APPENDIX A

Minor Project Report

On

An Analysis of Epilepsy Detection and Classification using Machine Learning Techniques

Submitted to Amity University, Uttar Pradesh

in partial fulfilment of the requirement for the award of the degree of

Bachelor of Technology in

Computer Science and Engineering (2020-24) By

Mantra Jain

(A2305220412)

Ansh Srivastav

(A2305220390)

Under the guidance of

Dr Harshit Bhardwaj

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY
AMITY UNIVERSITY NOIDA UTTAR PRADESH

July-October 2023

DECLARATION

We, Mantra Jain and Ansh Srivastav, students of B.Tech (7-CSE-8Y) hereby declare that the project titled “An Analysis of Epilepsy Detection and Classification using Machine Learning Techniques” which is submitted by me to Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, in partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

The Author attests that permission has been obtained for the use of any copyrighted material appearing in the Dissertation / Project report other than brief excerpts requiring only proper acknowledgement in scholarly writing and all such use is acknowledged.

Date:
Mantra Jain
A2305220412
7-CSE-8Y (2020-24)

CERTIFICATE

On the basis of declaration submitted by **Mantra Jain (A2305220412) and Ansh Srivastv (A2305220390)**, student of B.Tech. CSE, I hereby certify that the in-house project titled "**An Analysis of Epilepsy Detection and Classification using Machine Learning Techniques**", submitted to Department of Computer Science & Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology in Computer Science & Engineering, is an original contribution with existing knowledge and faithful record of work carried out by him under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: NOIDA
Date:

Dr. Harshit Bhardwaj
Assistant Professor-II
(Guide)
Department of Computer Science and Engineering
Amity School of Engineering and Technology
Amity University Uttar Pradesh, Noida

An analysis of Epilepsy detection and classification using machine learning techniques

Abstract

Epilepsy, a mental disorder characterized by seizures and uncertainty, remains a significant medical problem. Timely and accurate detection of epilepsy is very important for diagnosis, treatment and patient management. Considering that seizures can occur suddenly and without warning, it is important to have a system that can detect seizures. A comprehensive review of the electroencephalogram (EEG) recording is required to accurately identify these seizures. In recent years, the intersection of machine learning and medicine has shown promise in improving the diagnosis and classification of epilepsy. This summary provides a brief overview of the report on epilepsy diagnosis and classification analysis, which includes various machine learning algorithms such as K-Nearest Neighbor (KNN), Logistic Regression, Naive Bayes, Random Forest, Support Vector Machine (SVM) and Decision Trees. This study provides a brief summary of a report on epilepsy detection and classification through machine learning. This report explores the evolving field of epilepsy diagnosis and reviews the various machine learning algorithms, datasets, and computational techniques currently in use. The overall aim of this report is to demonstrate the potential of machine learning to improve our understanding and management of epilepsy.

Keywords: electroencephalogram, Epilepsy, Seizures

1. Introduction

We discovered at the outset of this research that it would be quite helpful to clarify a few things. To provide them a brief overview of what they will read about in the next chapters, as well as the nature of the examination's subject and the solution's structure. We will concentrate on identifying epileptic seizures in electroencephalogram (EEG) data. All users are welcome to utilize this information, which was gathered at the German university of Bonn. Several well-known machine learning techniques that have been suggested in the literature for comparable tasks will be used in the identification procedure. To evaluate them, seven di↵erent measures will be applied. Python 3.7 was used to implement the whole procedure. The objective is to contrast some of the approaches put out in the literature and expand them from patient-specific to datasets with numerous cases. Epilepsy is a mental disorder characterized by sudden and unpredictable events that a↵ects millions of people worldwide. These seizures are caused by electrical malfunctions in the brain and often present with symptoms that vary in intensity and duration. Accurate and timely diagnosis of epilepsy is important for e↵ective diagnosis, treatment and management. In recent years, the integration of machine learning techniques into medicine has opened new avenues in the analysis, diagnosis and classification of epilepsy. This integration provides a revolutionary way to increase diagnostic accuracy, predict epilepsy events, and tailor personalized treatment

Email addresses: mj.mantrajain@gmail.com (Mantra Jain), srivastavansh253@gmail.com (Ansh Srivastav)

strategies. This comprehensive document covers various areas of machine learning for search and classification and highlights the important role of artificial intelligence (AI) in improving our understanding of complex medical conditions. Although epilepsy diagnosis has historically relied on neurologists' clinical observations, physical examination, and electroencephalography (EEG) data analysis, machine learning algorithms such as nearest neighbor (KNN) and logistic regression promise to provide better and more e↵ective diagnosis. path. - E cient and scalable solution. Machine learning and various algorithms such as KNN and Logistic Regression have the potential to change the entire search and classification landscape. These algorithms use information from large, complex datasets from pattern recognition and data analysis to help us di↵erentiate epilepsy, predict seizure incidence, and develop personalized treatment plans.

In this report we will examine the current state of the art in detecting and classifying epilepsy, paying attention to the di↵erent types of machine learning algorithms used and the data and computational methods used. By examining the strengths and limitations of these algorithms, we aim to better understand their e↵ectiveness in diagnosing epilepsy and predicting seizures.

As we delve deeper into the integration of medicine and technology, we aim to see the potential of machine learning, including algorithms such as KNN and logistic regression, in search and classification. From the content of this report, readers will gain a deeper understanding of the current state of the field, future challenges, and deadlines for machine learning to improve epilepsy management and improve patient care.

2. Related Work

The extension of machine learning in epilepsy-focused sectors, including seizure detection and monitoring, has been the subject of numerous studies, including this paper. By utilizing methods such as multilayer artificial neural networks, support vector machine (SVM), and deep learning, machine learning shows potential in enhancing the ability to handle and evaluate EEG and imaging data that was once considered too complex for experts. Furthermore, this paper [1] supports applying machine learning techniques to optimize medication selection, improve the precision of clinical outcome predictions, and streamline surgical planning. Predictive models produced by machine learning are a source of concern for the authors due to the limited number of validation studies published. It's worth considering the applicability and generalizability of these models in light of this deficiency. Broader datasets that take into account greater diversity are recommended by the authors in order to fill this void. Furthermore, the expected increase in investment in external validation studies to make the application of machine learning in medicine, particularly in epilepsy, more reliable was highlighted.

U. Rajendra Acharya et al. explains that seizures that occur frequently are the hallmark of epilepsy, an electrophysiological brain condition. Epileptic seizures are often detected and studied using an electroencephalogram (EEG) [2] , a technique that tracks and analyzes electrical responses in the brain. Nevertheless, it is frequently challenging to spot minute but significant shifts in the EEG pattern by visual examination, opening up a wide study area for biomedical researchers to create and apply a number of clever algorithms enabling the detection of such modest alterations. The EEG signals are also irregular and unpredictable in character, which adds to the di culty of manually identifying both normal and aberrant

(interictal and ictal) activity. Therefore, a Computer Aided Diagnostic (CAD) method must be created in order to e↵ortlessly di↵erentiate between healthy and sick behaviors utilizing a minimal amount of highly distinguishing classifiers. It has been discovered that nonlinear characteristics can capture complicated physiological processes in the EEG signals, such as sudden shifts and unpredictable activity.

Akut, Rohan is talking about the avoidance of overfitting in modeling with a multiple validation of X folds to train each model. They did 150 epoch training loops and a batch size of 5 and used Batch Normalization to try and mitigate the e↵ect of any covariance variations happening in the model [3]. This modification includes another normalization e↵ect that allows for faster model training with less overfitting probability risk. Particular emphasis is on the Max Pooling Layer using a 1x2 filter which results in output size half of a input size. Its last layer of computation is passed through a SoftMax function in order to classify the data. A dropout layer value of 0.25 was used for more fine-tuning and overfitting prevention. Binary classified results were then compared with pre-existing models based on accuracy, sensitivity, and specificity, which suggests an overarching aim of assessing how well the model performs in small datasets. This procedure can help to get the detailed information about the capability of the model. Particularly, we assessed the validity of the model with respect to a two-level classification utilizing the Bonn dataset. The Bonn dataset is considered a standard benchmark for EEG data, as well as one of the benchmark datasets to evaluate classifier's performance on, which enables us to test our CNN based on a highly reliable source. The CNN model is trained using cross-validation of 10-folds and the optimal batch sizes/epochs that can generate best performance.

Amin et al. points out that reversing the epilepsy diagnosis is complex and requires examining the "unusual" EEG, which can be tricky. One of the primary causes for regular EEGs being mistakenly interpreted as disturbed is the absence of required practical experience in neurology residency programs. Tests, such as the EEG, shouldn't take precedence over medical expertise [4]. Some kinds of seizures may go undiagnosed, and identification of epilepsy may prove di cult. Hypermotor seizures in the frontal lobe may be mistaken for psychogenic episodes. Focal oblivious cognitive seizures in the elderly may be misdiagnosed as dementia, and frontal and temporal lobe epilepsies may present with ictal or interictal psychosis that is misdiagnosed as a main mental disease. Diagnostic mistakes happen often in medicine and a↵ect all disciplines. Both patients and doctors may have severe repercussions. In neurology, mistakes frequently result from a focus on "assessments over the actual clinical situation. Epilepsy is often diagnosed clinically and based on a patient's medical history. Overdiagnosis of epilepsy occurs more frequently than underdiagnosis. Poor medical background and an unusual EEG can result in an incorrect epilepsy diagnosis. Patients who had been diagnosed previously with epilepsy, failed to get better after receiving their first antiepileptic medication as they don't actually have the condition. Most people who receive an incorrect epilepsy diagnosis end up having syncope or psychogenic nonepileptic events.

Amin, Ushtar and Benbadis, Selim R came to the conclusion that Regular EEG has a sensitivity of minimum 80 percent for epilepsy. Except for the neglected issue of over-reading, the specificity for interictal epileptiform emition is strong i.e. grater than 90 percent [5]. When they occur, typically aid in identifying the epilepsy disorder type. One of the independent recurrent indicators is the existence of epileptiform EEG discharges, although this information must be evaluated in a clinical setting. This is a reason to record an EEG since it helps with risk assessment before deciding whether to discontinue ASM (Anti Seizure

Medications) from patients who are seizure-free. For various applications, there are many EEG recording methods. The video-EEG recording of the alleged occurrences is the most conclusive examination to identify the seizure type. When the cause of changed mental state and atypical behaviors is unclear, an EEG in the intensive care unit is a helpful tool for investigating the root cause. They explained how EEG continues to be a crucial test for the diagnosis of epilepsy despite developments in imaging. It can clarify the kind of epilepsy as well as confirm the diagnosis. Depending on the length, the availability of video, and the inpatient or outpatient surgery scenario, there are many distinct forms of EEG recordings, each with advantages and disadvantages. Although interictal epileptiform aberrations are particularly unique to epilepsy, unskilled readers may overinterpret them. EEG has a role in the diagnosis of epilepsy, the decision to stop therapy in seizure-free individuals, and the evaluation of critically sick patients for potential encephalopathies and status epilepticus. EEG results must be somewhat uniform and understandable to the doctor who ordered the EEG.

Bajaj et al. have introduced to an innovative personalisable approach to detecting epileptic seizures with the aid of an individual. It is based on the analysis of EEG signal's analytical intrinsic mode functions (IMF). It [7] is directed towards classifying EEG recordings corresponding to seizure free as well as seizure events, which can be informative in terms of diagnostics and therapy. The method we propose uses the extraction of an instantaneous area using the curve of the analytical IMFs of EEG calculated with a movable window of the minimal width. According to the study, this method demonstrated promising results in detecting focal temporal lobe epilepsy from intracranial EEG signals using rule-based technique. In addition, the authors contrasted their approach against existing detection techniques examined using the same EEG dataset. As demonstrated the proposed approach achieved better detection performance which pave the way for future research in epilepsy diagnosis. Statistical metrics like sensitivity (SEN), specificity (SPE), Positive predictive value (PPV), Negative Predictive Value (NPV) and Error rate detection (ERD) were used to evaluate the performance of the proposed system. The indices were calculated according to correctly and incorrectly detect positive and negative events in total. The performance was good for seizure detection and suggests the feasibility of this method to be deployed in clinics for practical use. Therefore, the proposed work is an additional contribution in the personalized epilepsy detection literature. This highlights the possibility of utilizing instantaneous area of IMFs obtained from EEG signals and opens up an avenue towards better understanding and manipulation of epileptic seizures.

Chakraborti et al. analyses artificial neural networks (ANNs), a popular machine learning technique used to discover knowledge and patterns from exponentially growing data sets. This discussion [8] centers on two specific types of neural networks: back-propagation (perceptron) and back-propagation networks. The learning rule of the former is used for updating the weights and the biases of the network. This network has an output that's controlled by the interaction between a transfer function, the weight, the bias and the input. Then there's the multilayer feed-forward back-propagation network which generalizes the least mean square algorithm. Backward propagation is done for the sensitive parts of the model. They use the back-propagation network to detect the epileptic and non-epileptic signals (Figure 2). Functionality and application aspects of both networks are discussed in detail followed by a comparative study highlighting the broader aspects and benefits of deploying artificial neural networks in the context of machine learning applications.

Chen et al. reviewed how electroencephalogram (EEG) is linked to Epileptic seizures and provided physiologic basis of EEG and intracranial EEG studies. They talked about pointed contoured waveforms or complexes that are di↵erent from background waves and mimic those observed in a part of human people with epileptic diseases are referred to as interictal epileptiform discharges. The most widely characterized interictal epileptiform discharges are spikes and abrupt waves [9]. They explained about rhythmic discharge that typically has to last at least 10 seconds to qualify as an electrographic seizure. BIRDs are described as "Concisely, this refers to short bursts of rhythmic brain activity exceeding 4 Hz, which may appear abruptly and do not match any recognized normal or harmless patterns" Their research frequently identifies interictal or ictal abnormalities, and how EEG is still an essential tool for diagnosis of epilepsy. Yet, epilepsy cannot be ruled out in the lack of interictal epileptiform discharges or ictal symptoms. There are two types of seizures: namely, focal or generalized. Electrographic patterns may di↵er, and ictal activity often develops throughout a seizure. In order to properly diagnose and treat nonconvulsive status epilepticus i.e is a state of continuous seizure activity for at least 30 minutes, with cognitive or behavioral changes, continuous EEG monitoring is crucial. When scalp EEG results are ambiguous, intracranial EEG monitoring is extremely helpful for planning surgery since it frequently provides for earlier seizure identification and higher spatial resolution than scalp recordings.

Guo et al. indicated that Epilepsy a↵ects around 1 percent of the worldwide population. Recurring seizures are epilepsy's primary symptom. The processes behind epileptic diseases can be better understood by carefully analyzing electroencephalogram (EEG) samples [10]. Automatic seizure identification in EEG recordings is crucial because epileptic seizures happen erratically and without warning. An e cient analytic method for signals that are not stationary, such as EEGs, is the wavelet transform (WT). The line length characteristic is a measurement that is responsive to variations in frequency as well as amplitude and depicts shifts in waveform dimensionality. The automated epileptic seizure detection approach described in this study employs line length characteristics based on wavelet transform multiresolution decomposition in conjunction with an artificial neural network (ANN) to categorize the EEG signals about seizure presence or absence. The authors are aware of no other works in the literature that are comparable to this one. The suggested approach was assessed using an established public dataset. The excellent accuracy achieved for three separate categorization issues attested to the method's outstanding performance.

Jaiswal et al. presents detailed comparison of SpPCA (Sparse Principle Component Analysis) against SubXPCA (Sub space Extend Princpale Component Analysis) with SVM (Support vector machine) for classification purpose. It turns out that adding more projection vectors is positively correlated with how well they can predict the given task [11]. In terms of classification accuracy, the paper shows that the SubXPCA method yielded very slightly better results in most cases with respect to SpPCA. A great finding was: As the number of PVs grew larger in SpPCA, its accuracy didn't raise dramatically. On the other hand, SubXPCA had an increasing number of accurate features. In the paper we also included figures showing an out-of-sample performance comparison among PCA + SVM, SpPCA + SVM, and SubXPCA + SVM; these confirmed the best outcomes of subspace XPCA + SVM. The authors end with putting the results into perspective and compare them to what had been found in prior literature. Thusly this document is an incredible expansion to the computational documentation in class calculations issues, especially for Principle Components Analysis, Spare Principle Corresponding Matrix Evaluation and Subspace Stretched out Principle Corresponding

Matrix Assessment. It demonstrates how these algorithms perform di↵erently with varying conditions which is quite an excellent base for future research on these algorithms.

Lahmiri, Salim indicated how epilepsy is becoming more common, and its prevalence is rising. Designing precise computerized procedures for the identification and categorization of electroencephalogram (EEG) data from epileptic patients is therefore very helpful in the diagnostic process. There work aims to propose a machine-learning diagnosis method that can quickly and accurately identify between normal and abnormal EEG data with seizure-free periods using the extended Hurst exponent approaches, fractal features of EEG signals are computed at various scales to better describe their dynamics [12]. Generic Hurst exponent estimations between healthy and epileptic EEG signals with seizures uninterrupted durations are statistically di↵erent, according to parametric and nonparametric statistical tests. Support vector machine classifiers that were trained using extended Hurst exponent estimates. There suggested system has potential and can be expanded for other biomedical applications such as di↵erentiating between normal brain waves and those with intervals of seizure or between epileptic EEG signals with seizure free intervals because this problem is challenging and has not been addressed in the literature.

Lahmiri et al. told that epilepsy is becoming more common, and its prevalence is rising. They designed precise computerized procedures [13] for the identification and categorization of electroencephalogram (EEG) data from epileptic patients, therefore they showed its importance in the diagnostic process. There work proposed an automated diagnosis method that can quickly and accurately identify between normal and abnormal EEG data with seizure-free periods. The system is built on extended pointed exponent estimations at various scales that are used to describe EEG recordings, and it is then based on a support vector machine classifier that will be used for categorization and have various kernels. At last, cross-validating studies using the ten-fold classification was suggested.

Mesraoua et al. indicated how EEG in comparison to the conventional method of eye assessment alone, scalp electroencephalography has the potential to provide additional spatial and temporal information. Fortunately, this information is easier to acquire because to contemporary digital EEG technology and computer-assisted analysis. A potential method to enhance non-invasive EEG localisation in focal epilepsies is to look at the spike voltage topography of interictal spikes [14]. Another additional method for locating the epileptogenic zone in individuals who are candidates for epilepsy surgery is electrical source imaging. The vast amount of data in continuous EEG is streamlined by quantitative EEG by providing a static graphical depiction. These are just a few instances of the wealth of information that scalp EEG recordings may give that goes beyond simple eye assessment.

In dissecting epileptic seizures, this paper “A review on the pattern detection methods for epilepsy seizure detection from EEG signals” o↵ers an outline on the use of Discrete Wavelet Transform (DWT). The early portion of the article regards the struggles involved with raw EEG data due to poor spatial resolution and a low signal-to-noise ratio. It is therefore crucial to initiate pre-processing as a means of enhancing the resolution and ratio of the data. The wavelet transform method is presented as a useful technique based on multi-resolution analysis that addresses both of these problematic elements adeptly [15]. Signals from epileptic seizures can be analyzed more e↵ectively using a technique that captures both low-frequency and high-frequency time information. This technique involves decomposing the signal into various levels using a filter bank, as the authors explain in their discussion on the functionality of the method. The paper delves

into the importance of statistical time-domain features extracted from DWT sub-bands, specifically focusing on the mean absolute value (MAV) and standard deviation. To provide further insight on the e↵ectiveness of DWT in EEG pre-processing, the paper references an existing study from the University of Bonn, Germany, which utilizes DWT to identify epileptic seizures. Overall, the discussion in the paper is well-reasoned and emphasizes the potential benefits of DWT in enhancing signal-to-noise ratios and spatial resolution for analyzing epileptic seizures in EEG data.

Smith SJM says Hans Berger, a German psychiatrist, made the discovery of a person electroencephalogram (EEG) in 1929 [17]. When Gibbs and colleagues in Boston showed 3 per second spike wave discharge in what was then called petit mal epilepsy, its potential uses in epilepsy quickly became evident. Because it is an easy and reasonably priced compared to later treatments, way to show the physiological manifestations of abnormal cortical excitability that underlie epilepsy, EEG persists on playing a crucial role in the detection and treatment of patients with epileptic seizures. EEG assists patients with epilepsy in identifying the kind of seizure and epilepsy syndrome, which helps with antiepileptic drug selection and prognostic prediction. In order to determine is the seizure condition focal or generalized, idiopathic or symptomatic, or a component of a particular epilepsy syndrome, EEG results are important to the multi-axial diagnosis of epilepsy. The logical distinction between partial and generalized seizures/epilepsy types is relevant and helpful in the therapeutic setting. Based on the patient and witness accounts, the doctor will often be able to determine the kind of seizure. However, EEG can assist distinguish between a complex partial seizure with focal IED (Interictal Epileptiform Discharges) and an absence type seizure with generalized IED when the history is uncertain (unobserved ”blackouts” or transient impairment of awareness).

The research article “Automatic epileptic seizure detection in EEG signals using multi-domain feature extraction and nonlinear analysis” [18] describes a method of data analysis that is used to categorize segments, this tool is therefore highly e↵ective. It primarily utilizes the time domain, frequency domain, time-frequency domain, and nonlinear analysis. The outcomes are considered via the perspective of di↵erent classifiers, including K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Naive Bayes (NB), Logistic Regression (LR), and Support Vector Machines (SVM). Each of these models is evaluated via two di↵erent techniques: 5-fold CV and 10-fold CV. In the time domain, the majority of classifiers have a high degree of accuracy, with KNN and SVM having the greatest success. In the domain of frequency, the accuracy is slightly lower, with the LR and SVM models providing the greatest performance. Within the domain of time-frequency, KNN and SVM have a high degree of success, this demonstrates the e↵ectiveness of these classifiers. Overall, the e↵ort suggests that these methods of classification, combined with a wavelet denoising method, are capable of producing accurate and legitimate results in various disciplines.

3. Methodology

3.1. Freely acessible datasets

The utilization of datasets is crucial for data scientists and academics to evaluate the success of the models they have presented. The detection of a tumor should similarly pick up on our brain signals. The most popular way to track brain activity is through EEG recordings. These recordings are crucial for machine learning classifications that investigate novel techniques for detecting tumors in a variety of ways, including early tumor detection, quick tumor detection, patient tumor detection, and tumor localization. Data sets that are accessible to the general public are crucial for analysis, comparison, and inference. We'll go through the well-known datasets frequently utilized in epilepsy in the part after that.

1. **Bonn University—EEG dataset** The BONN EEG Time Series Epilepsy Dataset constitutes an important tool for epilepsy research and neurology. The dataset [6] was developed at the University of Bonn in Germany towards enhancing computational analysis of epileptic seizure and improve its detection. Here are some more detailed aspects of the dataset: Data Source: Two major sources of the dataset are; EEG recordings.
   (a) Epileptic Patients: Epileptic EEG data from people. These recordings are very valuable forunderstanding epileptic seizures because they document the activity at the level of the brain during such events.
   (b) Healthy Individuals: Control: Data from EEG recordings of humans who do not have epilepticseizures.

   Annotations: Annotation has been applied in this dataset, indicating epileptic seizures and other events worthy of note. Such annotations are important for training and testing of automated seizure detection software in EEG data. Contributions: With the introduction of the BONN EEG Time Series Epilepsy Dataset, it is possible to develop computer-aided tools for epilepsy diagnosis and management. This allowed refining the algorithms that had the e↵ect of giving better results when working with other patients having this condition. This dataset consists of 100 single-channel EEG recordings, each lasting 23.6 seconds and sampled at a rate of 173.61 Hz. The spectral bandwidth of the data ranges from 0.5 Hz to 85 Hz, and it was originally obtained using a 128-channel acquisition system. These EEG recordings were extracted from larger multi-channel EEG recordings of five patients and designated as Sets A, B, C, D, and E.
   (a) Sets A and B represent surface EEG recordings during periods of closed and open eyes, respectively, in healthy patients.
   (b) Sets C and D comprise intracranial EEG recordings, with C obtained from a seizure-freezone within an epileptic patient's brain and D from a non-seizure-generating area of the same patient.
   (c) Set E contains intracranial EEG data from an epileptic patient captured during epilepticseizures.

   Each set contains 100 text files, each with 4097 samples representing a single EEG time series in ASCII code format. The data has undergone bandpass filtering with cuto↵ frequencies at 0.53 Hz and 40 Hz. It is noteworthy that this dataset is devoid of artifacts, and thus, no prior preprocessing steps are necessary for classifying healthy (non-epileptic) and unhealthy (epileptic) signals. Strong eye movement artifacts have been removed. This dataset was made publicly available in 2001 and has

been extended as part of the EPILEPSIA project. Indeed, the dataset is very important because it provides an opportunity to conduct further investigations into epilepsy which translates into development of e↵ective computational tools meant

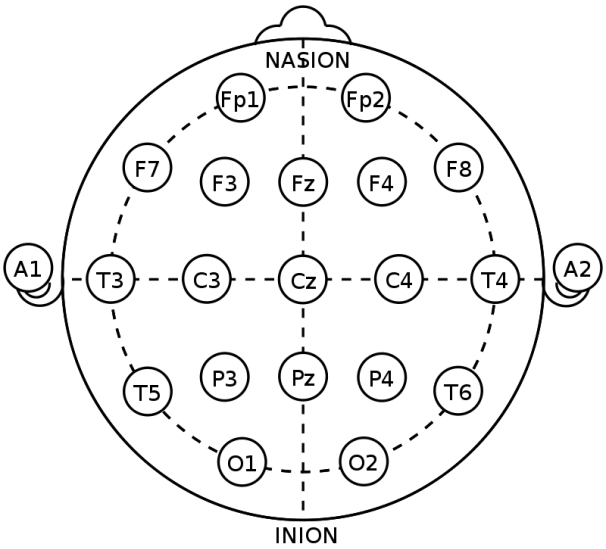

Figure 1: 10 - 20 Electrode System

	Set A	Set B	Set C	Set D	Set E
Subject	Vigorous	Vigorous	Epilepsy	Epilepsy	Epilepsy
Subject Conditions during Readings	Not asleep with eyes opened	Not asleep with eyes closed	Seizure-free (interictal)	Seizure-free (interictal)	Seizure-free (ictal)
Type of Electrode	Surface	Surface	Intracranial	Intracranial	Intracranial
Placement of Electrode	International 10 - 20 System	International 10 - 20 System	International 10 - 20 System	International 10 - 20 System	International 10 - 20 System
Channels	100	100	100	100	100
Duration (Secsond)	23.6	23.6	23.6	23.6	23.6

Table 1: Summary of BONN Dataset

2. **CHB-MIT Scalp EEG Database** The CHB-MIT Scalp EEG Database encompasses data extracted from EEG recordings of 22 young epileptic patients with unpreventable seizures. These [16] were followed by subjects undergoing withdrawal of antiseizure medicines with time spans ranging from two days to determine instance of seizures and operation suitable. 182 seizure startings and closures were recorded. The young epileptic patients were recorded on EEG at the Children's hospital in Boston. Following withdrawal of antiepileptic drugs, the patient had to be watched for several hours or even several days, in order for the type and frequency of epilepsy attacks could be identified and determine eligibility for surgical intervention. These 129 files which contain at least one episode of a seizure recorded in the book Records with Seizures form the subset of the recordings marked by (#) in the table on page two There are 198 seizures in these records as compared to the 182 listed among the first 23 instances in the seizure annotation files in RECORDS-WITH-SEIZURES. Files with names "chbnn-summary.txt" describe the montage used for each recording, as well as the duration in seconds between the commencement and the beginning of each seizure contained in di↵erent "edf(European Data Format)" files.

3.2. Classifiers Theory

1. Decision Tree

Using a decision tree technique, which is commonly used for classification jobs, epilepsy may be classified with EEG. Recursively partitioning the data according to distinct characteristics yields a tree-like structure that is used to identify the class labels of instances. By using the built-in structure as a decision-making tool, this algorithm operates. The decision tree for EEG epilepsy categorization and its mathematical justification are given below. The recursive process entails selecting the best characteristic to split the data at each node in order to construct a decision tree. The criteria measure may be used to reduce the disorder and impurity in the data, which is essential to achieving homogenous subsets. The recursive procedure continues for each subset until all samples in a node have the same class label, at which point it ends when a stopping requirement, such as the maximum depth or the minimum samples per leaf, is satisfied. The epileptic or non-epileptic status of fresh EEG data (X_new) may be determined by moving up the decision tree from the root node to a leaf node. This is the mathematical justification for the classification procedure that follows decision tree construction.

$$Entropy(S) = \sum_{i=1}^{N} p_i log_2 p_i$$

S- set of all instances in the dataset N -

number of distinct class values

p_i = *eventprobability*

2. K-Nearest Neighbors

KNN is k-nearest nearest algorithm otherwise known as a supervised learner with nonparametric characteristics. This involves determining an approximate class or value of a data point by comparing it to other data points. It is applicable in both regression and classification purposes but the general use of clustering similar points makes this tool mainly a classifier. “K” in KNN stands for the number of nearest neighbors, which is taken into account in case of a certain record classification. Choice of ‘K’ depends upon various parameters of the input data. Most of such data generally benefit from a higher ‘k’ value. For a classification technique, it’s usually advisable to use one ‘k’ value for this purpose; besides, some cross-validation methods can help choose the best ‘k’ for a dataset.

$$Euclidean = \sqrt{\sum_{i=1}^{i=k} (x_i \quad y_i)^2}$$

3. Logistic Regression

For binary classification issues, logistic regression is a popular statistical and machine learning technique. Based on a variety of input variables, it estimates the likelihood that an output will fall into one of two classifications. Logistic regression uses the logistic function to limit its output to a range between 0 and 1, reflecting probabilities, in contrast to linear regression, which predicts continuous values. Logistic regression measures the e↵ect of each input feature on the likelihood of class membership by calculating coe cients for each feature. It can be understood, is computationally e↵ective, and acts as a base for more sophisticated methods. Numerous industries, including as healthcare, finance, and marketing, use logistic regression because decision-making and predictive modeling require an understanding of the likelihood of binary outcomes.

4. Naive Bayes

   Na¨ıve Bayes is an easy-to-use probabilistic classification technique for simple applications. The latter relies on the Bayes theorem and provides the probability for a point to belong to a specific class. The naive assumption is that every attribute is independent and makes calculation easier, but this notion does not hold true in real cases. While it is somewhat oversimplified, still many techniques applied in the text classification of spam and opinion mining lose against naive Bayes. This particular algorithm is specifically ideal for high-dimensional datasets that have just enough labeled data. Naive Bayes is of considerable importance because of the capability of handling multiple classifications as well as the ease with which it can be trained and implemented for machine learning and many natural language processing applications.

$$P(C/X) = \frac{(P(X/C) \times P(C))}{P(X)}$$

   P(C/X) = Posterior Probability

   P(X/C) = Likelihood

   P(C) = Class Prior Probability

   P(X) = Predictor Prior Probability

5. Random Forest

   It is an excellent machine learning's ensemble learning technique. This method will result in good and accurate prediction by incorporating several decision making trees in it. The second point is that trees are randomly train-ed one at a time on a specifically chosen portion of the data, based on randomly selected features reducing thus overfitting and improving generalizedness. In this case, the final forecast is created using projections of di↵erent trees. Random Forest is able to provide accurate and highly reliable predictions on many applications which include both classification and regression. One such tool exists for numerous branches of study such as image classification, finances, and healthcare and needs minor changes done in respects with hyperparameters unlike a solitary decision tree.

6. Support Vector Machine

   SVM is a powerful supervised learning technique that can be used in classification and regression problems. Creating an ideal hyperplane that divides the data into two groups of two. SVM has advantage in high-dimensional domain; thus appropriate for tasks like text and image classification. The algorithm operates on such kernel functions as radial basis function (RBF) kernel for handling both linearly and non-linearly classified data. The biggest advantage of SVM is that it is very robust, particularly when working with small and unbalanced datasets; it can also cope with several non-linear relationships in data.

4. Experimental Results

This section details on the experimental results of the applied Machine Learning Algorithms. Table 2 shows training over 70 % of data and testing for 30%, Table 3shows training over 60 % of data and testing for 40%.

4.1. Performance Measures

Here's a brief explanation of the terms in a classification report:

1. Precision: Measures the accuracy of positive predictions.
$$Precision = \frac{TruePositives}{(TruePositives + FalsePositives)}$$

2. Recall: Measures the ability of the model to correctly identify all positive instances.
$$Recall = \frac{TruePositives}{(TruePositives + FalseNegatives)}$$

3. F1-score: A balanced metric that combines both precision and recall. It's the harmonic mean ofprecision and recall and is useful when you want to balance both false positives and false negatives.
$$F1 \quad score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4. Support: The number of samples in each class, which can help you understand the dataset's classdistribution.

5. Accuracy: It measures the overall correctness of predictions made by a model. It is calculated asthe ratio of correctly predicted instances to the total number of instances.
$$AccuracyScore = \frac{TruePositives + TrueNegatives}{TruePositives + FalseNegatives + TrueNegatives + FalsePositives}$$

6. Macro Average (Macro Avg): Macro average is a way to calculate an average of a metric (e.g., precision, recall, F1-score) across multiple classes in a multi-class classification problem. It calculates the metric independently for each class and then takes the unweighted average (arithmetic mean) of these class-wise metric values.
$$MacroAvg = \frac{1}{N}\sum_{i=1}^{N} Metric_i$$
   - N is the total number of classes.
   - Metric is the metric (e.g., precision, recall, F1-score) for class i

7. Weighted Average (Weighted Avg): Unlike macro average, weighted average takes into account theclass distribution. Classes with more instances have a greater influence on the weighted average than classes with fewer instances. It calculates the metric for each class, but the contribution of each class to the weighted average is proportional to the number of instances in that class.
$$WeightedAvg = \frac{1}{N}\sum_{1}^{N}(\frac{Metric_i \times Support_i}{TotalSupport})$$
   - N is the total number of classes.
   - Metric is the metric (e.g., precision, recall, F1-score) for class i.
   - Support is the number of instances in class i.
   - Total Support is the total number of instances in the dataset.

Metric	Classifier	precision	recall	f1-score	support
Baseline	Decision Tree	0.857	1.0	0.923	42.0
	K-Nearest Neighbors	0.737	1.0	0.848	42.0
	Logistic Regression	0.894	1.0	0.944	42.0
	Naive Bayes	0.933	1.0	0.966	42.0
	Random Forest	0.857	1.0	0.923	42.0
	Support Vector Machine	0.84	1.0	0.913	42.0
Seizure	Decision Tree	1.0	0.667	0.8	21.0
	K-Nearest Neighbors	1.0	0.286	0.444	21.0
	Logistic Regression	1.0	0.762	0.865	21.0
	Naive Bayes	1.0	0.857	0.923	21.0
	Random Forest	1.0	0.667	0.8	21.0
	Support Vector Machine	1.0	0.619	0.765	21.0
accuracy	Decision Tree			0.889	0.889
	K-Nearest Neighbors			0.762	0.762
	Logistic Regression			0.921	0.921
	Naive Bayes			0.952	0.952
	Random Forest			0.889	0.889
	Support Vector Machine			0.873	0.873
macro avg	Decision Tree	0.929	0.833	0.862	63.0
	K-Nearest Neighbors	0.868	0.643	0.646	63.0
	Logistic Regression	0.947	0.881	0.904	63.0
	Naive Bayes	0.967	0.929	0.944	63.0
	Random Forest	0.929	0.833	0.862	63.0
	Support Vector Machine	0.92	0.81	0.839	63.0
weighted avg	Decision Tree	0.905	0.889	0.882	63.0
	K-Nearest Neighbors	0.825	0.762	0.714	63.0
	Logistic Regression	0.929	0.921	0.918	63.0
	Naive Bayes	0.956	0.952	0.951	63.0
	Random Forest	0.905	0.889	0.882	63.0
	Support Vector Machine	0.893	0.873	0.864	63.0

Table 2: Seizure Detection on BONN dataset using 70-30 split ML Classifiers

Metric	Classifier	precision	recall	f1-score	support
Baseline	Decision Tree	0.96	1.0	0.98	48.0
	K-Nearest Neighbors	0.889	1.0	0.941	48.0
	Logistic Regression	0.98	1.0	0.99	48.0
	Naive Bayes	1.0	0.979	0.989	48.0
	Random Forest	0.96	1.0	0.98	48.0
	Support Vector Machine	0.96	1.0	0.98	48.0
Seizure	Decision Tree	1.0	0.917	0.957	24.0
	K-Nearest Neighbors	1.0	0.75	0.857	24.0
	Logistic Regression	1.0	0.958	0.979	24.0
	Naive Bayes	0.96	1.0	0.98	24.0
	Random Forest	1.0	0.917	0.957	24.0
	Support Vector Machine	1.0	0.917	0.957	24.0
accuracy	Decision Tree			0.972	0.972
	K-Nearest Neighbors			0.917	0.917
	Logistic Regression			0.986	0.986
	Naive Bayes			0.986	0.986
	Random Forest			0.972	0.972
	Support Vector Machine			0.972	0.972
macro avg	Decision Tree	0.98	0.958	0.968	72.0
	K-Nearest Neighbors	0.944	0.875	0.899	72.0
	Logistic Regression	0.99	0.979	0.984	72.0
	Naive Bayes	0.98	0.99	0.985	72.0
	Random Forest	0.98	0.958	0.968	72.0
	Support Vector Machine	0.98	0.958	0.968	72.0
weighted avg	Decision Tree	0.973	0.972	0.972	72.0
	K-Nearest Neighbors	0.926	0.917	0.913	72.0
	Logistic Regression	0.986	0.986	0.986	72.0
	Naive Bayes	0.987	0.986	0.986	72.0
	Random Forest	0.973	0.972	0.972	72.0
	Support Vector Machine	0.973	0.972	0.972	72.0

Table 3: Seizure Detection on BONN dataset using 60-40 split ML Classifiers

5. Conclusion

The increasing prevalence of epilepsy underscores the growing importance of accurate detection. A significant challenge lies in e↵ectively identifying seizures from extensive datasets. Given the intricate nature of EEG signals within such datasets, machine learning classifiers prove to be a fitting solution for precise seizure detection. However, the critical aspects are the judicious choice of classifiers and features.

This research paper has conducted a comprehensive examination of machine learning methodologies for seizure detection. Consequently, it is concluded that "non-black-box" classifiers, specifically the decision forest (an ensemble of decision trees), exhibit superior e↵ectiveness. This choice is motivated by their ability to generate multiple sensible and explanatory logic rules while maintaining a high prediction accuracy. Moreover, decision forests facilitate the exploration of valuable insights, including seizure localization and the investigation of various seizure types.

In contrast, "black-box" classifiers lack the capacity to generate explicit logic rules, although they can achieve notable predictive accuracy. Regarding feature selection, it is recommended to opt for features that yield logical outcomes. A review of existing literature reveals that features like entropy, line length, energy, skewness and standard deviation can attain a classification accuracy of 100%. It is advisable to avoid irrelevant features, especially as the dimensionality of the data increases, as this can lead to heightened computational costs and the emergence of unintelligible patterns. While employing only one or two features such as line length and energy may reduce the dataset's dimensionality, it may not be conducive to e↵ective knowledge discovery.

In essence, this paper o↵ers fresh insights for data scientists engaged in the domain of epileptic seizure detection through EEG signals. To sum up, the paper centers on the assessment of machine learning classifiers and the selection of appropriate features as key factors in enhancing seizure detection methodologies.

References

[1] Abbasi, B., & Goldenholz, D. M. (2019). Machine learning applications in epilepsy. *Epilepsia*, *60*, 2037–2047.

[2] Acharya, U. R., Vinitha Sree, S., Swapna, G., Martis, R. J., & Suri, J. S. (2013). Automated eeg analysis of epilepsy: A review. *Knowledge-Based Systems*, *45*, 147–165. URL: https://www.sciencedirect.com/science/article/pii/S0950705113000798. doi:https://doi.org/10. 1016/j.knosys.2013.02.014.

[3] Akut, R. (2019). Wavelet based deep learning approach for epilepsy detection. *Health information science and systems*, *7*, 8.

[4] Amin, U., & Benbadis, S. R. (2019). The role of eeg in the erroneous diagnosis of epilepsy. *Journal of Clinical Neurophysiology*, *36*, 294–297.

[5] Amin, U., & Benbadis, S. R. (2019). The role of eeg in the erroneous diagnosis of epilepsy. *Journal of Clinical Neurophysiology*, *36*, 294–297.

[6] Andrzejak, R., Lehnertz, K., Mormann, F., Rieke, C., David, P., & Elger, C. (2002). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical review. E, Statistical, nonlinear, and soft matter physics*, *64*, 061907. doi:10.1103/PhysRevE.64.061907.

[7] Bajaj, V., & Pachori, R. B. (2013). Epileptic seizure detection based on the instantaneous area of analytic intrinsic mode functions of eeg signals. *Biomedical Engineering Letters*, *3*, 17–21.

[8] Chakraborti, S., Choudhary, A., Singh, A., Kumar, R., & Swetapadma, A. (2018). A machine learning based method to detect epilepsy. *International Journal of Information Technology*, *10*, 257–263.

[9] Chen, H., & Koubeissi, M. Z. (2019). Electroencephalography in epilepsy evaluation. *Continuum: lifelong learning in neurology*, *25*, 431–453.

[10] Guo, L., Rivero, D., Dorado, J., Rabunal, J. R., & Pazos, A. (2010). Automatic epileptic seizure detection in eegs based on line length feature and artificial neural networks. *Journal of neuroscience methods*, *191*, 101–109.

[11] Jaiswal, A. K., & Banka, H. (2018). Epileptic seizure detection in eeg signal using machine learning techniques. *Australasian physical & engineering sciences in medicine*, *41*, 81–94.

[12] Lahmiri, S. (2018). An accurate system to distinguish between normal and abnormal electroencephalogram records with epileptic seizure free intervals. *Biomedical Signal Processing and Control*, *40*, 312–317.

[13] Lahmiri, S., & Shmuel, A. (2018). Accurate classification of seizure and seizure-free intervals of intracranial eeg signals from epileptic patients. *IEEE Transactions on Instrumentation and Measurement*, *68*, 791–796.

[14] Mesraoua, B., Deleu, D., Al Hail, H., Melikyan, G., Boon, P., Haider, H. A., & Asadi-Pooya, A. A. (2019). Electroencephalography in epilepsy: look for what could be beyond the visual inspection. *Neurological Sciences*, *40*, 2287–2291.

[15] Sharmila, A., & Geethanjali, P. (2019). A review on the pattern detection methods for epilepsy seizure detection from eeg signals. *Biomedical Engineering / Biomedizinische Technik*, *64*, 507–

517. URL: https://doi.org/10.1515/bmt-2017-0233. doi:doi:10.1515/bmt-2017-0233.

[16] Shoeb, A. (2000). Chb-mit scalp eeg database. URL: https://doi.org/10.13026/C2K01R.

[17] Smith, S. J. (2005). Eeg in the diagnosis, classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry*, *76*, ii2–ii7.

[18] Wang, L., Xue, W., Li, Y., Luo, M., Huang, J., Cui, W., & Huang, C. (2017). Automatic epileptic seizure detection in eeg signals using multi-domain feature extraction and nonlinear analysis. *Entropy*, *19*, 222.