Machine Learning Framework for Identification of Abnormal EEG Signal

A. Prabhakara Rao(✉), J. Bhaskar, and G. Prasanna Kumar

Department of ECE, Vishnu Institute of Technology, Bhimavaram, AP, India
prabhakararao.a@vishnu.edu.in

Abstract. Epilepsy is a chronic, noncommunicable disease (NCD) causing disorder in the brain's neurological activity. It may be due to genetic disorder or brain injuries caused by some accidents. This may cause seizures, loss of awareness, unusual sensations and behavior. Globally 50 million people are suffering from epilepsy, so it is one of the most prominent neurological diseases globally according to World Health Organization (WHO) statistics in 2021. It is estimated that up to 70% of people are suffering with epilepsy and they can be saved by timely diagnosis and proper treatment of epilepsy. Electroencephalograms (EEGs) are universally used to detect this chronic non-communicable disease. Furthermore, assessing a specific type of abnormality by visual examination of an EEG signal is an intuitive process that can vary from radiologist to radiologist. It is a challenging task for the radiologists to visually examine the EEG signal by looking for a shift in frequency or amplitude in long-duration signals. It may give rise to inaccurate categorization. Identification of epileptic seizure from the recorded EEG signal is a primary task in the treatment of epilepsy. In this work, wavelets were used to obtain the appropriate features from EEG signals. These features were fed to different classifiers. This work proposes a machine learning (ML) framework to detect the abnormality in the EEG signal automatically to assist the radiologists in their diagnosis. The ML framework uses 7 classifiers (KNN, SVM, Random Forest, Logistic Regression, Decision Tree, AdaBoost, and Bagging). Among these classifiers, Bagging Classifier was shown better performance in terms of accuracy and ROC.

Keywords: Electroencephalogram (EEG) · Epilepsy · Seizure · Wavelets · Machine learning (ML) framework · ML classifiers/

1 Introduction

Digital image and signal processing applications are spread in many fields [1, 2], among these, the application of image processing and signal processing in the medical field is a trending technique [3] in the current scenario. According to World Health Organization (WHO) estimates from 2021, epilepsy is one of the most prevalent neurological illnesses worldwide. Approximately 50 million people will be impacted by it, according to WHO figures until 2021 [4]. It results in muscle stiffness, seizures, etc. Prolonged seizures

N. Gupta et al. (Eds.): IC4S 2022, LNICST 472, pp. 42–54, 2023.
https://doi.org/10.1007/978-3-031-28975-0_4

can harm the brain. The brain's ability to operate will probably be negatively affected by isolated, brief seizures, and some brain cells may even be lost. More than half of the population suffering from epilepsy could live seizure-free lives if the condition was adequately recognized and treated in its early stages. It is a severe neurological condition with distinctive traits that is prone to recurrent seizures. This illness affects all mammal species, including rats, dogs, and cats in addition to people. However, the term "epilepsy" is unremarkable and consistently distributed around the world; it offers no hints as to the kind or severity of the seizures [5]. The classification of seizures is shown in Fig. 1.

Based on the symptoms and signs, seizures are classified into two major groups – focal and generalized [6]. Focal seizures affect one side of the brain (hemisphere) and the patient may lose consciousness. Partial or focal seizures are classified into simple and complex. Simple focal seizures are characterized by staring spells, automatisms, and sensory phenomena. Complex focal seizures involve confusion and disorientation. Generalized seizures influence both hemispheres affecting both the sides of the brain simultaneously and are accompanied by tonic-clonic movements. These include absence, myoclonic, tonic, and tonic-clonic seizures [7]. Absence seizures are characterized by sudden loss of consciousness. Myoclonus refers to jerking movements. Tonic seizures are characterized by stiffening of muscles. Tonic-clonic seizures may lead to rhythmic contractions of muscles. Generalized seizures basically classified into two categories namely convulsive and non-convulsive.

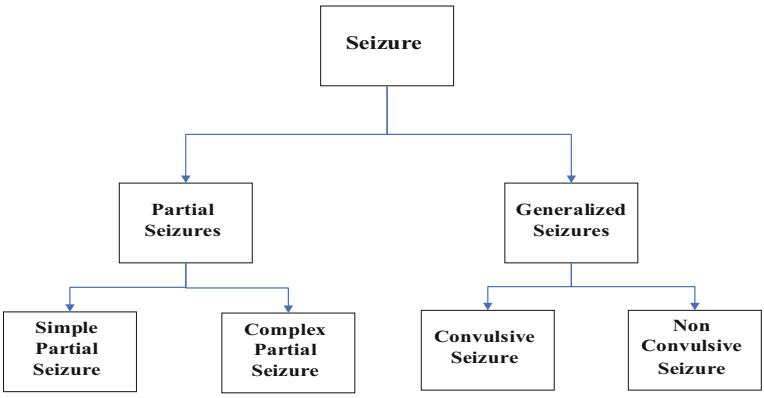

Fig. 1. Classification of seizures

Various non invasive approaches such as functional magnetic resonance imaging (fMRI), positron emission tomography (PET), nuclear magnetic resonance fluorescence, single-photon emission computed tomography, near-infrared spectroscopy, electroencephalogram (EEG) are commonly used to study the function of the brain. Among all these techniques, EEG is widely used because it records the electrical behaviors of the brain very accurately. Also, EEG is a simple, safe and less painful test. In diagnostic applications, priority is frequently given to spectral data. Electroencephalography (EEG), a method of electrophysiological monitoring, captures the brain's electrical activity. The electrodes are typically non-invasive and placed throughout the scalp, while occasionally invasive electrodes are used in specific applications. EEG records voltage changes in the brain's neurons due to ionic current. In therapeutic settings, an EEG is a long-term recording of the electrical activity of the brain made with many electrodes positioned on the scalp [8]. Electroencephalograms (EEGs) are universally used to detect epilepsy. Assessing a specific type of abnormality by visual examination of an EEG signal is an intuitive process that can vary from radiologist to radiologist. It is a challenging task for the radiologists to do visual examination of the recorded EEG signal and identify the shift in frequency or amplitude in the EEG signals of long-duration. It may give rise to inaccurate categorization. Determination of epileptic seizure is an essential task in the treatment of epilepsy. In this work, wavelets were used to obtain the appropriate features from EEG signals. Among the various signal processing techniques, wavelet transforms have the ability to efficiently identify the subtle changes in the EEG signal [9]. The wavelet features were fed to different classifiers. This paper proposes a machine learning (ML) framework consisting of 7 ML classifiers to detect the abnormality in the EEG signal automatically to assist the radiologists in their diagnosis.

The work presented in this paper is organized in the following manner. Section 1 gives brief introduction to Epileptic Seizures, EEGs, and the objective of the presented work. Section 2 deals with some of the related work. The methodology of the work is presented in Sect. 3. The results and discussions are included in Sect. 4, and the conclusions are described in Sect. 5.

2 Related Works

Several studies have been carried out to develop a system that can reliably identify abnormal EEGs in humans. This is because epilepsy [10], sleep disorders [11], and other conditions may be identified if these EEGs are correctly classified. These studies give equal attention to seizure detection and seizure prediction. The EEG signals which are non-linear in nature and are dynamic are difficult to analyze through linear techniques to produce consistent, accurate results. As a result, alternative machine learning (ML) or deep learning (DL) methodologies are applied in diverse investigations. Different features are extracted in the ML investigations. Wavelet transform, Hilbert-Huang transform, Eigen value decomposition, higher-order spectra, and cumulate features are some of the common examples. Depending on the study, one channel or multi-channel signals may be used. Different classifiers are employed in ML to categorize signals based on their signature or extracted features, such as k-nearest neighbor (KNN), support vector machine (SVM), random forest (RF), bagged trees, etc. Component analysis can also

be used to categorize EEGs, in addition. Principal component analysis (PCA) was performed by Lopez et al. [12] with the KNN and RF classifiers, yielding accuracy rates of 58.2% and 68.3%, respectively. After pre-processing the data to remove various artifacts and noise, signal processing tasks begin with normalizing of the signals. The attributes are extracted after pre-processing. The collected attributes are then supplied to the classifiers, and the effectiveness of the classification is evaluated. If a model performs as expected, it is put to further test using a fresh, unrelated set of data. Studies presently use a variety of deep learning-based techniques that don't necessitate feature extraction and selection. Using the same database and a one dimensional convolutional neural network (1D-CNN) with single-channel signals lasting 60 s, Yildirim et al. [13] identified the aberrant EEG signals and discovered an error rate of 20.6%. Diego et al. [14] proposed a system that combined 2D CNN with ML on four-channel signals and achieved an error rate of 21.2% in the detection of abnormal EEG signals. Acharya et al.[10] suggested a CNN-based method for automatically distinguishing between seizure and non-seizure EEG patterns (13 layers). 300 signals from 5 patients were employed in the investigation, and the classification accuracy was 88.67%. A 13-layer CNN model was used by Oh et al. [15] to provide a strategy for the detection of Parkinson's disease (PD). They were able to attain an accuracy of 88.25% in their investigation by using EEG data from 20 healthy people and 20 patients with Parkinson's disease. Existing works on automatic identification of abnormal EEG signal using deep learning (DL) methods showed better accuracy but the computational load and memory requirements are high. Objective of the proposed work in this paper is to identify a suitable classifier and set of wavelet features that will identify abnormal EEG signal efficiently with minimum time and minimum computational load.

3 Methodology

This section describes the various steps involved in the implementation of the proposed machine learning framework for identification of abnormal EEG signal. The entire procedure was illustrated in a flowchart as shown in Fig. 2.

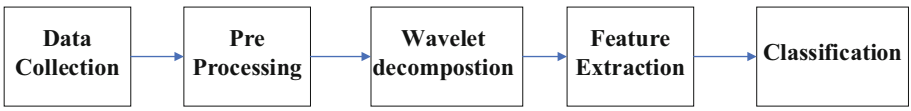

Fig. 2. Methodology of the proposed work

3.1 Data Collection

Epileptic seizure EEG data was collected from the Bonn dataset, an open-source data repository. The process of how these EEG recordings were obtained was explained in [16]. This data set contains five folders, each of which has 100 files that each represents a different subject or individual. An EEG recording of brain activity for duration of 23.6

s was stored in each file. The relevant time series 4097 data points are sampled in this analysis. As a result, we have collected 4097 data points across 23.5 s for a total of 500 candidates. Every 4097 data points were randomly divided into 23 pieces for each of the 178 sets of data that made up each chunk. Each data point displays the value of the EEG recording at a particular moment. Thus, we get $23 \times 500 = 11\,500$ bits of information, each of which contains 178 datasets for a time of one second, with the last column 179 designating the response label y: 1, 2, 3, 4, and 5. The response variable (y) indicates the conditions under which the patient's EEG signal was recorded. Condition-5 (eyes open): The patient's eyes were open while the brain's EEG data was being captured, Condition-4 (eyes closed): The patient's eyes were closed while the EEG signal was being recorded, Condition-3 (tumor located, EEG in the normal area): tumor location in the brain was determined, and the healthy brain area activity was recorded through the EEG signal. Condition-2 (tumor area): brain activity in the tumor region. Condition-1: Seizure activity recording. Thus, the conditions-(1, 2, 3) and conditions-(4, 5) indicate unhealthy and healthy people, respectively. One healthy record and one unhealthy record which represent seizure activity were considered in this work from the given data.

3.2 Data Preprocessing

Initially the sampling frequency of the dataset is 178.3 Hz. It was resampled to 128 Hz frequency using bandpass filter and notch filter. All frequencies falling inside the pass-band would be sent to the output without being amplified or attenuated, whereas all frequencies falling outside the passband would be totally attenuated in an ideal band-pass filter. No bandpass filter is perfect in real life. As a result, we have filter roll-off. To eliminate this filter roll-off, notch filter was used as it blocks a specific band of frequencies and allows all frequencies outside the band.

3.3 Multilevel Wavelet Decomposition

Fourier Transforms, Fast Fourier Transform, Long Time Fourier Transform, Short Time Fourier Transform, Wavelet Transform, etc. can be used in the analysis of the EEG signal in frequency domain. But Wavelet transforms are highly efficient and robust while dealing with discrete signals. An orthogonal wavelet decomposition technique can be used to break down a signal into its component parts. A basic hierarchical framework is provided by a multi-resolution representation to examine the signal at various resolution levels. This is comparable to the idea of breaking down a signal into Walsh, Haar, or Fourier transform components. A signal is uniquely and entirely represented by its orthogonality. According to the Mallat theory, a signal's multi-resolution representation can be used to analyze its information content at various levels of detail [17]. A signal can be approximated by this operator at a specific resolution. Figure 3 depicts the wavelet's decomposition process used in the proposed work.

In the process of a wavelet transformation of a signal (S) is first decomposed into approximate coefficients and detailed coefficients. The signal's approximate [A1] (low frequency components) coefficient is the output of a low pass filter, and the signal's detailed [D1] (high frequency components) coefficient is the output of a high pass filter. This Approximate coefficient [A1] is again passed through a low pass to get approximate

coefficient [A2] and A1 is passed through a high pass filter to obtain the detailed coefficient [D2]. Further A2 is decomposed into approximate coefficient [A3] and detailed coefficient [D3]. The number of decomposition levels depends on the length of signal and our requirements.

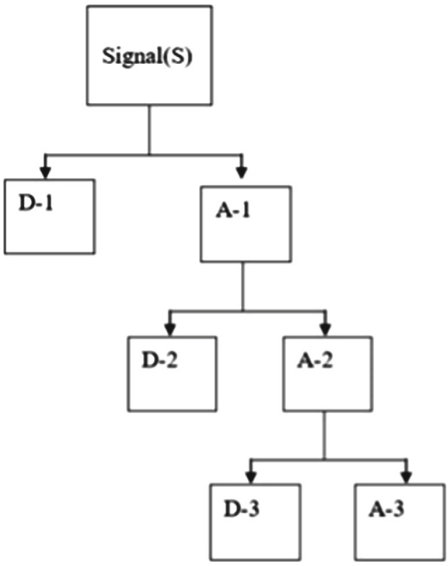

Fig. 3. Wavelet decomposition

The Original Signal S can be reconstructed with the help of A3, D3, D2 and D1.

With the decomposed wavelet coefficients the original signal can be reconstructed. The number of samples in next decomposition level is half as compared to previous stage. A1 and D1 will have N/2 samples if the original signal S had N samples, while A2 and D2 will have N/4 samples. The investigation of local signal behavior, such as spikes or discontinuities, is hence well suited for the wavelet transform. Because the frequencies change quickly and for a brief period at the site of discontinuity, we can investigate or analyze these abrupt shifts by selecting an appropriate time scale.

Db4 wavelet was applied on the two selected records to get the 4-level wavelet decomposition. As a result, 5 wavelet coefficients (a, d1, d2, d3, d4) were obtained for each record i.e., healthy and unhealthy. The dimensions of each coefficient record after applying the 4-level wavelet decomposition is 4098×100.

3.4 Feature Extraction

In this work, six features were calculated from each of the wavelet coefficient record obtained after wavelet decomposition. They are mean, variance, skewness, kurtosis,

max_svd, entropy_svd. Definition and mathematical equation of the six features were illustrated below.

Mean
It is the ratio of the sum of all the compliances in the data to the total number of compliances. Therefore, the mean is a number surrounding which the entire data is spreading. It can be calculated as shown in Eq. 1.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1}$$

where N represents the total number of observations and $\sum xi$ = sum of the observation

Variance
Variance measures how much variation there is within a group of data points. A low variance means that the data points are close together, while a high variance implies the data points are spread apart. It can be calculated using the formula shown in Eq. 2.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{2}$$

Skewness
Asymmetry in a probability distribution function is measured as Skewness. It can be calculated using the formula shown in Eq. 3.

$$skewness = \frac{3(mean - median)}{standard\ \ deviation} \tag{3}$$

Kurtosis
Kurtosis describes the tagging of data whether it is lightly tagged, heavily tagged when compared to a normal distribution.

Max_svd
It is the maximum of singular value decompositions.

Entropy_SVD
It is a measure of the dimensionality of the data.

The SVD entropy of a signal X is defined as

$$H = \sum_{I=1}^{M} -(P_i * lnP_i) \tag{4}$$

where

- P_i = normalized value of i^{th} singular value of X,
- M = Total number of singular values in the embedded matrix X,

3.5 Classification of Abnormal EEG

In the proposed ML frame work 7 classifiers were used. They are namely Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor (KNN), Logistic Regression, Random Forest, AdaBoost, Bagging classifiers.

3.6 Performance Parameters

ML framework was built by using different classification models based on train data and predicts the results and compares them with test data using some parameters. They are Accuracy, Precision, Recallf1, Score, Confusion Matrix and ROC Curve.

Confusion Matrix
It serves as a performance indicator for classification problems using machine learning. It is a table containing four separate sets with actual and anticipated values. Recall, Precision, Specificity, Accuracy, and most critically area under the curve-receiver operating characteristic curve (AUC-ROC) are all very well measured by it. Accuracy and AUC are considered as performance measures in this work.

Accuracy
It describes the percentage of accurate predictions from the test records. It can be calculated from Eq. 5.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

ROC Curve
Receiver operating characteristic (ROC) curve is a graphical representation of the performance of different classification models at all thresholds. It is plot of true positive rate versus false positive rate.

4 Results and Discussions

The classification report in terms of confusion matrix and RoC curve of the seven classifiers (SVM, Decision Tree, KNN, Logistic Regression, Random Forest, AdaBoost, and Bagging) were obtained and their performance was compared using accuracy and AUC-ROC. Confusion matrix and ROC Curve of the seven classifiers were shown in Figs. 4, 5, 6, 7.

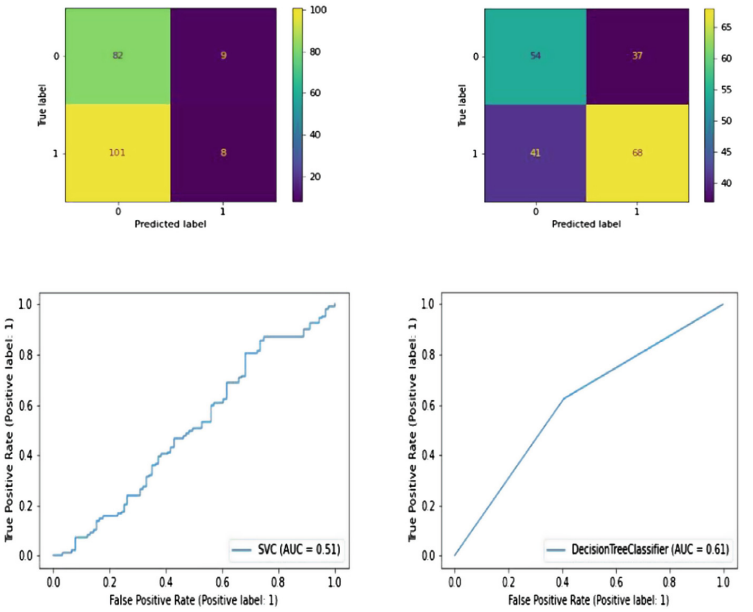

Fig. 4. Confusion matrix and ROC Curve of the SVM, Decision Tree Classifiers

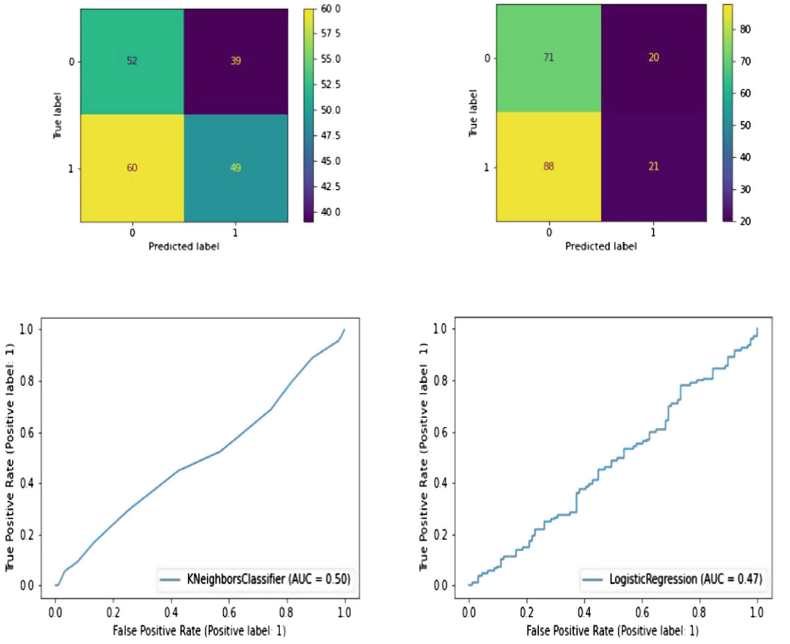

Fig. 5. Confusion matrix and ROC Curve of KNN and Logistic Regression Classifiers

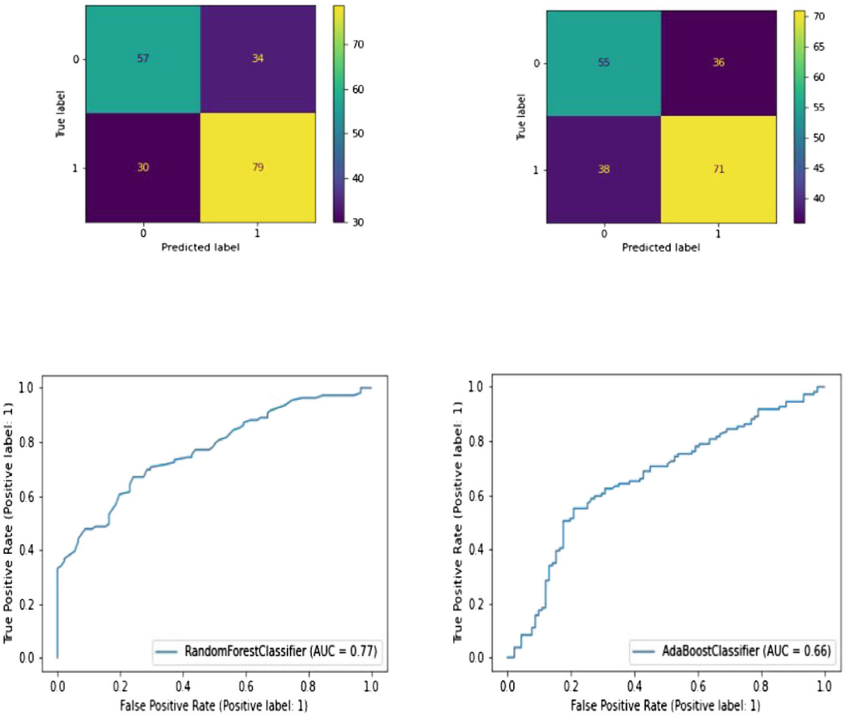

Fig. 6. Confusion matrix and ROC Curve of Random Forest and AdaBoost Classifiers

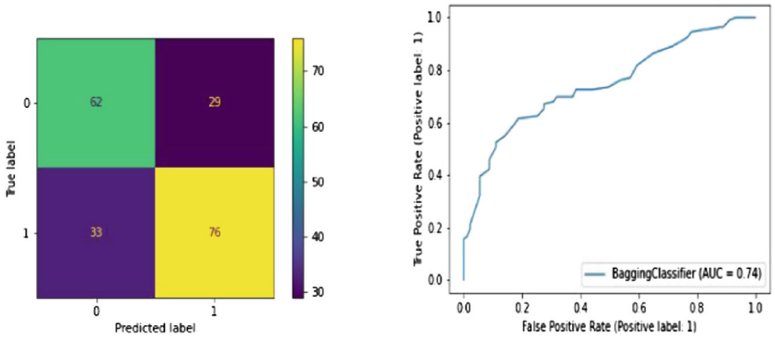

Fig. 7. Confusion matrix and ROC Curve of Bagging Classifier

Performance comparison of all the seven classifiers in terms of accuracy and % AUC-ROC were summarized in Table 1 and Fig. 8.

From Table 1 and Fig. 8, it was observed that among all the seven classifiers used in the ML framework, bagging classifier has shown better performance with a classification accuracy of 70% and AUC-ROC curve of 74%. This classifier can be used effectively in

Table 1. Performance of various classifiers used in the ML framework

Classifier	Accuracy (%)	AUC-ROC (%)
SVM	48	51
Decision Tree	65	61
Random Forest	68	77
AdaBoost	58	66
KNN	56	50
Logistic Regression	48	47
Bagging	70	74

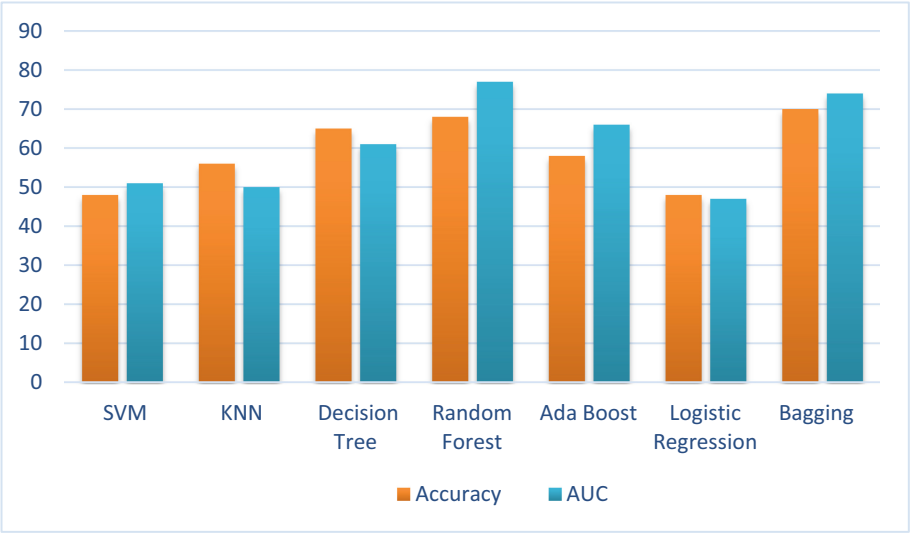

Fig. 8. Performance plot of the various classifiers

the detection of abnormal (seizure activity in EEG) EEG. For long-duration transmissions, it will be difficult for the clinicians to visually examine the EEG signal to identify the frequency or amplitude changes. Therefore the ML framework presented in this paper can be used in the automatic diagnosis of EEG signal. It improves the diagnostic accuracies as compared to the manual examination reducing the load on the clinicians.

This model's primary goal is to assist the radiologist in making an accurate diagnosis of EEG abnormalities. So, to categorize normal and abnormal EEG, which reflects seizure activity, an ML Framework with a variety of classification methods was used in the presented work. The EEG input was first fed into the ML framework model, which then underwent 4-level wavelet decomposition. The mean, variance, skewness, kurtosis, max_svd, and entropy characteristics are extracted from all the coefficients obtained after wavelet decomposition and then these features are given to several classification

algorithms in the ML framework model, such as SVM, Decision Tree, KNN, Logistic Regression, Random Forest, AdaBoost, and Bagging. The bagging classifier in the proposed system performed well for the given EEG data with 70% accuracy and 76% AUC among all the classification algorithms. This framework model can be employed in a practical setup to analyze the EEG Signals in real time because it requires less computational resources and time as compared to the CNN models. Although the performance of the proposed model is less compared to CNN models it is good in terms of less computational load, memory requirement and time. In addition, performance of the ML framework presented in this work can be improved by analyzing on the suitable wavelet features among the five wavelet coefficients obtained from the four level wavelet decomposition. This ML framework can also be used to treat other anomalies, such as sleep disorders and other neurological disorders.

5 Conclusions

Information processing in the brain signal was recorded through EEG. Dynamic changes in the brain activity can be recorded through EEG which produces electrical signals varying in time, frequency and space. Various non linear and time-frequency analysis methods were used to analyze the EEG. Among these time-frequency analysis techniques, wavelet transforms were proven to be better as they efficiently capture the dynamic and subtle changes in the EEG signal. Therefore in the proposed work five wavelet features were extracted by using db4 wavelets on the EEG signal and these features were given as inputs to the seven classifiers to detect abnormal EEG. Performance of these classifiers was analyzed and it was observed that bagging classifier was showing better performance for the given EEG data with an accuracy of 70% and area under ROC curve of 74%. The proposed ML framework was used to automatically detect the abnormal EEG, which will be used to assist the clinicians. This approach can lighten the workload of the radiologist who is responsible for manually and visually identifying seizures on long-duration EEG signals.

References

1. Budumuru, P.R., Kumar, G.P., Raju, B.E.: Hiding an image in an audio file using LSB Audio technique. In: International Conference on Computer Communication and Informatics (ICCCI). IEEE (2021)
2. Sahu, S., Rao, A.P., Mishra, S.T.: Fingerprints based gender classification using adaptive neuro fuzzy inference system. In: International Conference on Communications and Signal Processing (ICCSP). IEEE (2015)
3. Rao, A.P., Bokde, N., Sinha, S.: Photoacoustic imaging for management of breast cancer: a literature review and future perspectives. Appl. Sci. **10**(3), 767 (2020)
4. World Health Organization: Epilepsy. Epilepsy Key facts 2022 [cited 2022 7 Oct 2022]. Available from: https://www.who.int/news-room/fact-sheets/detail/epilepsy
5. Reynolds, E.H., Rodin, E.: The clinical concept of epilepsy. Epilepsia **50**, 2–7 (2009)
6. Sharma, M., Patel, S., Acharya, U.R.: Automated detection of abnormal EEG signals using localized wavelet filter banks. Pattern Recogn. Lett. **133**, 188–194 (2020)

7. Mayoclinic: Eplilepsy overview (2021) [cited 7 Oct 2022]. Available from: https://www.mayoclinic.org/diseases-conditions/epilepsy/symptoms-causes/syc-20350093
8. HealthMedia: EEG (Electroencephalogram) overview. EEG Overview 2022, 9 Nov 2021 [cited 7 Oct 2022]. Available from: https://www.healthline.com/health/eeg
9. Faust, O., et al.: Wavelet-based EEG processing for computer-aided seizure detection and epilepsy diagnosis. Seizure **26**, 56–64 (2015)
10. Acharya, U.R., et al.: Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. Comput. Biol. Med. **100**, 270–278 (2018)
11. Michielli, N., Acharya, U.R., Molinari, F.: Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals. Comput. Biol. Med. **106**, 71–81 (2019)
12. Lopez, S., et al.: Automated identification of abnormal adult EEGs. In: IEEE Signal Processing in Medicine and Biology Symposium (SPMB). IEEE (2015)
13. Yıldırım, Ö., Baloglu, U.B., Acharya, U.R.: A deep convolutional neural network model for automated identification of abnormal EEG signals. Neural Comput. Appl. **32**(20), 15857–15868 (2018). https://doi.org/10.1007/s00521-018-3889-z
14. Diego, S.L.d.: Automated identification of abnormal EEGs. In: Electrical Engineering, p. 63. Temple University (2017)
15. Oh, S.L., et al.: A deep learning approach for Parkinson's disease diagnosis from EEG signals. Neural Comput. Appl. **32**(15), 10927–10933 (2018). https://doi.org/10.1007/s00521-018-3689-5
16. Andrzejak, R.G., et al.: Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. Phys. Rev. E **64**(6), 061907 (2001)
17. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**(7), 674–693 (1989)