International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 10 | Oct 2023 www.irjet.net p-ISSN: 2395-0072

CLASSIFICATION OF ELECTROENCEPHALOGRAM SIGNALS USING XGBOOST ALGORITHM AND SUPPORT VECTOR MACHINE

Oladeji S.O[1], Emuoyibofarhe J. O[2]., Ganiyu R. A[3], Akerele B.A[4]

[1]*Research Scholar, Ladoke Akintola University of Technology, Ogbomoso, Nigeria.*
[2]*Professor, Department of Computer Science, Ladoke Akintola University of Technology, Ogbomoso, Nigeria.*
[3]*Reader, Department of Computer Engeering, Ladoke Akintola University of Technology, Ogbomoso, Nigeria.*
[4]*ECloud Engineer, Vista Entertainment, South Africa.*

ABSTRACT

The automatic identification of epilepsy seizures through the analysis of Electroencephalogram (EEG) data has been an active area of investigation within the biomedical science field. Numerous studies have proposed various methods for classifying EEG signals in recent years. While many of these approaches have demonstrated promising performance, they have often been associated with a high rate of false positives and have also posed computational intensity challenges. This study conducted a comparative assessment of the performance of the Extreme Gradient Boost Algorithm and the Support Vector Machine for the purpose of classifying epileptic seizures within human EEG data. The results of this investigation indicated that the XGBoost Algorithm exhibited superior classification capabilities when compared to the SVM model for EEG signal analysis.

INTRODUCTION

In recent times, significant attention has been directed towards leveraging computer analysis for the examination of bio-electric signals within the human body. Various health conditions in humans can be identified by analyzing these electrical signals, including critical signals governing functions such as heartbeats, brain activity, and those within the central nervous system. Notably, advancements in soft computing and artificial intelligence have substantially enhanced the development of more effective methods for classification, diagnostics, and improvements in treatment approaches (Chen et. al., 2020; Chakole, et al., 2019; Tzallas et al., 2012). Soft computing techniques have played a pivotal role in the extraction and categorization of bio-signals like Electromyography (EMG), electroencephalogram (EEG), Electrooculography (EOG), and electrocardiogram (ECG) to aid in ailment detection and treatment. Diverse methodologies and techniques have emerged for distinguishing and categorizing electroencephalogram (EEG) signals as either normal or indicative of epilepsy. However, the visual analysis of EEG signals in its entirety presents considerable challenges, necessitating the development of automated detection methods.

Epilepsy is characterized by sudden, recurring, and transient disruptions in perception or behavior, stemming from the excessive synchronization of cortical neuronal networks. Epileptic seizures are categorized based on their clinical manifestations into partial or focal, generalized, unilateral, and unclassified seizures (Bhattacharyya and Pachori 2017; Tzallas, Tsipouras, and Fotiadis, 2009). The adoption of classification systems in medical diagnosis has witnessed a substantial rise. It is undeniable that the evaluation of patient data and expert decisions constitute the most critical elements in the diagnostic process. Classification systems play a pivotal role in reducing potential errors that may arise due to fatigue or lack of experience on the part of physicians. Automated diagnostic systems have found application in diverse medical data domains, including electroencephalograms (EEGs), electromyograms (EMGs), ultrasound signals/images, X-rays, electrocardiograms (ECGs), and computed tomographic images (AlZubi, Islam, and Abbod, 2011). This study focuses on conducting a comparative analysis between the extreme gradient boost and support vector machine approaches for the detection and classification of EEG signals as either indicative of epilepsy seizures or non-epileptic seizures, aiming to facilitate more effective management of patients afflicted by epilepsy seizures.

Consequently, the remainder of this paper is structured as follows: an examination of pertinent EEG signal research, a delineation of the methodologies employed in Xboost and SVM, followed by the presentation of results and ensuing discussion. The final section concludes the paper and provides recommendations for future research endeavors.

2. REVIEWS ON ELECTROENCEPHALOGRAM

In recent times, considerable effort has been directed towards harnessing computer-based analysis of the bio-electric signals within the human body. While various methods for examining brain function, such as positron emission tomography (PET), functional magnetic resonance imaging (fMRI), and magnetoencephalography (MEG), have been introduced, the Electroencephalogram (EEG) signal remains a valuable tool for monitoring brain

activity, primarily due to its cost-effectiveness and patient-friendliness. However, the advent of breakthroughs in soft computing and artificial intelligence has significantly advanced the development of more effective techniques for classification, diagnostics, and treatment methodologies (Sharma et. al., 2017; Tzallas et al., 2012). Diverse methodologies and techniques have emerged for the detection and categorization of electroencephalogram (EEG) signals as either normal or indicative of epilepsy. Nevertheless, the comprehensive visual analysis of EEG signals presents considerable challenges, necessitating the essential adoption of automated detection methods.

Subasi et al. (2005) introduced an innovative approach to EEG signal analysis using discrete wavelet transform and subsequent classification with artificial neural networks (ANNs). Their method involved decomposing the signal into five levels using the Daubechies order 4 (DB4) wavelet filter, with input features derived from the energy of details and approximation. Adeli et al. (2007) proposed a methodology combining wavelet analysis, chaos theory, and neural networks for classifying electroencephalograms (EEGs) into healthy, ictal, and interictal EEGs. They employed wavelet analysis to decompose the EEG into delta, theta, alpha, beta, and gamma sub-bands, and used three parameters for EEG representation: standard deviation, correlation dimension, and largest Lyapunov exponent. Their research comprised two phases, aimed at optimizing computing time and output analysis through band-specific and mixed-band analyses. The outcomes indicated the significance of all three key components in enhancing EEG classification accuracy within the wavelet-chaos-neural network methodologies.

Ganesan et al. (2010) proposed a technique for automatically detecting spikes in long-term 18-channel human electroencephalograms (EEGs) using a limited dataset. Their approach for detecting epileptic and non-epileptic spikes in EEGs was founded on a multi-resolution, multi-level analysis and Artificial Neural Network (ANN) approach. However, it is important to note that the results obtained from various research endeavors have consistently demonstrated a high false positive rate.

3. METHODOLOGY

The implementation of this study was conducted using the Python 3.6 software package, specifically within the Spyder 3.5.1 development environment. The operating system utilized was Windows 10 Enterprise, a 64-bit version, running on a system equipped with a Core i5 CPU T4500@2.30GHZ Central Processing Unit, 8GB of RAM, and a 500 Gigabytes hard disk drive with sufficient speed to ensure optimal performance.

To assess and validate the performance of each technique employed, statistical tools, specifically t-test values, were utilized.

The selection of the Python programming language for implementing the system was motivated by its versatility, offering support for multiple programming paradigms and dynamic features, particularly well-suited for machine learning applications.

The methodology employed in this study comprised five principal steps:

a. Data Acquisition
b. Removal of Artifacts
c. Extraction of Features and Decomposition of Extracted Features
d. Classification of Decomposed Features using XGboost and SVM: The final step involved the classification of the decomposed features using the Extreme Gradient Boost (XGboost) and Support Vector Machine (SVM) techniques, which were presumably chosen for their effectiveness in this context.

A. Acquisition of datasets

We utilized a publicly accessible dataset from the Clinique of Bonn University as the foundation for our research. This dataset was generated through the utilization of a 128-channel 12-bit EEG system operating at a rate of 173.5 samples per second. The complete dataset consisted of 500 segments, which were organized into five distinct sets. Each of these segments had a duration of 23.6 seconds. To ensure data quality, rigorous preprocessing was performed to eliminate artifacts originating from eye and muscle movements.

The EEG data we obtained encompassed three distinct scenarios: data collected from individuals without any neurological conditions (healthy subjects), data from individuals with epilepsy during seizure-free intervals (interictal), and data from individuals with epilepsy during active seizure events (ictal). Each of these cases was further divided into five specific segments denoted as Z, O, N, S, and F, which were employed for training and testing the Extreme Gradient Boost (XGboost) and Support Vector Machine (SVM) models.

Segments Z and O were derived from recordings of healthy individuals, with Z corresponding to data acquired when the subjects had their eyes open and O representing data when their eyes were closed. These recordings were captured using a standardized electrode placement scheme on the external surface of the scalp.

Segments N and F were extracted from individuals experiencing interictal phases. Specifically, segment F

originated from epileptogenic regions of the brain, signifying focal activity, while segment N was sourced from the hippocampal region of the brain, indicating non-focal interictal activity.

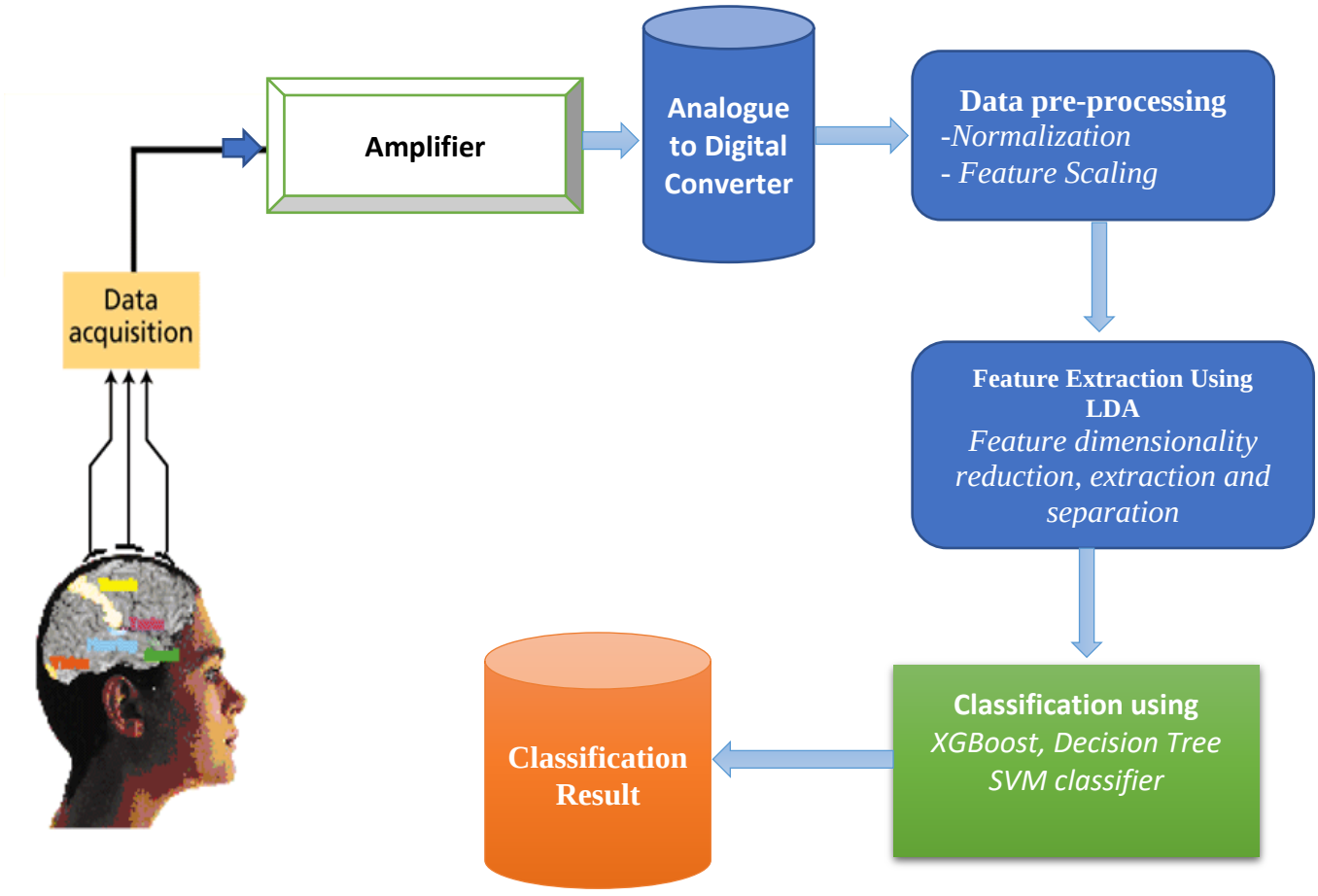
Segment S was obtained from an individual with epilepsy during an active seizure event, providing data relevant to this critical phase of neurological activity.

Within the EEG data, the clinically significant frequency bands of interest include delta, theta, alpha, beta, and gamma, each offering valuable insights for our analysis.

B. Removal of Artifacts

The EEG signal underwent normalization by adjusting the features in such a way that the signal exhibited the characteristics of a standard normal distribution with a mean (average) of μ=0 and a standard deviation of σ=1, where μ represents the mean and σ represents the standard deviation calculated from the mean.

$$z = \frac{x_i - mean(x)}{stdev(x)}$$



C. Feature Extraction Technique and Dimensional Reduction Using LDA

After data normalization, the feature was extracted using LDA.

Given the EEG signals, linear discriminant analysis feature extraction is obtained by performing the following steps:

Step 1: Given a set of N samples $[x_i]_{i=1}^{N}$, each of which is represented as a row of length M as in Figure 2.4 (step (A)), and $X(NxM)$ as given by,

$$\begin{bmatrix} x_{(1,1)} & x_{(1,1)} & \dots & x_{(1,1)} \\ x_{(2,1)} & x_{(2,2)} & \dots & x_{(2,M)} \\ & & \vdots & \\ x_{(N,1)} & x_{(N,2)} & \dots & x_{(N,M)} \end{bmatrix} \quad (3.1)$$

Step 2: Compute the mean of each $\mu_i(1\ x\ M)$ as in Equation 2.2

Step 3: Compute the mean of each $\mu_i(1\ x\ M)$ as in Equation 2.3

Step 4: Calculate between-class matrix $S_B(M\ x\ M)$, as in Equation 3.2 below

$$S_B = \sum_{i=1}^{c} n_i(\mu_i - \mu\ )(\mu_i - \mu\ )^T \quad (3.2)$$

Step 5: for all Class $i, i = 1,2, \dots \dots c\ do$

where S_i represents coefficients of signal s in an orthonormal basis.

Step 6: Compute within-class matrix of each class $S_{W_j}(M\ x\ M)$, as follows:

$$Sw_j = \sum_{x_i \in \omega_j} (x_i - \mu_j)(x_i - \mu_j)^T \quad (3.3)$$

Step 7: Construct a transformation matrix for each class $(W_i)\ as\ follows$:

$$(W_i) = S_{W_i}^{-1} S_B \quad (3.4)$$

Step 8: The eigenvalues (λ^i) and eigenvector (V^i) of each transformation matrix(W_i), are then calculated, where $\lambda^i$and V^i represent the calculated eigenvalues and eigenvectors of the ith class respectively.

Step 9: Sorting eigenvectors in descending order according to their corresponding eigenvalues. The first k eigenvectors are then used as a lower dimensional space for each class (V_k^i).

Step 10: Project all original samples (ω_i) onto their lower dimensional space (V_k^i), as:

$$\Omega_j = x_i V_k^i, x_i\ \epsilon\ \omega_j \quad (3.5)$$

Where Ω_j represents the projected samples of the class ω_j.

Step 11: end for

The linear discriminant analysis of the EEG signals was implemented by using Python 3.7 (Spyder 3.5.1). Using the above procedure, 4-dimension features are extracted from EEG signals.

D. Classification of decomposed feature using XGboost and SVM.

i. Support Vector Machine for classification of extracted features

Gradient boosting serves as the foundational model for XGBoost, where it progressively amalgamates weak base learning models into a more robust learner through iterative steps. During each iteration of the gradient boosting process, the residual error is employed to rectify the preceding predictor in such a way that the specified loss function can be optimized.

Step 1: input data $(x, y)^{N_{i=1}}$
Step 2: Select number of iterations M
Step 3: Choice of the loss-function $\Psi(y, f)$
Step 4: choice of the base-learner model $h(x, \theta)$
Step 5: Initialize $\hat{f}_0$ with a constant
Step 6: for $t = 1\ to\ M\ do$
Step 7: compute the negative gradient $g_t(x)$
Step 7: fit a new base-learner function $h(x, \theta_t)$
Step 9: find the best gradient descent step-size ρ_t:

$$\rho_t = \arg min_\rho \sum_{i=1}^{N} \Psi[y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t)]$$

Step 10: update the function estimate:

$$\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho h(x_i, \theta_t)$$

Step 11: end for

In this research project, the machine learning models were trained using Python, harnessing various scientific computing libraries like NumPy and Pandas. These libraries offer efficient data structures and preprocessing techniques that are crucial for data analysis and model training. Additionally, the project relied on Scikit-learn version 0.18.1 and XGBoost, which were imported to facilitate the implementation and support of Linear Discriminant Analysis (LDA) and XGBoost learning models.

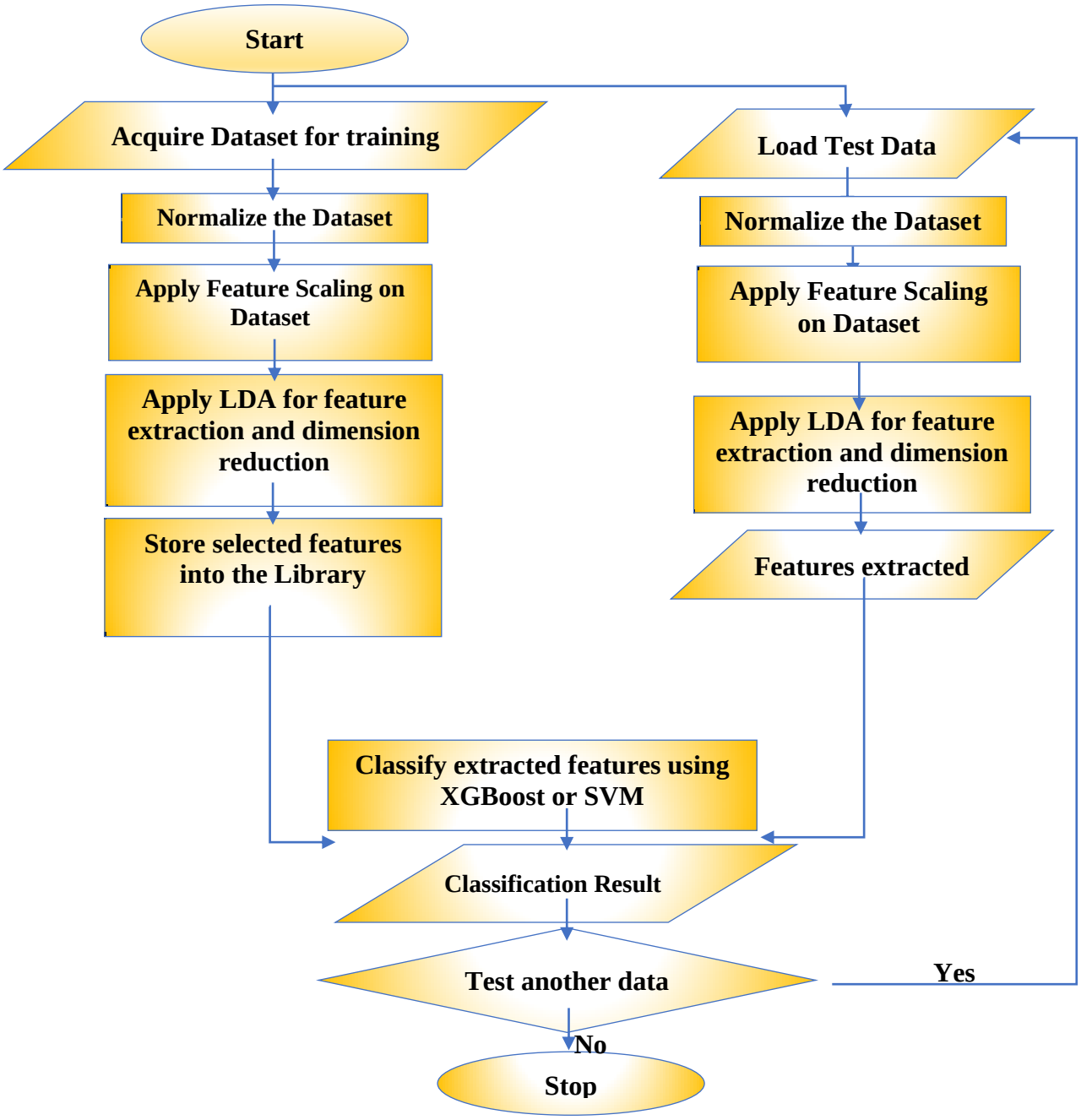


Figure 3.2: Flowchart showing trained and tested EEG signal with XGBoost, SVM and Decision Tree

ii. Support Vector Machine for classification of extracted features

Support Vector Machine (SVM) is employed in this study to classify between normal and seizure activities based on continuous EEG signal recordings. Feature vectors are generated for both seizure and non-seizure activities. These feature vectors are constructed using Linear Discriminant Analysis (LDA), which decomposes the EEG signal into its amplitude and frequency modulated components. Parameters such as the area and mean frequency of these components are estimated and then provided as input for the LDA-SVM classification process.

The Selected best global position (P_i) of the LDA output with the detected feature subset mapped by P_i and modelled with the optimized parameters C and σ using equation (3.20).

$$\min \ \frac{1}{2}\|P_i\|^2 + C\sum_{i=1}^{N}\xi_i \qquad Such\ that \ \sum_{i=1}^{N} P_i x_i \geq \left(\frac{1-\xi_i}{y_i}\right) - b$$

$$i = 1, 2, \ldots, N, \ \xi_i \geq 0, \ i = 1, 2, \ldots, N, \quad (14)$$

Equation (14) was applied to obtain the final classification of each case:

$$y_i = arg\ max_{k(1,\ldots k)}(P_i^T y_i(x_i) + b_i) \quad (15)$$

Where N is the size of the dataset, C is the cost function. I , ξ are the slack variables, x and b is an offset scalar.

iii. The Implementation Phase

The implementation process for both the Gradient Boost algorithm and Support Vector Machine is illustrated in Figure 3.2. The initial stage involved acquiring signal data from human EEG, which underwent preprocessing using the Standard Scaler. The subsequent stage comprised feature selection and dimensional reduction accomplished through Linear Discriminant Analysis. The features extracted were then classified into two categories: epileptic seizures and non-epileptic seizures, employing the XGBoost and SVM algorithms. This entire process was carried out using version 3.2.4 of the Spyder software.

4. RESULTS AND DISCUSSION

The average training times were examined for both the XGBoost and Support Vector Machine (SVM) algorithms across five datasets (Z-O-N-F-S), each consisting of 20,490 data points. The results indicated that the average training time for XGBoost was 2.817 seconds, whereas for SVM, it was 0.610 seconds. This demonstrates that the time required for training increases with larger datasets, suggesting a dependence on the dataset's features for both the XGBoost and SVM models.

For XGBoost, classification results revealed that as the dataset size increased, the computation time also increased. Specifically, for the five datasets (Z-O-N-F-S), XGBoost achieved a false positive rate of 0.33%, sensitivity of 99.68%, specificity of 99.45%, precision of 98.48%, accuracy of 99.06%, and an F-measure of 99.07% at a classification time of 0.01 seconds. For three datasets (O-N-S) of dimension 12,294x100, XGBoost achieved a false positive rate of 1.86%, sensitivity of 92.52%, specificity of 98.14%, precision of 96.53%, accuracy of 96.12%, and an F-measure of 94.49% at a classification time of 0.03 seconds. Similarly, for the five datasets (Z-O-N-F-S), XGBoost had a false positive rate of 2.04%, sensitivity of 86.92%, specificity of 97.96%, precision of 91.55%, accuracy of 95.72%, and an F-measure of 89.17% at a classification time of 0.11 seconds.

The SVM algorithm exhibited a similar trend, with computation time increasing as the dataset size grew. For the five datasets (Z-O-N-F-S), SVM achieved a false positive rate of 1.47%, sensitivity of 99.03%, specificity of 98.53%, precision of 98.55%, accuracy of 98.78%, and an F-measure of 98.79% at a classification time of 0.004 seconds. For three datasets (O-N-S) of dimension 12,294x100, SVM achieved a false positive rate of 3.24%, sensitivity of 94.24%, specificity of 96.76%, precision of 93.47%, accuracy of 95.93%, and an F-measure of 93.86% at a classification time of 0.042 seconds. Similarly, for the five datasets (Z-O-N-F-S), SVM had a false positive rate of 3.54%, sensitivity of 88.99%, specificity of 96.46%, precision of 85.66%, accuracy of 95.02%, and an F-measure of 87.29% at a classification time of 0.016 seconds.

Overall, the results indicated that XGBoost outperformed SVM in terms of classification metrics, but it was computationally more expensive in terms of classification time.

Statistical analyses were conducted to compare XGBoost and SVM. A paired t-test was performed on the False Positive Rate (FPR) and F-Measure between the two models. The analysis showed that XGBoost was statistically significant at a significance level of alpha (α) = 0.05, with a negative mean difference indicating a reduced FPR compared to SVM. This confirms that XGBoost outperformed SVM in terms of FPR.

Similarly, the paired t-test on the F-Measure showed that there was not a significant distinction in the test result, with a mean difference close to zero. However, it confirmed that XGBoost was statistically significant at a significance level of alpha (α) = 0.05, indicating superior performance over SVM in terms of F-Measure.

Table 4.1a: Classification Scheme Results for XGBoosting Algorithm

Dataset	FPR (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	F-Score (%)	Classification Time
Z-S	0.33	99.68	99.45	98.48	99.06	99.07	0.01
O-N-S	1.86	92.52	98.14	96.53	96.12	94.49	0.03
Z-O-N-F-S	2.04	86.92	97.96	91.55	95.72	89.17	0.11

Table 4.1b: Classification Scheme Results for SVM Algorithm

Dataset	FPR (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	F-Score (%)	Classification Time
Z-S	1.47	99.03	98.53	98.55	98.78	98.79	0.004
O-N-S	3.24	94.24	96.76	93.47	95.93	93.86	0.042
Z-O-N-F-S	3.54	88.99	96.46	85.66	95.02	87.29	0.16

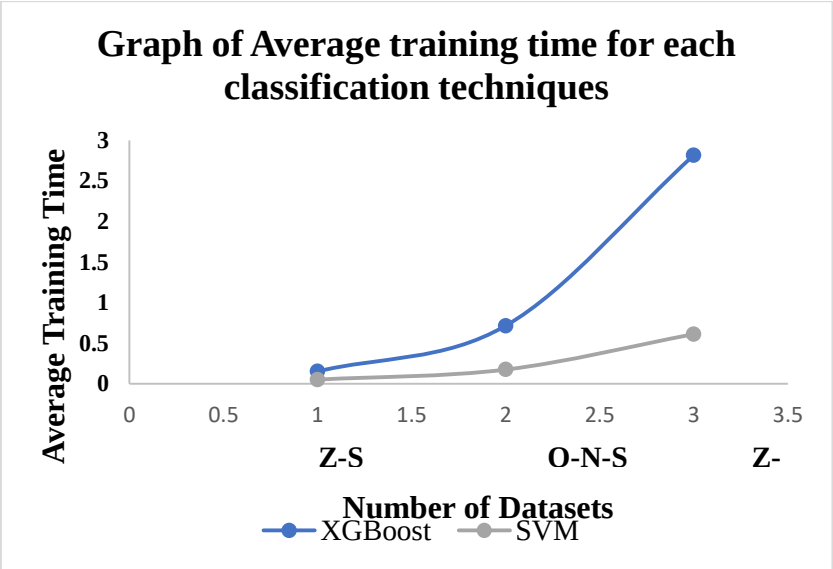


Fig 3.1: Graph of Average time for each techniques.

CONCLUSION AND FUTURE WORKS

The findings derived from this study's experiments demonstrate that the XGBoost model consistently delivered superior results in terms of recognition precision, accuracy, sensitivity, specificity, false positive rate (FPR), and F-measure when compared to the SVM. Consequently, an EEG signal classification system based on the XGBoost model emerges as a more dependable means of detecting seizures or identifying seizure-free states compared to the SVM model.

Future research endeavors may involve exploring the performance of a hybrid approach that combines the XGBoost model with other classifiers. This approach would help ascertain their collective performance across various aspects such as classification accuracy, recognition, and average response time.

REFERENCE

[1] Tzallas (2012). Automatic seizure detection based on time-frequency analysis and artificial neural networks. Computational Intelligence and Neuroscience, 1-13.

[2] Tzallas, A. T., Tsipouras, M. G., and Fotiadis, D. I. (2009). Epileptic seizure detection in EEGs using time-frequency analysis. IEEE Trans Inf Technol Biomed, 13(5), 703-710.

[3] AlZubi, S; Islam, N. and Abbod M. (2011): Multiresolution Analysis Using Wavelet, Ridgelet, and Curvelet Transforms for Medical Image Segmentation. Int J Biomed Imaging. 11(4): 1-4.

[4] Subasi, A., Alkan, A., Koklukaya, E., and Kiymik, M. K. (2005). Wavelet neural network classification of EEG signals by using AR model with MLE preprocessing. Neural Netw, 18(7): 985-997.

[5] Adeli, H., Ghosh-Dastidar, S., & Dadmehr, N. (2007). A wavelet-chaos methodology for analysis of EEGs and EEG subbands to detect seizure and epilepsy. IEEE Trans Biomed Eng, 54(2):205-211.

[6] Ganesan. M, Sumesh. E.P and Vidhyalavanya R, (2010). Multi-Stage, Multi-Resolution Method for Automatic Characterization of Epileptic Spikes in EEG., International Journal of Signal Processing, Image Processing and Pattern Recognition, 3(2): 33-40

[7] Chakole, A.R., Barekar, P.V., Ambulkar, R.V., Kamble, S.D. (2019). Review of EEG Signal Classification. In: Satapathy, S., Joshi, A. (eds) Information and Communication Technology for Intelligent Systems . Smart Innovation, Systems and Technologies, vol 107. Springer, Singapore. https://doi.org/10.1007/978-981-13-1747-7_11

[8] Chen, Z. Lu, G. Xie, Z. and Shang, W. "A unified framework and method for EEG-based early epileptic seizure detection and epilepsy diagnosis," IEEE Access, vol. 8, pp. 20080–20092, 2020, doi: 10.1109/ACCESS.2020.2969055.

[9] Bhattacharyya A. and Pachori, R. B. "A multivariate approach for patientspecific EEG seizure detection using empirical wavelet transform," IEEE Trans. Biomed. Eng., vol. 64, no. 9, pp. 2003–2015, Sep. 2017, doi: 10.1109/TBME.2017.2650259.

[10] Sharma, R. Kumar, M. Pachori, R. B. and Acharya, U. R. "Decision support system for focal EEG signals using tunable-Q wavelet transform," J. Comput. Sci., vol. 20, pp. 52–60, May 2017, doi: 10.1016/j.jocs.2017.03.022.