Epileptic Seizures Detection from EEG Recordings Based on a Hybrid System of Gaussian Mixture Model and Random Forest Classifier

Garineh S. Ohannesian and Esraa J. Harfash
E-mail: ohannesiangarena@gmail.com and esra.harfash@uobasrah.edu.iq
College of Computer Science & Information Technology, University of Basrah, Iraq

Keywords: epileptic seizures, Electroencephalogram (EEG), Discrete Wavelet Transform (DWT), random forest, Principle Component Analysis (PCA), Gaussian Mixture Model (GMM)

Received: May 27, 2022

Epilepsy is the most common neurological disease defined as a central nervous system disorder that is characterized by recurrent seizures. While electroencephalography (EEG) is an essential tool for monitoring epilepsy patients' brain activity and diagnosing epilepsy, Visual detection of the EEG signal to identify epileptic seizures is a time-consuming approach that might result in human error. Therefore, an early and precise epilepsy diagnosis is critical to reducing the risk of future seizures. This paper aims to increase epileptic seizure detection accuracy in a balanced dataset while reducing the execution time. To address this, we proposed a hybrid system of supervised and unsupervised machine learning algorithms to construct a computationally efficient and scalable model for the early detection of epileptic seizures from two-class EEG datasets. First, Discrete Wavelet Transform (DWT) was applied to the EEG signal to decompose it into frequency sub-bands. Then these EEG extracted features were fed into the Gaussian Mixture Model (GMM) for partitioning these features into two clusters: epilepsy or not. Lastly, the clusters' output was evaluated with the random forest classifier. In addition, Principal Component Analysis (PCA) was used to reduce the EEG features and to reduce further the features obtained after conducting DWT on the EEG signal to determine the impacts of dimension reduction on this system performance. The experimental results show that the highest accuracy was achieved by the hybrid system of GMM with random forest with DWT features with an accuracy of 93.62 %.

Povzetek: Razvita je bila izvirna metoda strojnega učenja za zaznavanje epileptičnih napadov iz EEG signalov.

1 Introduction

Epilepsy is one of the most severe neurological diseases that affect people's lives[1]. It is a long-term brain disease marked by an aberrant nervous system imbalance produced by the sudden, repeated discharge of the total neuron population from the brain, which leads to recurrent seizures [2]. According to the World Health Organization (WHO), around 50 million individuals suffer from epilepsy globally. Many patients are children and seniors aged 65 to 70 [3]. Although the exact source of this disease is unknown, the majority of epilepsy seizures may be managed medically. Antiepileptic medications might cure just two-thirds of total epilepsy patients, while surgical procedures could help 7-8% of patients. Overall, 25% of people with epilepsy suffer from a lack of possible treatments [4].

Epileptic seizures are characterized by many signs and symptoms, including loss of awareness and consciousness, jerking movements, strange behavior, and disorientation. These symptoms may lead to severe injuries, such as falling and biting one's tongue, and sometimes death [5]. The chance of fatality will be reduced if seizures are managed, and medical treatment is provided when seizures occur [6].

One of the most important tools for detecting epileptic seizures is the electroencephalogram (EEG) test, which aids in the early detection, treatment and soothing of patients [7]. EEG is a clinical method for imaging the human brain while the brain is engaged in cognitive activity. Electrodes are placed on the patient's scalp to record the patient's EEG. The electro-activity produced by the brain along the scalp may then be recorded [3]. EEG recordings provide a large quantity of multichannel EEG signal data that is very complex in nature, including non-stationary, chaos, and aperiodicity. To date, specialists or physicians have mostly used visual analysis to discover and comprehend abnormalities in the brain and how they spread.

Visual labeling of EEG recordings by human specialists to discover evidence of epilepsy is not a suitable process for a trustworthy diagnosis and interpretation since such analysis is time-consuming, expensive, onerous, and vulnerable to mistakes and prejudice. As a result, one of the foremost biomedical research challenges is determining how to classify time-varying EEG data as correctly as possible to aid in the diagnosis of epileptic seizures [8].

Therefore, the development of an automated, computer-aided approach to epilepsy diagnosis is critical [9]. As a result, several approaches for detecting epileptic seizures using EEG recordings have been developed, and Machine-learning algorithms were used for this task,

including gathering Electroencephalography (EEG) signals, preprocessing, feature extraction from the data, and ultimately classification of epileptic seizures [10]. In recent years, researchers have attempted to discover a more effective solution in the machine-learning field to increase prediction performance [4].

Inspired by this, this paper proposed a new hybrid system of supervised and unsupervised machine learning algorithms. The main aim of this study is to improve the accuracy of epileptic seizure detection in a balanced dataset and reduce the execution time by applying the DWT as a feature extraction technique to extract the most important features in EEG signals. Then to choose the most important features, the number of derived features is reduced using the PCA technique. The selected features are clustered with GMM into two clusters. Finally, the output is evaluated using the Random Forest classifier to identify whether it is an epileptic seizure or not. The proposed method was tested by an Epileptic EEG dataset collected by the Bonn University in Germany.

To the best of our knowledge, so far no study has used GMM for clustering as a hybrid with the random forest classifier for epileptic seizure detection. In addition, for the first time in this paper a downsampling technique was performed on this EEG dataset to balance the dataset, make the work process more challenging and the results more realistic.

This paper is organized into several sections as follows: section 2 provides some related works. Section 3 provides a brief explanation of the theoretical background of some machine learning algorithms and techniques that were relied upon in this study. Section 4 describes the methodology of the proposed system. In section 5, the experimental results of the proposed system are presented. The experimental results are discussed in section 6. Section 7 concludes with the conclusions and future work.

2 Literature review

For epilepsy detection, numerous studies have focused on EEG signal classification. This section provides several recent studies on epileptic seizure detection using EEG data. Table 1 illustrates the related works on the detection of epileptic seizures using various methods.

The authors in [11] utilized the Boon university EEG dataset to propose a technique for diagnosing epileptic patients' from EEG signals. First, they used expectation-maximization features to reduce the dimensions of the EEG dataset. These reduced characteristics were then fed into five classifiers for epilepsy classifications including: nonlinear models like the Gaussian mixture model (GMM), logistic regression, firefly algorithms, and hybrid models like cuckoo search with GMM and firefly algorithm with GMM. The hybrid classifier combining Cuckoo search with GMM got the greatest accuracy of 92.19 %, with a lower error rate than the other four classifiers according to this research. While in [12], they suggested a technique for detecting epileptic seizures in long-term EEG recordings. The EEG data were acquired from Adnan Menderes University's Department of Neurology and Clinical Neurophysiology's EEG laboratory. The detection was performed with two classification techniques, support vector machine (SVM) and linear discriminant analysis (LDA), and their results were compared. First, they decomposed the EEG dataset into multiple frequency sub-bands using the DWT. The extracted characteristics are either immediately fed into the classification algorithms or tested using two dimension reduction approaches, PCA and ICA. The experimental results show that while employing the SVM with radial basis function (RBF) kernel without dimension reduction to categorize EEG signals as normal or epileptic, they achieved the greatest accuracy rate of 88.9%. The efficacy of the KNN classifier and K-means clustering to identify epilepsy risk levels from EEG data is investigated by the authors in [13]. The objective was to develop a classification algorithm with a high-performance index, low false alarm rate, and low missed classification rate. The EEG recordings of twenty individuals are analyzed in this research. To begin, Detrend analysis is performed to identify data nonlinearity. Second, the power spectral density is calculated, and the EEG dimensionality is reduced. Finally, the data is classified as they used the KNN classifier and K-means clustering. According to this research, the KNN classifier and K-means clustering have performance indices of 78.31 % and 93.02%, respectively. The KNN classifier produced a lousy value of 18.02, but K-means clustering produced a high-quality rating of 22.37 with a false alarm rate of 0%. The authors of [14] employed a variety of dimension reduction approaches to decrease the EEG features, including Singular Value Decomposition (SVD), Principal Component Analysis (PCA), Independent Component Analysis (ICA), Fast ICA, and Linear Discriminant Analysis (LDA). The dimensionally reduced EEG characteristics are then supplied into a hybrid classifier called the Artificial Bee Colony-Particle Swarm Optimization (ABC-PSO) Classifier, which uses EEG data to categorize epilepsy risk levels. In this study, they used an EEG dataset of twenty epileptic patients who were receiving epilepsy medication. The Fast ICA with ABC-PSO Classifier produced the best results, with an accuracy of 97.42%, the greatest quality value of 22.76, and a time delay of roughly 2.9 seconds, according to the trial data. Different classifiers were used in [5] to classify the Epileptic Seizure dataset. With 97.08%, ROC = 0.996, and RMSE = 0.1527, the Random Forest classifier outperformed the K-Nearest Neighbor (K-NN), Nave Bayes, Logistic Regression, Decision Tree (D.T.), Random Tree, J48, and Stochastic Gradient Descent (S.G.D.) classifiers. Also in this study, Sensitivity analysis was done on several of these classifiers to see how well they performed in classifying the Epileptic Seizure dataset when some of their parameters were changed. After that, a dataset prediction was made using feature selection based on attribute variance. In [15], they demonstrate the use of wavelet transform (WT) for feature extraction of EEG data, using Artificial Neural Network (ANN) and Support Vector Machine (SVM) as classifiers. The EEG signal is decomposed using the Daubechies wavelet for feature extraction. The experimental results of this study show that the ANN presents the best performance with an accuracy of 96.00%.

Table 1: A summary of the above previous researches**.**

Research	Year	Dimension Reduction technique	Feature Extraction technique	Classifiers for the diagnosis of epilepsy	Best Accuracy
[11]	2022	Expectation maximization	---	GMM, logistic regression, firefly, and hybrid model such as cuckoo search with GMM and firefly with the GMM.	92.19%
[12]	2018	PCA and ICA	DWT	Support Vector Machine (SVM) and linear discriminant analysis (LDA).	88.9%
[13]	2016	---	---	KNN classifier and K-means clustering.	78.31% 93.02%
[14]	2018	SVD, PCA, ICA, Fast ICA and LDA	---	Artificial Bee Colony-Particle Swarm Optimization (ABC-PSO).	97.42%
[15]	2017	---	Wavelet Transform (WT)	Artificial Neural Network (ANN) and Support Vector Machine (SVM).	96.00%

3 Theoretical background

The purpose of this section is to provide the theoretical background and processes necessary to comprehend the approaches employed in the next section.

3.1 Feature extraction technique

Feature extraction techniques reduce the amount of data that must be processed while still properly and thoroughly characterizing the original dataset by selecting and/or combining variables into meaningful features [10]. This study presents one feature extraction technique applied on the EEG dataset, namely Wavelet Transform (WT).

3.1.1 Wavelet Transform (WT)

Jean Morlet, a French geophysicist, introduced WT in 1982 [16], and it compresses the time-varying biomedical signal, which consists of numerous data points, into a limited number of parameters that characterize the signal. Because the EEG signal is nonstationary, time-frequency domain approaches such as Wavelet Transform (WT), which is a spectral estimating technique in which every general function may be described as an infinite sequence of wavelets, are the best way to extract features from raw data. Since WT allows for variable-sized windows, it offers additional flexibility in signal time-frequency representation. This approach is designed to deal with nonstationary signals like EEG. The original EEG signal is represented in the WT technique by wavelets, which are secure and simple building blocks. Through translation and dilation, or (shifting) and (compression and stretching) operations along the time axis, the mother wavelet generates these wavelets as part of derived functions. The WT may be classified into two categories: continuous and discrete [17]. The Discrete Wavelet Transform (DWT) will be explained since it was used in this study.

3.1.2 Discrete Wavelet Transformation (DWT)

Mallat developed the DWT, which was specified using multi-scale feature representation. Each DWT scale represents a unique set of brain signals. Convolution is a two-function multiplication technique that employs the low-pass or high-pass filter coefficients, which are subsequently processed by downsampling. Downsampling is the process of halving the size of a sample signal (reduction). Wavelets have two sorts of signals: approximation and detail. A signal acquired via the convolution process of the original signal to the low-pass filter is an approximation coefficient. In contrast, a signal obtained through the convolution process of the original signal to the high-pass filter is a detail coefficient [18, 19]. Figure 1 shows DWT decomposition for EEG data at various levels.

DWT can also be utilized for signal noise reduction, preprocessing, and feature extraction. There are several DWT types grouped into families based on frequency components; these are: The Discrete Meyer (dmey), Reverse biorthogonal (rbio), Daubechies (db), Coiflets (coif), Symlets (sym), and Haar, which are mathematical and statistical functions. DWT performance is influenced

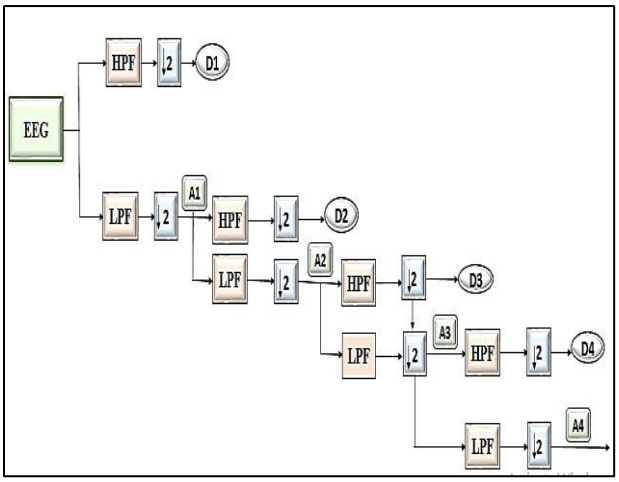

Figure 1: Four-level EEG signal decomposition.

by four major factors: DWT coefficient characteristic, mother wavelet, frequency band, and decomposition level [10].

3.2 Dimensionality reduction

The feature space of datasets may be rather extensive, with thousands of measurements per sample, which will make the analysis of these features very challenging. The analysis of high-dimensional datasets can be related to a phenomenon known as the "curse of dimensionality." It is worth noting that the "curse of dimensionality" makes most data analysis methodologies, particularly machine learning, challenging. Therefore, dropping redundant features might increase the model's performance and convergence time [20]. There are several algorithms for dimensionality reduction, but only the PCA will be explained, which is implemented in this study.

3.2.1 Principle Component Analysis (PCA)

Principal component analysis (PCA) is an unsupervised linear technique. It is a typical data reduction approach used in statistical pattern identification and signal processing [21]. PCA's goal is to decrease the dimensionality of a dataset that contains many correlated variables while maintaining as much variation as possible in the dataset. Hence, this approach creates a collection of uncorrelated features called "Principal Components (PCs)" to reduce the original data's characteristics. The components are defined as the components that cover the largest variation in the dataset and give statistically significant information about the original data. Where the first PC has the greatest variation and the second PC contains the second most variance, and so on [22, 23]. The other benefit of PCA is not losing important information by identifying the patterns by reducing the number of dimensions [21]. Assume a dataset vector $X = x^1, x^2, \dots\dots\dots x^n$ contains n-dimension inputs. Using PCA, the n-dimensional data space will be reduced to a d-dimensional space Y ($d \leq n$). The following are the steps for implementing this algorithm [24].

1. Compute the mean of each vector.

$$\mathbf{X} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{k=1}}^{\mathbf{n}} \mathbf{x^k} \quad (1)$$

Calculate the covariance matrix.

$$\mathbf{C} = \frac{\mathbf{1}}{\mathbf{n-1}} \sum_{\mathbf{k=1}}^{\mathbf{n}} (\mathbf{x^k - X})(\mathbf{x^k - X})^{\mathbf{T}} \quad (2)$$

2. Eigenvalues and Eigenvectors are computed by applying the eigenvalues decomposition to the **C**. The Eigenvectors should be sorted based on their eigenvalues in decreasing order to generate the matrix $\boldsymbol{S(d \times n)}$, which transforms the original n-dimensional space (X) into a new d-dimensional space (Y).
3. Finally, using $\boldsymbol{S^T}$, transform $\boldsymbol{x^k}$ to get the new subspace by calculating:

$$\boldsymbol{Y = S^T x^k} \quad \textbf{For each} \quad \boldsymbol{x^1, x^2 \dots\dots, x^n} \quad (3)$$

3.3 Clustering method

Clustering is a fundamental unsupervised learning technique. It is a method for grouping data points into distinct clusters based on similarity measures. Such that data points in one cluster are similar but distinct from data points in other clusters. Several data clustering techniques are available [25]. Among these techniques, Gaussian Mixture Model (GMM) was used in this study, which will be explained below.

3.3.1 Gaussian Mixture Model (GMM)

Mixture models use a probabilistic form of "soft clustering." Data points in every cluster reflect samples from some kind of probability distribution in a d-dimensional spatial space.

A GMM is an unsupervised clustering algorithm that uses probability density estimates to produce "ellipsoidal-shaped clusters". Every data point to be clustered is taken from a mixture of Gaussian distributions with unknown parameters, according to Gaussian Mixture Model. So, in order to calculate the values of these unknown parameters and to build the distinct clusters, a learning technique is used. The Gaussian distribution, commonly known as the Normal distribution, is a continuous probability distribution that is defined by the following Equation:

$$\boldsymbol{N(X|\mu,\Sigma)} = \frac{\mathbf{1}}{\boldsymbol{(2\pi)^{\frac{D}{2}}\sqrt{|\Sigma|}}} \mathbf{exp}\left\{-\frac{\mathbf{(X-\mu)^T\Sigma^{-1}(X-\mu)}}{\mathbf{2}}\right\} \quad (4)$$

Where, $(\boldsymbol{\Sigma})$ indicates the Gaussian representation were it is a $(D \times D)$ covariance matrix, and its determinant is indicated by $(|\Sigma|)$. Whereas (μ) is a $\mathrm{D}-$dimensional mean vector.

A GMM is a linear combination of the basic Gaussian probability distribution, and it is represented by the following equation:

$$\mathbf{p(X)} = \sum_{\mathbf{k=1}}^{\mathbf{K}} \boldsymbol{\pi_k}\, \boldsymbol{N(X|\mu_k,\Sigma_k)} \quad (5)$$

From the above equation, $(\boldsymbol{K})$ is the number of components in the mixture model and $(\boldsymbol{\pi_k})$ is called the mixing coefficient, and $\boldsymbol{N(X|\mu_k,\Sigma_k)}$ is the Gaussian density, is known as a component of the mixture model.

Gaussian distribution with covariance $\boldsymbol{\Sigma_k}$, mean $\boldsymbol{\mu_k}$, and the mixing coefficient $\boldsymbol{\pi_k}$ is used to explain each component $(\boldsymbol{K})$ [26, 27].

As mentioned above, to find the parameters of the Gaussian distribution for each cluster or to build the distinct clusters, the GMM uses a learning technique (also known as an optimization technique), called the expectation-maximization (EM) technique. It determines the statistical model's maximum likelihood parameters. The unknown model parameters are estimated repeatedly in two steps. First is the E (expectation) step, which involves calculating the posterior distribution of the latent variables using the present model parameters. Data points are fractionally distributed across clusters based on this value. Second, the M (maximization) step, which determines the fractional assignment by recalculating the model parameters using the maximum likelihood rule [28]. The EM algorithm is shown in Figure 2.

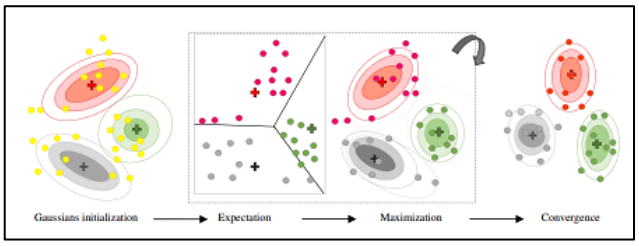

Figure 2: shows the expectation-maximization approach for Gaussian mixtures [28].

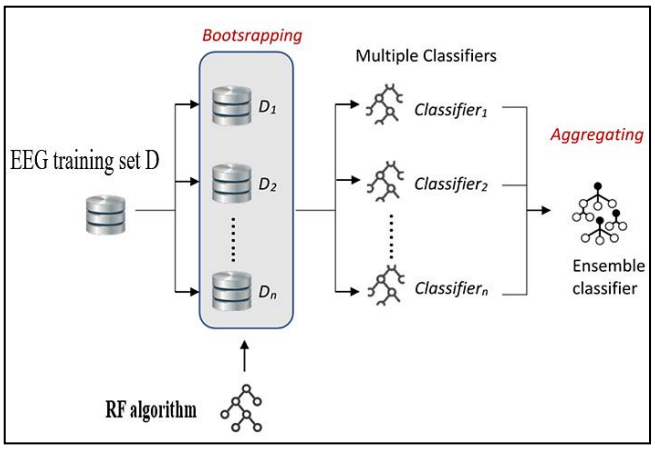

Figure 3: Illustrates the random forest bagging strategy decision process [35].

3.4 Classification algorithm

The goal of classification algorithms is to anticipate discrete outcomes. Classification algorithms are recommended if the data can be categorized or labeled into specific classes. Classification models are used to classify input data, with output values or the goal (Y) being categorical. Example of classification used to determine whether or not a patient has epilepsy. There are many supervised machine-learning algorithms used in medicine classification [29]. The Random Forest algorithm is described below since it is implemented in this study.

3.4.1 Random Forest algorithm

Random forest (RF) is a robust supervised method that performs classification and regression problems [30]. Leo Breiman of the University of California presented it originally in 2001 [31]. This algorithm, also known as Random Decision Forests, is a machine learning approach that uses ensemble learning. It is a "forest" that made up of numerous independent and unpruned decision trees that aggregate the classification results of distinct trees. It is also referred to as a bagging-type ensemble classifier. Combining a few decision trees minimizes the likelihood of overfitting by reducing the variance and bias [32, 33]. The Random Forest algorithm classifies data using a voting mechanism that incorporates the outcomes of individual trees. Direct voting counts the number of trees that have a certain characteristic categorized under a specific class [34]. It is worth mentioning that the RF classifier has been extensively employed in a variety of medical research studies [32]. Figure 3 depicts the main structure of the Random forest bagging strategy using the EEG training data.

The following are the major steps of this algorithm:

1. The EEG training dataset D is segmented into n datasets D1, D2, ... , Dn by using the bootstrap approach.
2. These various datasets are used to train the random forest algorithm.
3. An unpruned classification tree is built for each bootstrap sample.
4. After that, many classifiers (multiple decision trees) were created (Classifier1, Classifier2,..., Classifiern).
5. The ensemble classifier is created by combining the predictions of many classifiers (majority votes for classification) [35].

3.5 Evaluation metrics

Classification model performance is evaluated using unseen data (testing data) by the following metrics.

3.5.1 Confusion Matrix

The confusion matrix is a table that shows the classification results in detail, including whether they were correctly or incorrectly classified. A (2*2) matrix is used for binary classification [36]. Table 2 shows a confusion matrix for binary classification.

- **True Positive (TP):** The model properly recognized positive samples.
- **False Negative (FN):** A positive sample that the model incorrectly classifies.
- **False Positive (FP):** A negative sample that the model incorrectly classifies.
- **True Negative (TN):** The model properly categorized negative samples.

Table 2: Illustrates the confusion matrix.

Confusion Matrix		Actual Class	
		Positive (p)	Negative (N)
Predicted Class	Positive (p)	True Positive (TP)	False Positive (FP)
	Negative (N)	False Negative (FN)	True Negative (TN)

3.5.2 Accuracy

It is the most extensively used metric. It is the proportion of properly categorized samples to the total number of samples for a particular test data set [36], and it is denoted mathematically as:

$$\textbf{Accuracy} = \frac{\textbf{TP+TN}}{\textbf{TP+TN+FP+FN}} \quad (6)$$

3.5.3 Precision

It calculates the ratio of all ‘‘correctly detected items’’ to all ‘‘actually detected items’’ [36]. It is denoted mathematically as:

$$\textbf{Precision} = \frac{\textbf{TP}}{(\textbf{TP+FP})} \quad (7)$$

3.5.4 Recall

The ratio of accurately predicted positive values to the total number of positive values in the dataset is also known as “true positive rate (TPR)” [20]. It is denoted mathematically as:

$$\textbf{Recall} = \frac{\textbf{TP}}{(\textbf{TP+FN})} \quad (8)$$

3.5.5 F1-score

It calculates the harmonic mean of the precision and the recall [36], [37]. It is denoted mathematically as [38]:

$$\textbf{F1} = \textbf{2} * \frac{\textbf{1}}{\frac{\textbf{1}}{\textbf{precision}} + \frac{\textbf{1}}{\textbf{recall}}} \quad (9)$$

4 The proposed system

In this section, the architecture of the proposed system is illustrated, and Figure 4 shows the basic steps for building this system and each part of the scheme will be explained in detail in the following paragraphs.

4.1 Data collection stage

In this study, the experiments were conducted using an EEG epileptic seizure dataset gathered from Bonn University in Germany. There are 500 patients in the dataset, with 4097 electroencephalograms (EEG) measurements collected during 23.5 seconds. The 4097 data points are divided into 23 chunks, each containing 178-voltage signals equivalent to one second of brain activity. As a result, there are 11500 instances in this multivariate time series dataset, each with a brain activity label for 178 features. The last column denotes the label y, which has five classes: 1, 2, 3, 4, and 5. Table 3 contains information about the target classes.

Table 3: Information about the target class.

Target class	Explanation	Number of cases
1	Recording of seizure activity.	2300
2	EEG signal from the tumor region has been recorded.	2300
3	The E.E.G. activity was measured in a healthy brain area.	2300
4	Eyes closed.	2300
5	Eyes open.	2300

4.2 Preprocessing stage

As shown in Figure 4, there are four primary steps in this stage, which are explained as follows:

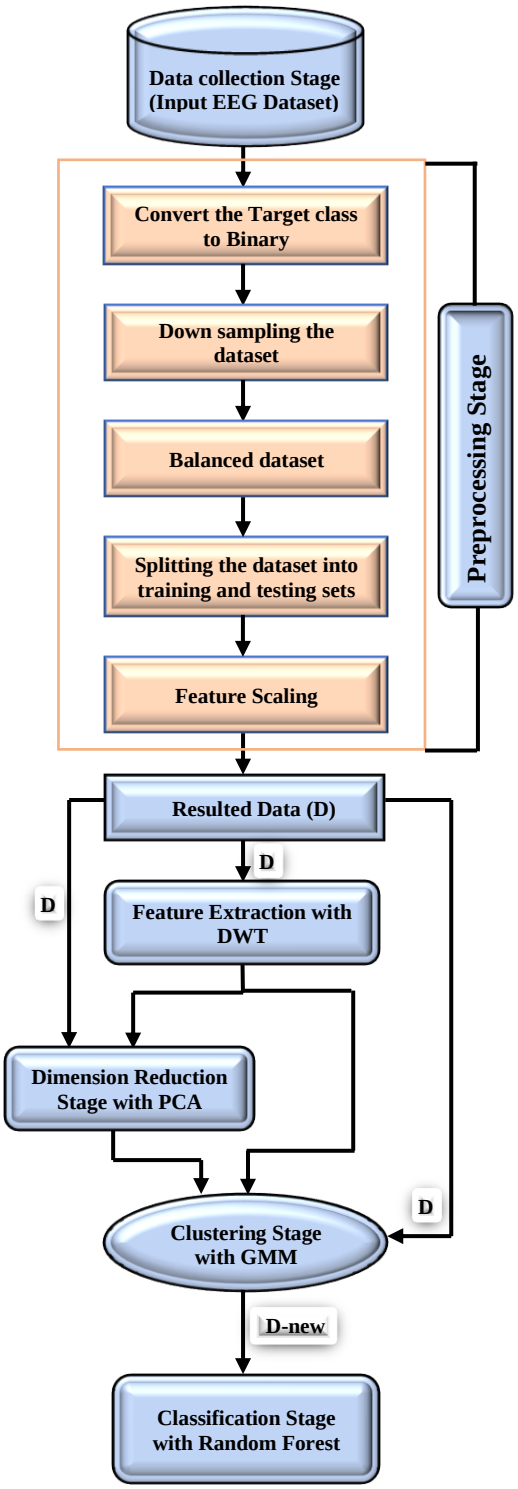

Figure 4: Shows the General Proposed System For Epileptic Detection.

4.2.1 Convert the target class (y)

Since this study is for a binary classification assignment, the first step is to transform the target class from five classes to a binary class. Therefore, when the target class value is larger than 1, all of these values will be set to 0. As a result, Class 1 represents patients with epileptic seizures, while Class 0 represents patients who do not have

epileptic seizures, with a total of 2300 and 9200 individuals, respectively.

4.2.2 Down sampling the dataset

The second step in this stage is to balance the EEG dataset by using the down-sampling approach to make the work process more challenging and the results more realistic. As a result, the dataset has been adjusted so that, the number of patients with epilepsy and without epilepsy equals 2300 the total is 4600 patients. To the best of our knowledge, this is the first study that applies the downsampling technique to this EEG dataset to make it balanced.

4.2.3 Dataset splitting

In this step, the EEG dataset is split into two independent sets; the testing set and the training set. The selected splitting ratio is 70% for training the model and 30% for testing the model. Therefore, total training data is 3220, and total testing data is 1380.

4.2.4 Feature scaling

The final step in the preprocessing stage is applying the feature scaling procedure. The reason for applying this technique is that the EEG dataset contains variables with highly varied scales. Therefore, the dataset should be scaled to transform the feature vectors into a format that machine-learning algorithms can understand. The standard-scaler was utilized in this study as it converts a dataset into a distribution with a mean of 0 and a standard deviation of 1. Equation 6 is used to represent it mathematically.

$$\mathbf{Z} = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \qquad (\mathbf{10})$$

Where, $\boldsymbol{\mu}$ is the mean, $\boldsymbol{\sigma}$ is a standard deviation, $\boldsymbol{x}$ is an original value.

The resulted data (D) from the preprocessing stage has three paths:

1. Enter directly to the feature extraction stage.
2. Or enter directly to the clustering stage.
3. Alternatively, enter the dimension reduction stage.

4.3 Feature extraction stage

In this study, the Discrete Wavelet Transformation (DWT) was used for decomposing the EEG signal into several frequency sub-bands. The EEG signal is decomposed by using the Haar wavelet. The number of decomposition levels is selected to be three. Therefore, the EEG signal is decomposed into cD1, cD2, and cD3 details and cA3 coefficients are chosen as wavelet features with a length of 23. Figure 5 illustrates the decomposition process. In addition, Figure 6 shows the details and approximation coefficients at level three with 23 dimensions.

We utilized DWT to extract the EEG features since it is a highly effective technique compared to other techniques in terms of accuracy of feature extraction to assure the efficacy of the following processes. The output of this step (the extracted features) has two paths:

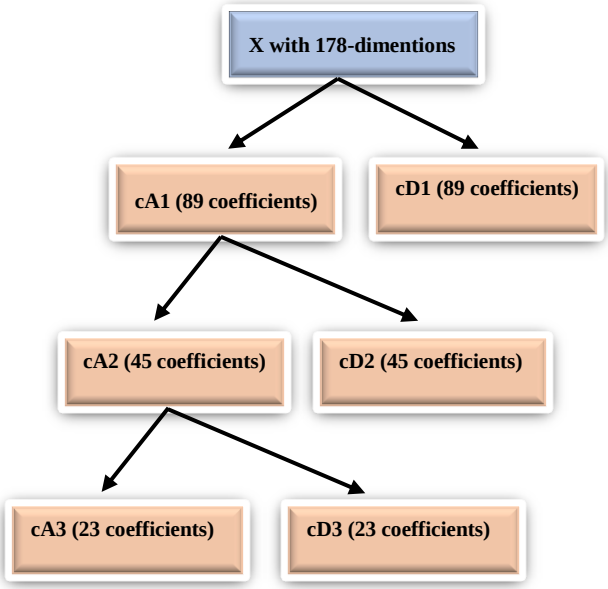

Figure 5: show three levels decomposition with DWT for EEG signal.

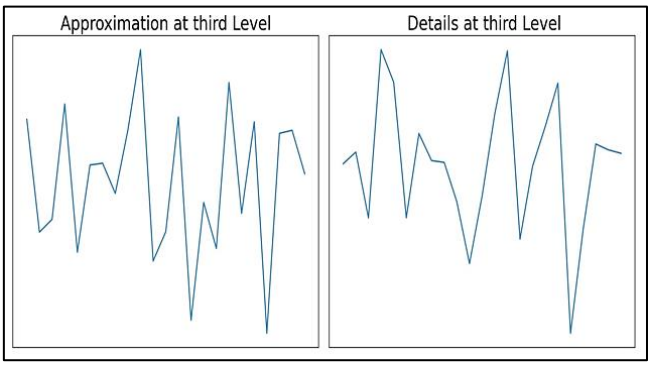

Figure 6: The approximation and Details coefficients with 23-dimension**.**

1. Enter directly to the clustering stage.
2. Alternatively, enter to dimension reduction stage to reduce data further.

4.4 Dimension Reduction

In this study, the Principle Component Analysis (PCA) was used to reduce and detect highly correlated features in the large EEG dataset into fewer independent variables while maintaining the EEG signal's characteristics. The input EEG data to the PCA is either from the DWT step with 23 features or the resulting data (D) from the preprocessing stage with 178 features. When applying the PCA, the number of these features is reduced to only 11 components. The charting of the first two and four PCs of the new reduced EEG data is illustrated in Figure 7.

The reduced features dimensions from this stage fed directly into the clustering stage.

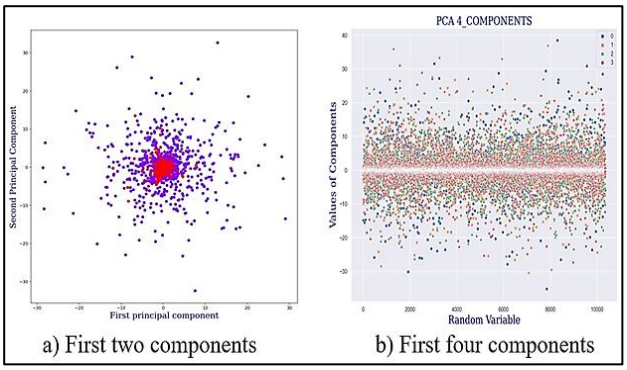

Figure 7: shows an example of the distribution of PCA components.

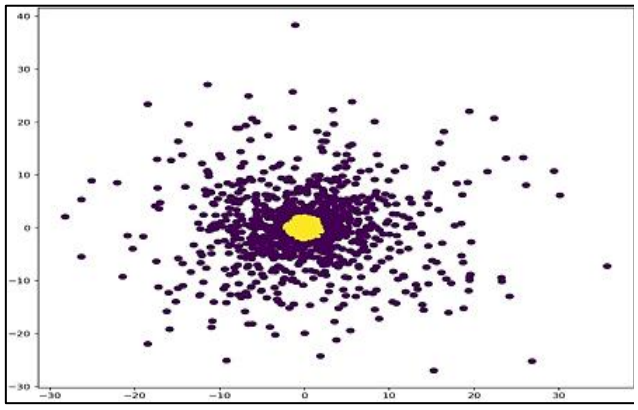

Figure 8: GMM uncertainty cluster assignment.

4.5 Clustering stage

The Gaussian Mixture Model (GMM) was used in this study. It is a probabilistic algorithm to cluster (group) or separate the EEG data from the previous stages into two clusters: epileptic seizures or non-epileptic seizures. The input to this stage is the resulted EEG data from the preprocessing stage with 178 features or the extracted EEG from the DWT with 23 features, or the EEG reduced features from the PCA with 11 features.

The fundamental concern with the GMM in this study was the continual fluctuation of the centroid and its lack of stability. Thus, various parameters were examined to ensure the centroid's stability, but none substantially affected the model. Consequently, just the number of clusters was used which is n_components = 2. In section 5, we will discuss how we have overcome the cluster instability issue. As shown in Figure 8, there is some degree of ambiguity in the clustering process.

The output clusters from this stage enter directly into the classification stage.

4.6 Classification stage

The final stage of building the system is the classification stage. The Random Forest classifier is implemented in this study, and the input to this stage is the clustering result from the previous stage. Multiple decision trees are used to predict the outputs based on the EEG training dataset features. In the end, the outcomes of all outputs were gathered utilizing a voting mechanism. Before the model training process, the first and most important step is defining the optimal parameters for obtaining the best model performance. Selecting the parameters was challenging, so each parameter in the random forest was evaluated until the best parameters were determined. The effective parameters utilized in this study for the random forest are illustrated in Table 4.

Table 4: The random forest effective parameters

Random Forest Parameters	Parameter Value
n_estimators	1000
max_depth	10
random_state	42

Now after selecting the best parameters the Random Forest can be trained on the training data. After training the Random Forest classifier, the testing data is supplied to this model to assess the classification results.

5 Experimental results

This section illustrates the results obtained from the proposed hybrid system of all experiments. The results of the proposed hybrid system will display in a table. It is worth mentioning that this table includes the experiment results for the following combinations: (DWT+GMM, DWT+PCA+GMM, Normal EEG+PCA+GMM, and Normal EEG+GMM); all these combinations are combined with the Random Forest classifier to define the cluster accuracy. The performance of this classifier is evaluated with many metrics, including testing accuracy, F1- score, recall, precision, and confusion matrix, to determine how well the model performed on the testing data. Because this study is about a medical dataset, the execution time is very important, and it is already included in each experiment. Finally, after displaying all the experiment results in the table, the best model with the best performance will be selected from among all the experiments. Table 5 illustrates the final results of the proposed hybrid system experiments.

Table 5: The Final Results of the proposed hybrid system with all experiments.

GMM		Reduction with PCA	Without Reduction
		Random Forest	
Accuracy	DWT	92.39	93.62
F1-score		0.92	0.94
Recall		0.92	0.94
Precision		0.93	0.94
Time-consuming in seconds		0.3587	0.4264
Samples correctly labeled		1275 out of 1380	1292 out of 1380
		Reduction with PCA	Without Reduction
		Random Forest	
Accuracy	Normal EEG Data	92.53	61.01
F1-score		0.93	0.55
Recall		0.93	0.61
Precision		0.93	0.77
Time-consuming in seconds		0.3276	0.7811
Samples correctly labeled		1277 out of 1380	842 out of 1380

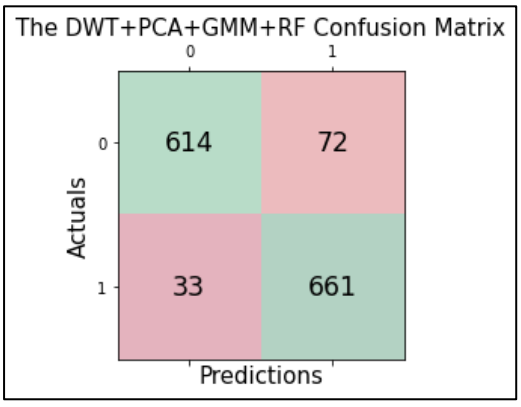

Figure 9: The confusion matrix for DWT and PCA features.

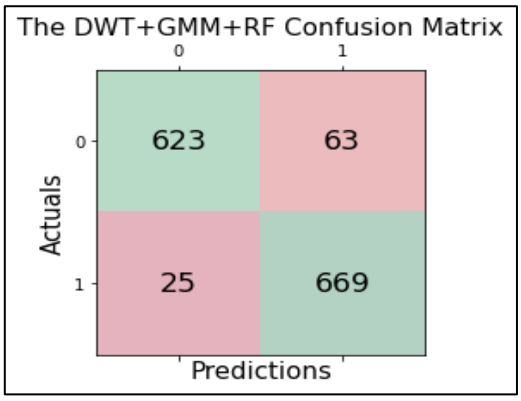

Figure 10: The confusion matrix for DWT features.

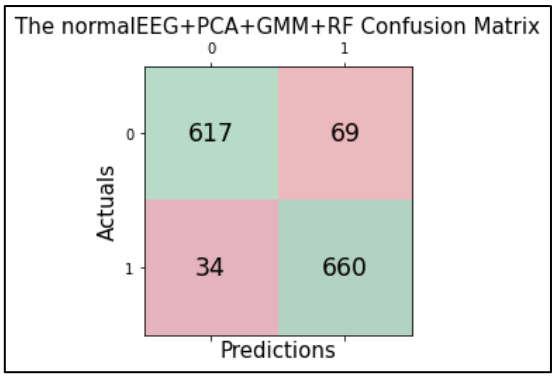

Figure 11: The confusion matrix for normal EEG and PCA.

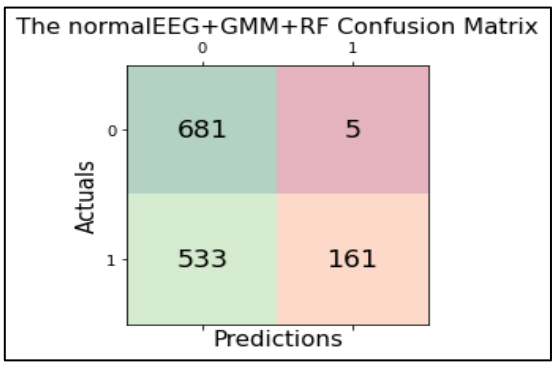

Figure 12: The confusion matrix for normal EEG data features.

6 Discussion

In this section, the experimental results that are referred to in Table 5 are discussed. The proposed hybrid system provides the best performance among all the experiments with the DWT features as it was able to accurately categorize (1292) samples from the testing dataset out of a total of (1380) samples, with an accuracy of (93.62%). Moreover, when the PCA was applied to the DWT features to reduce the data further, the results were good but slightly less than the case without the PCA reduction. As for the case when the PCA was applied to the normal EEG data, the proposed system performed well, but the results in this case are slightly less than the DWT features result. However, the worst performance was recorded when the EEG data were entered directly into the system. In addition, it was wrong in terms of clustering as the GMM fails to distinguish between both classes (epileptic seizure or non-seizure) equally. So, the above table shows the importance of using feature extraction and dimension reduction techniques with this system, as they reduce the number of features in the dataset while preserving the most essential features and information in the signal, leading to a significant improvement in the results. As for in terms of execution time, the proposed hybrid system was fast with the DWT features and with a slight improvement in time when applying the PCA technique

over the DWT. However, the system had the longest execution time with the normal EEG data. Once again, it becomes clear the importance of using the feature extraction and dimension reduction techniques, whether in terms of time or accuracy.

As mentioned before, the main problem faced in this study with the GMM is that the clustering is extremely different with each program execution. In addition, Since the GMM is combined with the Random Forest; different accuracies are obtained after every program execution. To overcome this issue, several experiments and tests were undertaken in this study, and the problem was solved by using the random-state parameter, as it considerably aided in stabilizing the clustering centroids and obtaining the same accuracy from the random forest after each program execution. The random-state parameter was employed extensively in this study:

1. In the test-train-split section, and the parameter value is (random_state = 1).
2. With the PCA technique, and the parameter value is (random_state = 1).
3. Finally, it is employed as a parameter with the GMM, and the parameter value is (random_state=25).

Moreover, selecting the values of the random-state parameter was another issue. After many experiments, the random-state parameter values were determined so that one corresponds to the other and corresponds to the value of the PCA component.

And As for the comparison with the related works [11], [12], [13], [14], and [15] that are illustrated in Table 1, it is important to say that our proposed hybrid system work is different from the previous studies presented in the field of detecting epileptic seizures (mentioned previously). In [11], the GMM was implemented as a classifier, in contrast to our proposed method, where it was used as a clustering algorithm. In [13], the EEG data were entered directly into the classifiers, while in our system, dimension reduction and feature extraction were used before the classification phase. Overall, our proposed hybrid system achieved the highest accuracy of 93.62% compared with [11], [12], and [13]. In addition, the two studies in [14] and [15] used a different EEG dataset. It is hard to compare with the related work since this study used a balanced dataset, which will make the results more realistic. In contrast, previous studies used imbalanced or different EEG datasets.

7 Conclusion

Epilepsy is one of the most dangerous diseases that affect human lives, so they need to diagnose. An important tool used for diagnosing epileptic seizures is the EEG test. There are many traditional approaches for analyzing EEG data for epilepsy detection, which are time-consuming and inaccurate. This paper proposes a new hybrid supervised and unsupervised system for detecting epileptic seizures from the EEG signals. The aim is to increase epileptic seizure detection accuracy in a balanced dataset while reducing the execution time. The process of detecting epileptic seizures goes through many stages. The first stage is the EEG signal preprocessing, which is the most important step in improving the system's performance. This stage aims to balance the EEG dataset and convert it into a more suitable format for the machine-learning algorithm. The second stage is features extraction with DWT to decompose the EEG signal into different sub-bands and extract the most important features from this signal. The output of this stage is the extracted features that will be used later in the dimension reduction or clustering stage. The third stage is the dimension reduction stage with PCA to select the best features from large number of features.

The output of the previous stages will be fed into the clustering stage with GMM to cluster the data into two clusters: epileptic and not epileptic seizures. The result of this stage will be entered directly into the EEG classification stage, where the random forest algorithm was used. The proposed system was evaluated with different metrics, including Accuracy, F1-score, Recall, Precision, and confusion matrix. The results showed that the proposed hybrid system achieved good results with the DWT and PCA features in terms of these metrics. In addition, the results of the experiments revealed that the DWT features combined with this hybrid system produced the highest result, with an accuracy of 93.62%. The new automated hybrid system can detect epilepsy with high accuracy and short execution time. Finally, in this study, we were able to solve the problem of centroid instability by using the random state parameter.

For future work, we are particularly interested in building a smart system application that may be employed in wearable devices to detect epileptic seizures.

References

[1] W. Mardini, M. M. B. Yassein, R. Al-Rawashdeh, S. Aljawarneh, Y. Khamayseh, and O. J. I. A. Meqdadi, "Enhanced detection of epileptic seizure using EEG signals in combination with machine learning classifiers," vol. 8, pp. 24046-24055, 2020. Doi: 10.1109/ACCESS.2020.2970012.

[2] N. S. Kadhim, D. M. Bachi, A. M. T. M. T. J. I. J. o. F. M. Abed, and Toxicology, "Knowledge of Primary School's Teachers about Epilepsy in Al-Basra City Center," vol. 15, no. 2, p. 1597, 2021.

[3] Z. Lasefr, S. S. V. Ayyalasomayajula, and K. Elleithy, "Epilepsy seizure detection using EEG signals," in 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), 2017, pp. 162-167: IEEE. https://doi.org/10.1109/UEMCON.2017.8249018.

[4] K. J. E. S. w. A. Akyol, "Stacking ensemble based deep neural networks modeling for effective epileptic seizure detection," vol. 148, p. 113239, 2020. https://doi.org/10.1016/j.eswa.2020.113239.

[5] K. M. J. I. i. M. U. Almustafa, "Classification of epileptic seizure dataset using different machine learning algorithms," vol. 21, p. 100444, 2020. https://doi.org/10.1016/j.imu.2020.100444.

[6] T. I. Rohan, M. S. U. Yusuf, M. Islam, and S. Roy, "Efficient approach to detect epileptic seizure using machine learning models for modern healthcare system," in 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 1783-1786: IEEE. https://doi.org/10.1109/TENSYMP50017.2020.9230731.

[7] [7] M. Zubair et al., "Detection of Epileptic Seizures From EEG Signals by Combining Dimensionality Reduction Algorithms With Machine Learning Models," vol. 21, no. 15, pp. 16861-16869, 2021. https://doi.org/10.1109/JSEN.2021.3077578.

[8] E. Kabir, S. Siuly, J. Cao, and H. J. I. J. o. C. I. S. Wang, "A computer aided analysis scheme for detecting epileptic seizure from EEG data," vol. 11, no. 1, pp. 663-671, 2018. https://www.atlantis-press.com/journals/ijcis/25892519.

[9] M. Zhou et al., "Epileptic seizure detection based on EEG signals and CNN," p. 95, 2018. https://doi.org/10.3389/fninf.2018.00095.

[10] C. Feudjio, V. D. Noyum, Y. P. Mofendjou, and E. J. a. p. a. Fokoué, "A Novel Use of Discrete Wavelet Transform Features in the Prediction of Epileptic Seizures from EEG Data," 2021. https://doi.org/10.48550/arXiv.2102.01647.

[11] R. Deepa, R. Anand, D. Pandey, B. K. Pandey, and B. J. M. P. i. E. Karki, "Comprehensive Performance Analysis of Classifiers in Diagnosis of Epilepsy," vol. 2022, 2022. https://doi.org/10.1155/2022/1559312.

[12] H. Ozturk, M. Ture, N. Kiylioglu, I. K. J. M. M. Omurlu, and D. Journal, "The Comparison of Different Dimension Reduction and Classification Methods in Electroencephalogram Signals/ Elektroensefalografi Sinyallerinde Farkli Boyut Indirgeme ve Siniflandirma Yontemlerinin Karsilastirilmasi," vol. 19, no. 4, pp. 336-345, 2018.

[13] M. Manjusha and R. Harikumar, "Performance analysis of KNN classifier and K-means clustering for robust classification of epilepsy from EEG signals," in 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2016, pp. 2412-2416: IEEE. https://doi.org/10.1109/WiSPNET.2016.7566575.

[14] H. Rajaguru and S. K. Prabhakar, "Analysis of Dimensionality Reduction Techniques with ABC-PSO Classifier for Classification of Epilepsy from EEG Signals," in Computational Vision and Bio Inspired Computing: Springer, 2018, pp. 625-633. 10.1007/978-3-319-71767-8_54.

[15] N. Kumar, K. Alam, A. H. J. B. Siddiqi, and P. Journal, "Wavelet Transform for Classification of EEG Signal using SVM and ANN," vol. 10, no. 4, pp. 2061-2069, 2017. https://dx.doi.org/10.13005/bpj/1328.

[16] R. Panda, P. Khobragade, P. Jambhule, S. Jengthe, P. Pal, and T. Gandhi, "Classification of EEG signal using wavelet transform and support vector machine for epileptic seizure diction," in 2010 International conference on systems in medicine and biology, 2010, pp. 405-408: IEEE. https://doi.org/10.1109/ICSMB.2010.5735413.

[17] A. S. Al-Fahoum and A. A. J. I. S. R. N. Al-Fraihat, "Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains," vol. 2014, 2014. http://dx.doi.org/10.1155/2014/730218.

[18] S. Phadikar, N. Sinha, R. J. I. J. o. B. Ghosh, and H. Informatics, "Automatic eyeblink artifact removal from EEG signal using wavelet transform with heuristically optimized threshold," vol. 25, no. 2, pp. 475-484, 2020. https://doi.org/10.1109/JBHI.2020.2995235.

[19] H. Hindarto, A. Muntasa, and S. Sumarno, "Feature Extraction ElectroEncephaloGram (EEG) using wavelet transform for cursor movement," in IOP Conference Series: Materials Science and Engineering, 2018, vol. 434, no. 1, p. 012261: IOP Publishing. doi:10.1088/1757-899X/434/1/012261.

[20] S. Badillo et al., "An introduction to machine learning," vol. 107, no. 4, pp. 871-885, 2020. https://doi.org/10.1002/cpt.1796.

[21] E. J. J. I. J. o. C. S. E. HARFASH and I. T. Research, "Face Recognition System Using PCA, LDA, Kernel PCA and Kernel LDA," vol. 6, no. 5, pp. 9-20, 2016.

[22] L. Farsi, S. Siuly, E. Kabir, and H. J. I. S. J. Wang, "Classification of alcoholic EEG signals using a deep learning method," vol. 21, no. 3, pp. 3552-3560, 2020. https://doi.org/10.1109/JSEN.2020.3026830.

[23] S. Ayesha, M. K. Hanif, and R. J. I. F. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," vol. 59, pp. 44-58, 2020. https://doi.org/10.1016/j.inffus.2020.01.005.

[24] A. Matin, R. A. Bhuiyan, S. R. Shafi, A. K. Kundu, and M. U. Islam, "A hybrid scheme using pca and ica based statistical feature for epileptic seizure recognition from eeg signal," in 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 2019, pp. 301-306: IEEE. https://doi.org/10.1109/ICIEV.2019.8858573.

[25] T. Xie, R. Liu, Z. J. A. M. Wei, and N. Sciences, "Improvement of the Fast Clustering Algorithm Improved by-Means in the Big Data," vol. 5, no. 1, pp. 1-10, 2020. https://doi.org/10.2478/amns.2020.1.00001.

[26] V. Vashishth, A. Chhabra, and D. K. J. C. C. Sharma, "GMMR: A Gaussian mixture model based

unsupervised machine learning approach for optimal routing in opportunistic IoT networks," vol. 134, pp. 138-148, 2019. https://doi.org/10.1016/j.comcom.2018.12.001.

[27] E. Patel and D. S. J. P. C. S. Kushwaha, "Clustering cloud workloads: K-means vs gaussian mixture model," vol. 171, pp. 158-167, 2020. https://doi.org/10.1016/j.procs.2020.04.017.

[28] J. Anitha, I.-H. Ting, S. A. Agnes, S. I. A. Pandian, and R. Belfin, "Social media data analytics using feature engineering," in Systems Simulation and Modeling for Cloud Computing and Big Data Applications: Elsevier, 2020, pp. 29-59. https://doi.org/10.1016/B978-0-12-819779-0.00003-4.

[29] G. Shobha and S. Rangaswamy, "Chapter 8-Machine Learning Handbook of Statistics," ed: Elsevier, 2018.

[30] A. K. Ali, A. M. J. I. J. o. E. Abdullah, and C. Engineering, "Fake accounts detection on social media using stack ensemble system," vol. 12, no. 3, 2022. DOI: 10.11591/ijece.v12i3.pp3013-3022.

[31] A. Parmar, R. Katariya, and V. Patel, "A review on random forest: An ensemble classifier," in International Conference on Intelligent Data Communication Technologies and Internet of Things, 2018, pp. 758-763: Springer.doi: 10.1007/978-3-030-03146-6_86.

[32] J. L. M. Kumar et al., "The classification of EEG-based winking signals: a transfer learning and random forest pipeline," vol. 9, p. e11182, 2021. https://doi.org/10.7717/peerj.11182.

[33] K. Singh, J. J. J. o. A. I. Malhotra, and H. Computing, "IoT and cloud computing based automatic epileptic seizure detection using HOS features based random forest classification," pp. 1-16, 2019. https://doi.org/10.1007/s12652-019-01613-7.

[34] P. Resque, A. Barros, D. Rosário, and E. Cerqueira, "An investigation of different machine learning approaches for epileptic seizure detection," in 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), 2019, pp. 301-306: IEEE. https://doi.org/10.1109/IWCMC.2019.8766652.

[35] P. Y. Taser, "Application of Bagging and Boosting Approaches Using Decision Tree-Based Algorithms in Diabetes Risk Prediction," in Multidisciplinary Digital Publishing Institute Proceedings, 2021, vol. 74, no. 1, p. 6. https://doi.org/10.3390/proceedings2021074006.

[36] Y. Xin et al., "Machine learning and deep learning methods for cybersecurity," vol. 6, pp. 35365-35381, 2018. https://doi.org/10.1109/ACCESS.2018.2836950.

[37] S. F. Raheem and M. J. I. Alabbas, "Dynamic Artificial Bee Colony Algorithm with Hybrid Initialization Method," vol. 45, no. 6, 2021. https://doi.org/10.31449/inf.v45i6.3652.

[38] N. M. A.-M. M. Al and R. S. J. I. Khudeyer, "ResNet-34/DR: A Residual Convolutional Neural Network for the Diagnosis of Diabetic Retinopathy," vol. 45, no. 7, 2021. https://doi.org/10.31449/inf.v45i7.3774.