirza Muntasir Nishat, Muntequa Imtiaz Siraji, Lamim Ibtisam Khalid, Md Rezaul Hoque Khan, and Md Taslim Reza. (2021). “Detection of Epileptic Seizure from EEG Signal Data by Employing Machine Learning Algorithms with Hyperparameter Optimization.” In 2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART). IEEE. https://doi.org/10.1109/biosmart54244.2021.9677770.

Siddiqui, Mohammad Khubeb, Ruben Morales-Menendez, Xiaodi Huang, and Nasir Hussain. (2020). “A Review of Epileptic Seizure Detection Using Machine Learning Classifiers.” Brain Informatics 7 (1): 5.

Torlay, L., M. Perrone-Bertolotti, E. Thomas, and M. Baciu. (2017). “Machine learning–XGBoost Analysis of Language Networks to Classify Patients with Epilepsy.” Brain Informatics 4 (3): 159–69.

Vickram, A. S., Samad, H. A., Latheef, S. K., Chakraborty, S., Dhama, K., Sridharan, T. B., ... & Gulothungan, G. (2020). Human prostasomes an extracellular vesicle–Biomarkers for male infertility and prostrate cancer: The journey from identification to current knowledge. International journal of biological macromolecules, 146, 946-958.

“Website.” n.d. https://doi.org/10.31083/j.jin.2020.01.24.

Worley, Matthew. (2016). “Oi! Oi! Oi!” Fight Back. https://doi.org/10.7765/9781847799616.00011.

Wu, Jiang, Tengfei Zhou, and Taiyong Li. (2020). “Detecting Epileptic Seizures in EEG Signals with Complementary Ensemble Empirical Mode Decomposition and Extreme Gradient Boosting.” Entropy 22 (2). https://doi.org/10.3390/e22020140.

Yao, Lijun, Mengting Cai, Yang Chen, Chunhong Shen, Lei Shi, and Yi Guo. (2019). “Prediction of Antiepileptic Drug Treatment Outcomes of Patients with Newly Diagnosed Epilepsy by Machine Learning.” Epilepsy & Behavior: E&B 96 (July): 92–97.

Zhou, Mengni, Cheng Tian, Rui Cao, Bin Wang, Yan Niu, Ting Hu, Hao Guo, and Jie Xiang. (2018). “Epileptic Seizure Detection Based on EEG Signals and CNN.” Frontiers in Neuroinformatics 12 (December). https://doi.org/10.3389/fninf.2018.00095