Comparison of Different Machine Learning Algorithms to Classify Epilepsy Seizure from EEG Signals †

Pankaj Kunekar *, Chanchal Kumawat, Vaishnavi Lande, Sushant Lokhande, Ram Mandhana and Malhar Kshirsagar

Abstract: Recurrent seizures are a symptom of a central nervous system disease called epilepsy. The duration of these seizures lasts less than a few seconds or sometimes minutes. There are very few ways to record seizures, and one of them is EEG. EEG systems mainly consist of scalp electrodes that record electrical activity. These EEG data are often complex signals containing noise and artifacts. Accurate classification of epileptic seizures is a major challenge, as manual seizure identification is a laborious and challenging endeavor for neurologists. An automated method for seizure detection and categorization was required to address this issue. In this paper, we used machine learning and proposed a model that predicts the behavior of these signals and classifies seizures. The Epileptic Seizure Recognition Data Set from the UCI Machine Learning Repository was the dataset used in this work. The model is evaluated on various models such as XGboost, Extra Tree Classifier, Random Forest, etc. Using measures like F1 score, recall, and precision, the proposed approaches have been assessed. The results indicate that Random Forest produced the superior result of 0.943 F1 score, and XGB achieved a slightly lower F1 score of 0.933. Moreover, Random Forest has the highest accuracy of 0.977.

Keywords: Musa epilepsy; EEG; ETC; XGBoost; F1 score

1. Introduction

In the present years, the brain–computer interface with clever signal decomposition has played a major role in the fields of medicine, defense services, etc. [1]. It is desirable to have a skilled approach to feature extraction from EEG signals that will make brain–computer interface building easier. The communication between neurons in the brain takes place using electrical signals. The brain activities of living beings are based on these electrical signals. At some point, a failure to communicate causes a brain seizure [2]. The electroencephalogram (EEG) data reveal the brain’s electrical activity. Epilepsy, for instance, can be identified by examining the EEG waves [3]. Seizures are characterized by abnormal brain activity brought on by an epileptic disturbance that interferes with normal brain or body function and affects the central nervous system. The neurological condition epilepsy is characterized by repeated seizures, which are short (lasting less than minutes) bursts of uncontrollable movement that can affect one or all body parts and are occasionally followed by loss of consciousness and control over bowel or bladder function. High electrical performance in a bunch of cerebrum chambers is the result of seizures. Such discharges can occur in many areas of the brain. Muscle jerks used to prolong seizure episodes might be different in seizures. The frequency of seizures also varies. The progressive neurobiological process, known as “epileptogenesis”, is the cause
of epilepsy. Its signs include disorientation, strange behavior, and awareness loss [4]. These symptoms can lead to injury from falls or tongue-biting.

A WHO survey notes that almost 7 million of the globe’s population have seizures, leading to the most common neurological disease [5]. It could be difficult to predict when someone could experience a seizure. Therefore, we can say that an epileptic seizure is a dangerous disorder, as it is not inculcated when it arrives. So, for this reason, EEG signals play a major role in the detection of epileptic disorders. To diagnose and conduct a thorough investigation of the brain during an epileptic seizure event, electroencephalography (EEG) is useful. The study of epilepsy recovery strategies involves utilizing EEG, a technique that examines non-Gaussian and non-stationary signals representing electrical activity in the brain. These signals are employed to detect various types of brain disorders [6].

In [7], the authors proposed a fusion approach using deep learning. They used a multi-modal approach for each data stream. And the method uses ensemble fusion on the outputs. EEG waves are written down using the metal plates attached to the scalp. There are several channels that collect brain activity using these electrodes. The physical placement of these electrodes is followed by the International 10–20 systems, as shown in Figure 1.

![Figure 1. International 10–20 system for EEG electrodes.](image)

Although there are many models in earlier papers, the detection of seizures can still be improved. In this research, we build on the previous work and explore the viability of employing machine learning methods for automatically classifying seizures. Epileptic Seizure Recognition is the dataset used in this model. The origin of this dataset can be traced back to the UCI Machine Learning Repository. The model is evaluated on various models such as XGboost, Extra Tree Classifier, Random Forest, etc. Using measures like the F1 score, recall, and precision, the proposed approaches were assessed. The F1 score was considered as the evaluation factor as the dataset was mostly imbalanced. The detailed methodology of the model is explained further in the paper.

2. Materials and Method

2.1. Materials

The points addressed in this section are the dataset used, data pre-processing techniques, and machine learning classification algorithms that were explored.

Before jumping into further analysis and classification, there is one thing that is very crucial in any machine learning research. Good and clean data have no substitute. This research focuses on detecting epileptic seizures from multichannel electroencephalogram signals. Often, EEG signals contain various noises and artifacts, and to detect the occurrence of seizures, the collected data need to be clean, as it reflects brain activity and important electrical impulses [8]. The EEG records the continuous electrical impulses in the form of signals, and it records the abnormal electrical activity that leads to seizures. Therefore, the first step is always to find the correct dataset [9]. In this paper, we studied various
epileptic seizure databases that were released by various research institutes around the world. There are mainly 6 datasets that we have included in our analysis, considering different factors that our research primarily needs. Table 1 contains an overview of the datasets available publicly. Although multiple databases are widely available, we focused on a few parameters that we needed in our study, such as sampling rate. It is one of the necessary parameters when it comes to EEG signals. As the human brain is functioning constantly, the generated voltages are fluctuating continuously, and on the other hand, EEG systems produce samples of data by taking discrete photos of this ongoing activity, much like a camera would do. Various EEG systems have different sampling rates, indicating the number of samples they can capture per second. Like oscillations, the unit of the sampling rate is also samples per second, which is also called Hertz (Hz). For instance, a 250 Hz sampling rate EEG equipment may record 250 samples per second. Hence, the higher the sampling rate is, the better would be the precision of signals. The highest frequency in an EEG signal that may be studied is equal to half of the sampling rate. Only consider frequencies up to \( 256/2 = 128 \) Hz, for example, if your data were collected at 256 Hz. Research that is even more focused suggests staying with frequencies that are roughly one-third the sampling rate (for example, \( 256/3 = 85.3 \) Hz). Since the brain typically generates lower frequencies, you can be certain that the results will be clear even with low EEG sampling rates of roughly 100 Hz (for example, gamma between \([25–80 \text{ Hz}]\) and delta between \([1–4 \text{ Hz}]\)).

Table 1. Comparing Different datasets.

<table>
<thead>
<tr>
<th>Sr No</th>
<th>Dataset Name</th>
<th>Year of Release</th>
<th>Sampling Rate</th>
<th>No. of Channels</th>
<th>Duration of Recordings</th>
<th>Processed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Epileptic EEG Dataset, Mendeley data</td>
<td>2021</td>
<td>500 Hz</td>
<td>21</td>
<td>-</td>
<td>NO (mat and EDF files)</td>
</tr>
<tr>
<td>2.</td>
<td>Pediatric EEG dataset</td>
<td>2021</td>
<td>2000 Hz</td>
<td>52</td>
<td>-</td>
<td>NO (contains EDF files)</td>
</tr>
<tr>
<td>3.</td>
<td>Siena Scalp EEG Database</td>
<td>2020</td>
<td>512 Hz</td>
<td>34</td>
<td>Total 128 h</td>
<td>NO (contains EDF format files)</td>
</tr>
<tr>
<td>4.</td>
<td>The Bonn-Barcelona micro- and macro-EEG database</td>
<td>2020</td>
<td>-</td>
<td>16</td>
<td>32 s each</td>
<td>NO (contain MATLAB-files)</td>
</tr>
</tbody>
</table>

This model utilized the Epileptic Seizure Recognition Data Set sourced from the UCI Machine Learning Repository. In the original dataset, there were 100 files per folder, each of which represented a distinct subject or individual. In each file, there is a recording of brain activity lasting 23.6 s. The associated time series comprises 4097 sampled data points, each indicating the EEG record’s value at a particular moment in time. Therefore, the data consist of a total of 500 individuals, each of whom has 4097 data points collected during 23.5 s. Every data point represents the EEG recording value at a distinct moment, as each 23-unit chunk contains 178 data points per second, randomly distributed within the 4097 data points. Therefore, \( 23 \times 500 = 11,500 \) rows of informational elements make up the data, with the last column serving as the output label. In each row, there are 178 data points, corresponding to a duration of 1 s. The dataset comprises 2300 instances of EEG signals with seizures and 9200 instances of EEG signals without seizures. Due to this imbalanced nature of the dataset, the F1 score, recall, and precision were chosen as the performance metrics. Table 2 gives the description of the dataset.

Table 2. Description of the dataset.

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Class Label</th>
<th>Number of Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epileptic Seizure</td>
<td>1</td>
<td>2300</td>
</tr>
<tr>
<td>Non-Seizure</td>
<td>0</td>
<td>9200</td>
</tr>
</tbody>
</table>
2.2. Pre-Processing

The proposed model’s progression is illustrated in Figure 2. The initial stage involves data pre-processing, wherein standardization is carried out. Subsequently, classification algorithms are employed to distinguish between epileptic seizure and non-seizure conditions.

\[ \text{Standardization} (z) = \frac{(x - \mu)}{\sigma} \]  

\[ \text{Mean} (\mu) = \frac{1}{N} \sum_{i=1}^{N} (x_i) \]  

\[ \text{Standard Deviation} (\sigma) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2} \]

Figures 2. Flowchart of proposed model.

To avoid overfitting, any machine learning model must go through a crucial step where the data are split into two halves [10]. One for testing, the other for training. Hold out cross validation is a technique where the data are divided into a train test set. After trying different combinations of train and test size, this system appeared to exhibit improved performance when utilizing a 70–30 split, meaning 70% of the data were allocated for training and 30% for testing.

The model fitting and model learning functions are not equally affected by variables assessed at different scales, and they may even lead to bias. The idea is to standardize the features individually (mean = 0 and standard deviation = 1).

For classification, we have used algorithms such as k-Nearest Neighbor (KNN), Naive Bayes, Random Forest, Gradient Boost, Extreme Gradient Boost (XGB) and Extra Tree Classifier (ETC). For all these classifiers, a train-test split ratio of 70–30 was chosen.

2.3. Machine Learning Classification Models

2.3.1. K-Nearest Neighbor

KNN is the supervised machine learning algorithm capable of serving purposes in both classification and regression tasks. The outcome of KNN is c class membership. The class to which an object should belong is determined by the majority of votes of k nearest neighbors [11]. The positive integer k is typically small and mostly odd. The KNN model is fitted to the training and validation set with the value of k as 3. The distance metric used...
was Minkowski distance of order 2 and rest parameters as default. The minkowski distance between two points \( X = (x_1, x_2, \ldots, x_n) \) and \( Y = (y_1, y_2, \ldots, y_n) \) is given by:

\[
D = \left( \sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}
\]

As order of two is used in the algorithm, then the Minkowski distance becomes the Euclidean distance.

2.3.2. Naive Bayes

It relies on the principles of Bayes theorem and hence utilizes probability for dependent variables. It is assumed that every independent variable has an independent and equal contribution to the outcome [12]. The parameters of NB were kept default and for better prediction.

\[
P(A|B) = \frac{P(A|B)P(A)}{P(B)}
\]

2.3.3. Random Forest

Random Forest combines the result of decision trees to give better accuracy. Here parallel trees are bagged or aggregated we can say, so it is a bagging or bootstrap aggregation ensemble learning algorithm with base learner as decision trees. It constructs and integrates various decision trees to improve accuracy. The names random and forest refer to the random selection of predictors and the usage of a variety of decision trees in prediction and decision-making. Through its features, random forest aids in preventing overfitting and strengthens the model. The model uses 100 estimators with Gini as the criterion [13].

\[
\text{Gini Impurity} = \sum_{i=1}^{C} f_i (1 - f_i)
\]

where, \( f_i \) = frequency of node, \( C \) = number of distinct labels.

2.3.4. Gradient Boost

It is a widely used boosting algorithm which uses Classification and Regression Trees as its base learner. It consists of \( n \) trees in which the first tree is randomly trained. Then, the next tree is given residual or errors of the first tree with preceding dataset and a similar pattern continues until all \( N \) trees forming ensembles are trained. It trains the base learners(trees) sequentially, that is models are trained one after the other. The model uses 100 boosting rounds. To increase the learning capability of the model, a learning rate of 1.0 was used. These two parameters are the hyperparameters of the gradient boost algorithm on which the performance of the model depends [14]. Other parameters such as max depth and random state were 3 and 69, respectively. This algorithm proved to be one of the best among all in predicting the seizures. This model keeps updating the predictions of previous models in the next model. Let us assume a gradient boost model with \( M \) stages. At each stage \( m \) such that \( 1 \leq m \leq M \), there is an unreliable model \( F_m \), so the model adds some new estimators \( h_m(x) \),

\[
F_{(m+1)}(x_i) = F_m(x_i) + h_m(x_i) = y_i
\]

In this way, the gradient boost model corrects the errors of its predecessor model in the successor.

2.3.5. Extreme Gradient Boost

This is an ensemble modelling technique of boosting type. It is faster than gradient boosting as regularization techniques, that is L1(Lasso) and L2(Ridge) Regularization. Weights are assigned to independent variables and then given to the decision tree which then increases the weight of wrongly assigned variables before going to the next decision
tree to improve accuracy of the model. It minimizes the bias and the penalizing of trees is conducted very cleverly. It has gained much popularity in recent times as it is the choice of many data scientists [15]. This algorithm uses the Newton–Raphson function space unlike gradient descent boosting that uses gradient descent function space.

2.3.6. Extra Tree Classifier

Extra Tree Classifier is pretty like the random forest classifier, with the building of the decision trees in the forest being the sole difference. Multiple de-correlated decision trees are formed using calculations based on Gini index or information gain and entropy. This very random tree classifier is another example of an ensemble learning technique. In this, 100 estimators were used [16]. The criterion was entropy and max features as 1.0. Minimum leaf samples were 3 and minimum sample split was 20. This model does not use bootstrap that means the samples of trees are made without replacement unlike Random Forest which uses bootstrap. Another difference between extra tree classifier and random forest is that extra trees use random splits of nodes unlike random trees which use best splits.

\[ \text{Entropy} (S) = \sum_{i=1}^{c} -p_i \log_2(p_i) \]  

In the above formula, \( c \) stands for unique class labels and \( p_i \) stands for the ratio of rows with output labels.

2.4. Performance Metrics

After analyzing the data, it was found that the data is biased towards the baseline class. Prevalence of seizure class is just 20% in the whole dataset which says that the dataset is mostly imbalanced. Hence it would not be wise to use accuracy as a performance indicator. Therefore, the F1 score is used to evaluate the classification models [17]. The F1 score is a metric that represents the harmonic mean of precision and recall, providing a more balanced summary of the model’s performance.

\[ F1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]  

The F1 score in terms of components of confusion matrix can be given as follows:

\[ F1 \text{ Score} = 2 \times \frac{TP}{TP + \frac{1}{2}(FP + FN)} \]  

3. Results and Discussion

The EEG signals are complex in nature and difficult to extract; they often contain a lot of noise, and the reduced availability of EEG data makes it difficult for researchers to study. Numerous datasets exist in various forms, but extracting the EEG signals is often a complicated task. The training data has been fed to several classifiers for the classification of seizures. Dealing with large amounts of data seemed to be one of the major challenges [18].

To assess the overall performance of classifiers, the UCI dataset is experimented with various classification models. To validate the proposed method, 30% of the dataset is allocated for training, while the remaining 70% is utilized for testing. As the dataset was unbalanced, instead of accuracy, the F1 score was calculated as a performance metric for all of the classifiers. In Figure 3, it is clear that Random Forest (RF) and Extreme Gradient Boost (EGB) achieved the highest F1 score for both training and testing. Other classifiers, such as KNN, Naive Bayes, and Gradient Boost, achieved the F1 score less than Random Forest and ETC. The accuracy, the F1 score, precision and recall of the entire classification are summarized in Table 3. For this binary classification, confusion matrices were plotted for visualization purposes.
Seizure and non-seizure activity were classified using several models, including Random Forest, K-nearest neighbor, naïve bayes, gradient boost, extra tree classifier and extreme gradient boost [20]. To validate the aforementioned models, 30% of the dataset is employed for training, while the remaining 70% is utilized for testing. As the dataset was imbalanced, instead of accuracy, the F1 score was calculated as a performance metric for all classifiers. The results show that Random Forest and Extreme Gradient Boost achieved the highest F1 score of 0.943 and 0.933, respectively. An accuracy of 0.977 was given by Random Forest, which was the highest among all. On an average, 97% accuracy was shown by all the algorithms.

<table>
<thead>
<tr>
<th>Methods</th>
<th>F1 Score</th>
<th>Recall</th>
<th>Specificity</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>0.943</td>
<td>0.928</td>
<td>0.990</td>
<td>0.958</td>
<td>0.977</td>
</tr>
<tr>
<td>XGB</td>
<td>0.933</td>
<td>0.897</td>
<td>0.993</td>
<td>0.972</td>
<td>0.974</td>
</tr>
<tr>
<td>ETC</td>
<td>0.930</td>
<td>0.894</td>
<td>0.993</td>
<td>0.969</td>
<td>0.973</td>
</tr>
<tr>
<td>NB</td>
<td>0.899</td>
<td>0.887</td>
<td>0.979</td>
<td>0.912</td>
<td>0.960</td>
</tr>
<tr>
<td>GB</td>
<td>0.889</td>
<td>0.861</td>
<td>0.981</td>
<td>0.920</td>
<td>0.957</td>
</tr>
<tr>
<td>KNN</td>
<td>0.768</td>
<td>0.625</td>
<td>0.999</td>
<td>0.995</td>
<td>0.924</td>
</tr>
</tbody>
</table>

When examining the output generated by the classifier models based on the F1 score, the lowest F1 score, which is 0.768, was attributed to the KNN model, whereas Gradient Boost reached a higher F1 score of 0.889. On the other hand, Random Forest outperformed all of the models and achieved an F1 score of 0.943. Eventually, the accuracy study of specified models showed that Random Forest performed better than others, with a maximum accuracy of 0.977, whereas Extreme Gradient Boost and Extra Tree Classifier achieved slightly lower accuracy of 0.974 and 0.973, respectively. Other performance metrics, such as recall, precision and specificity, of all of the classifiers are depicted in Table 3.

4. Conclusions

These EEG data are often complex signals containing noise and artifacts. Accurate classification of epileptic seizures is a major challenge, as manual seizure identification is a laborious and challenging endeavor for neurologists [19]. This work aimed to develop a potent machine learning model that could forecast when epileptic seizures will occur. This study utilized the Epileptic Seizure Recognition Data Set sourced from the UCI Machine Learning Repository. Standardization was applied during the data pre-processing phase.
Author Contributions: Conceptualization, V.L.; methodology, C.K. and V.L.; validation, M.K.; investigation, R.M.; resources, S.L.; data curation, S.L.; writing—original draft preparation, C.K. and R.M.; writing—review and editing, M.K. and C.K.; supervision, P.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data used are made available in the present work.

Conflicts of Interest: The authors declare no conflicts of interest.

References


5. Almustafa, K.M. Classification of epileptic seizure dataset using different machine learning algorithms. *Inform. Med. Unlocked* *2020*, 21, 100444. [CrossRef]


11. Guerrero, M.C.; Parada, J.S.; Espitia, H.E. Principal Components Analysis of EEG Signals for Epileptic Patient Identification. *Computation* *2021*, 9, 133. [CrossRef]


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.