A novel brain EEG clustering based on Minkowski distance to improve intelligent epilepsy diagnosis

Informatics in Medicine Unlocked 47 (2024) 101492

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

A novel brain EEG clustering based on Minkowski distance to improve intelligent epilepsy diagnosis

Dhiah Al-Shammary [a,*], Ekram Hakem [a], Ahmed M. Mahdi [a], Ayman Ibaida [b], Khandakar Ahmed [b]

[a] College of Computer Science and Information Technology, University of Al-Qadisiyah, Iraq
[b] Intelligent Technology Innovation Lab, Victoria University, Melbourne, Victoria, Australia

ARTICLE INFO

Keywords:
Particle Swarm Optimization (PSO)
Minkowski distance
Clustering
Feature selection algorithms
EEG data
Machine learning algorithms

ABSTRACT

This paper introduces a novel clustering approach based on Minkowski's mathematical similarity to improve EEG feature selection for classification and have efficient Particle Swarm Optimization (PSO) in the context of machine learning. Given the intricacy of high-dimensional medical datasets, feature selection plays a critical role in preventing disease and promoting public health. By employing Minkowski clustering, the objective is to group dataset records into two clusters exhibiting high feature coherence, thereby improving accuracy by applying optimization techniques like PSO to select the most optimal features. Furthermore, the proposed model can be extended to intelligent datasets, including EEG and others. As fewer features are needed for precise categorization, intelligent feature selection is an advanced step of machine learning. This paper investigates the key factors influencing feature selection in the EEG Bonn University dataset. The proposed system is compared against various optimization and feature selection methods, demonstrating superior performance in analyzing and classifying EEG signals based on accuracy measures. The experimental results have confirmed the effectiveness of the suggested model as a valuable tool for EEG data classification, achieving up to 100% accuracy. The outcomes of this research have the potential to benefit medical experts in related specialties by streamlining the process of identifying and diagnosing brain disorders. Technically, the machine learning algorithms RF, KNN, SVM, NB, and DT are employed to classify the selected features.

1. Introduction

In machine learning, the feature selection process holds significant importance as it involves identifying and removing non-effective features, resulting in several benefits such as model simplification, reduced training time, improved prediction accuracy, efficient memory storage, and avoidance of high-dimensional data [1]. As a high-dimensional dataset, the EEG dataset has many features that are employed in the classification procedure. However, not all of these features are essential for distinguishing between normal and epileptic seizures [2]. Therefore, selecting effective features has become crucial as it enables early detection, reduces data size, and mitigates the complexity of execution time due to the dataset's wide dimensions. Feature selection algorithms aim to identify the most compelling features that adequately explain the entire dataset without compromising the performance of the classification model [3]. Furthermore, two standard Filter and wrapper approaches are employed for feature selection as shown in Fig. 1.

Filter approaches in feature selection do not rely on machine learning methods to determine the selection of features. Instead, they utilize basic data qualities as a scoring mechanism for feature selection. Statistical indicators such as the Laplacian score and entropy are employed to calculate this score. Although Filter approaches require less processing time, they are only suitable for independent features [5].

On the other hand, Wrapper approaches encompass three essential components for feature selection: search strategy, prediction function, and fitness function. The search strategy selects the subset of features to be evaluated. The prediction function assesses the performance of the selected features compared to the fitness function, which can utilize any classification method. However, Wrapper techniques often face challenges with the time-consuming nature of the search strategy's processing. To overcome this limitation, metaheuristic techniques have emerged as a potential solution [4].

* Corresponding author.
E-mail addresses: d.alshammary@qu.edu.iq (D. Al-Shammary), com21.post1@qu.edu.iq (E. Hakem), ahmed.m.mahdi@qu.edu.iq (A.M. Mahdi), ayman.ibaida@vu.edu.au (A. Ibaida), Khandakar.Ahmed@vu.edu.au (K. Ahmed).

https://doi.org/10.1016/j.imu.2024.101492
Received 23 January 2024; Received in revised form 2 April 2024; Accepted 2 April 2024
Available online 3 April 2024

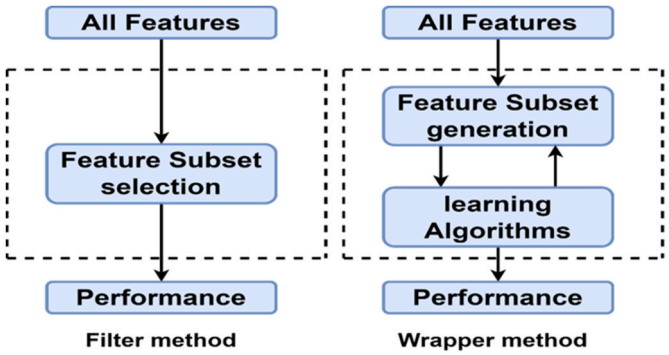

Fig. 1. Wrapper and Filter methods feature selection [4].

1.1. Motivation

The number of scientists studying the analysis of EEG signals has significantly increased in the last several years, primarily driven by the significance of these signals in the discovery and diagnosis of brain diseases [6]. With billions of interconnected neurons, the EEG signal is a complicated network that produces dozens of characteristics every second. When trying to classify EEG data, machine learning algorithms are faced with a significant problem due to this high feature rate. This is because they frequently require assistance in processing such a vast number of features, many of which may be undesired [6]. To address this challenge, feature selection algorithms have been developed to identify and select the best qualities for enhanced exploration and utilization [7]. Technically, effective extracting and reducing the data dimensions, feature selection techniques can lead to improvements in computational complexity, processing time, and memory storage. The main motivation for this paper is to increase feature selectors efficiency by determining features with high harmony of the dataset mainly by separate records into two clusters based on Minkowski's mathematical similarity. This would lead to more efficient Particle Swarm Optimization (PSO) as a feature selection example. However, it is important to acknowledge that feature selection and optimization algorithms have limitations. These limitations can be summarized in the following steps:

- Optimal Local Stagnation Probability: One notable limitation in optimization algorithms is the probability of encountering optimal local stagnation. This issue arises when there is a lack of diversity in the population, hindering the exploration of new solutions and impeding the extraction of essential features from previous iterations.
- High Time Requirement: Another limitation is the substantial processing time required by many iterative processes. This can lead to longer convergence rates and pose challenges in terms of computational efficiency.
- Inconsistent Results: In general, Particle Swarm Optimization (PSO) techniques tend to yield highly varied results, especially when applied to high-dimensional objective functions. This inconsistency can pose difficulties in achieving reliable and stable outcomes.

1.2. Problem statement

The high-dimensional medical dataset makes feature selection an essential process for early detecting diseases to protect people's health by reducing the number of features to be included in the classification function process [22]. High-dimensional medical data such as EEG burdens machine learning algorithms as they suffer greatly from high feature rates [23]. On the other hand, the present methodologies fall short of identifying the most useful features. In fact, search time complexity is technically a problem for the wrapper feature selection approaches. Moreover, the standard PSO algorithm for feature selection suffers from the stagnation effect in local optima, and the convergence rate of many iterative processes needs to be higher and produce consistent results. Potential feature selection algorithms must overcome traditional feature selection challenges and drawbacks. Therefore, The research problem is developing a new feature selection technique for analyzing and classifying EEG signals better to get consistent results to help diagnose and detect epileptic seizures from EEG signals.

1.3. Existing solutions

Main previous studies and achievements have concentrated on using metrics and algorithms to efficiently distribute and allocate with the aim increase the probability of having similar points in the same cluster. For instance, Minkowski Weighted K-means is optimized using Particle Swarm Optimization (PSO) in order to identify features [8]. This method has demonstrated an accuracy of 82.3% and 93.6%. Furthermore, a hybrid Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) is proposed [9] and based on *K*-means clustering technique for distinguishing two-class motor imagery (MI) tasks. The average accuracy achieved by the model is 60.42%. Additionally [10], the focus of this study was the application of clustering techniques utilizing the ACO-Decision Tree method, with the Euclidean distance measure employed for feature extraction. EEG classification accuracy is of 71.30%. In another study [11], the authors proposed utilization of Artificial Bee Colony (ABC) and Radial-Based Function Networks (RBFNNs) for the EEG signal analysis-based identification of epileptic seizure disorders. Resulting in a maximum accuracy of 82.3% for epilepsy identification.

1.4. Contribution

This research paper introduces a clustering approach based on Minkowski mathematical similarity to improve efficient Particle Swarm Optimization (PSO). The goal is to enhance machine learning by selecting more effective feature data and eliminating non-effective features. This process significantly reduces the number of features, consequently impacting system performance. Moreover, Minkowski feature selection demonstrates its applicability in handling high-dimensional intelligent data, such as EEG signals.

The proposed method utilizes the Minkowski scale to analyze EEG signals and identify the most effective features for Epileptic seizures detection in patient data. The research findings have demonstrated several promising achievements:

- **Low Stagnation Probability**: The Minkowski clustering technique aims to group dataset records into two clusters with high feature coherence, thereby promoting exploration and minimizing the likelihood of falling into stagnation.
- **Consistent Results**: Minkowski PSO clustering consistently produces similar outcomes when dealing with high-dimensional problems.
- **Reduced Time Requirement**: The Minkowski PSO clustering algorithm requires less time than traditional PSO methods to obtain the optimal solution for a given objective function.
- **Improved Accuracy**: Selecting highly optimal features through optimization techniques like PSO contributes to developing high-performance systems.

1.5. Paper organization

The rest of the paper is organized as Section 2 describes related work, Section 3 illustrates Minkowski distance metric, Section 4 presents the proposed method, Section 5 presents discussion and experimental results, Section 6 presents traditional evaluation of PSO, and finally conclusions and future work is presented in Sections 7 and 8 respectively.

2. Related works

Jamali-Dinan et al. [8] centered their investigation on identifying individuals suffering from Temporal Lobe Epilepsy (TLE), the most common type of focal epilepsy. This work presents a novel approach to optimize Minkowski Weighted K-means using Particle Swarm Optimization (PSO) (MWK) clustering was applied to identify the characteristics associated with temporal lobe epilepsy. The traditional K-means algorithm was found to be susceptible to noisy features. To address this limitation, the weighted K-means algorithm was optimized using the Minkowski distance metric. Technically, sensitivity to the initialization process is one of the challenges encountered with the weighted K-means algorithm. To mitigate this issue, the researchers incorporated Particle Swarm Optimization (PSO) to take advantage of the benefits of both MWK and PSO techniques and prevent local stagnation. The accuracy metric was the only one used to assess the suggested model. Additionally, the Silhouette criteria were employed to determine the optimal number of clusters. The researchers utilized standard datasets selected from the UCI Machine Learning Repository for their experiments. The proposed method demonstrated the ability to identify epilepsy with an accuracy of 82.3% and 93.6%. However, it is important to note that the evaluation of the model's performance was limited to the accuracy measure alone, neglecting other important criteria such as processing time and complexity.

Suraj et al. [9] conducted a study on diagnosing changes in brain cells using electroencephalogram (EEG) signals. The dynamic nature of the EEG signal prompted the researchers to explore the application of evolutionary algorithms (EA) in this context. In this paper, the authors proposed a hybrid Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) based *K*-means clustering technique for distinguishing two-class motor imagery (MI) tasks. The effectiveness of the proposed model was evaluated using two key metrics: accuracy and execution time. In order to validate the model, experimental evaluation was performed using standard datasets specifically designed for classifying two-class MI tasks. The average accuracy achieved by the model was reported to be 60.42%. However, it is worth noting that the researchers did not provide a comprehensive analysis of additional metrics such as precision, recall, and F1-score. This omission leaves the performance evaluation of the model somewhat ambiguous. In conclusion, the study conducted focused on diagnosing changes in brain cells through the analysis of EEG signals. The proposed hybrid GA-PSO-based *K*-means clustering technique showed promise in distinguishing two-class MI tasks. Nevertheless, the evaluation of the model's performance was limited to accuracy and lacked a comprehensive analysis of other important metrics. Future research should aim to address these limitations and provide a more thorough evaluation of the model's effectiveness.

Bursa et al. [10] conducted research on the processing of high-dimensional medical datasets using artificial intelligence techniques and evolutionary algorithms (EA). Applying clustering techniques utilizing the ACO-Decision Tree method, with the Euclidean distance measure employed for feature extraction is the main focus of this study. The effectiveness of the proposed model was evaluated using two key metrics: accuracy and Sensitivity-Specificity. To validate the model, experimental evaluation was conducted using standard datasets, specifically the MIT-BIH (ECG) database containing over 80,000 records, as well as EEG data consisting of approximately 4000 instances. The reported results indicated an ECG classification accuracy of 97.11% and an EEG classification accuracy of 71.30%. However, it is important to note that the researchers did not provide a comprehensive analysis of additional metrics, thereby leaving the performance evaluation of the model somewhat ambiguous. In conclusion, focused on the processing of high-dimensional medical datasets using artificial intelligence techniques and evolutionary algorithms. The proposed ACO-Decision Tree clustering method demonstrated promising results in ECG and EEG classification tasks. Future research should aim to address these limitations and provide a more thorough evaluation of the model's performance.

Satapathy et al. [11] conducted a study primarily focused on identifying and categorizing epileptic seizures in comparison to non-seizure patients. In this paper, the authors proposed the utilization of Artificial Bee Colony (ABC) and Radial-Based Function Networks (RBFNNs) for the EEG signal analysis-based identification of epileptic seizure disorders in the human brain. Several measures were used to assess the effectiveness of the suggested approach, including accuracy, recall, mean square error (MSE), and the discrete wave transformation (DWT) technique for feature extraction from the signal. Five sets of EEG data for epileptic seizure identification were collected from publicly accessible sources at the University of Bonn. The modified ABC algorithm was employed for the classification of EEG data, resulting in a maximum accuracy of 82.3% for epilepsy identification. However, in comparison to other existing methods, it is important to note that the outcomes of the proposed method are not highly efficient. Moreover, performance measurements like system complexity and computation time are missing. which could provide a more comprehensive assessment of the proposed method's performance. In conclusion, focused on the detection and classification of epileptic seizures using EEG signal analysis. The proposed approach utilizing ABC and RBFNNs showed potential, although it did not outperform other existing methods.

Wang et al. [12] conducted a study that focuses on using smart computing tools to diagnose epilepsy early. The study utilized the K-Nearest Neighbors (KNN) algorithm to determine the proximity between data points in both the training and validation datasets. The KNN method employed the Minkowski Distance metric, which takes into consideration three crucial factors: the distance metric itself, the selection of the K-value, and the decision-making process. The effectiveness of the classifier was assessed using various criteria, including accuracy, precision, recall, sensitivity, specificity, and the F1-score. To validate the model, the researchers utilized the Bonn EEG dataset, which is a widely accepted standard dataset for experimental evaluation in this domain. The experimental results have indicated an average classification accuracy of 100% for the proposed model. However, it is clear that this method has certain weaknesses. More specifically, lower feature counts in high-dimensional datasets can result from applying KNN algorithms with the Minkowski scale. Concluded with an emphasis on using smart computing tools to diagnose epilepsy early. The KNN algorithm with the Minkowski Distance metric demonstrated promising results, achieving a high average classification accuracy of 100%. However, it is important to consider the limitations of this method, particularly the potential loss of features in high-dimensional datasets when utilizing KNN algorithms with the Minkowski scale.

3. Minkowski distance

In the context of a normed vector space, Minkowski distance serves as a metric for measuring the similarity of distances between vectors. This distance metric can be employed in machine learning to assess the similarity of two or more vectors [13].

Consider a symmetric, open, strictly convex set X, which represents a bounded domain in R. The function d, defined by

$$\mathbf{D} = \left(\sum_{i=0}^{n} |\boldsymbol{P}_i - \boldsymbol{Q}_i|^k \right)^{\frac{1}{k}} \tag{1}$$

D represents the Minkowski distance. Here, k denotes a parameter, n represents the number of data vector values (attributes), P_i and Q_i denote the data points (data values).

The term ($|P_i\text{-}Q_i|$) represents the distance between two points in an n-dimensional space, as measured by the Minkowski metric. This metric adheres to the strict triangle inequality and serves as a generalization of both the Manhattan distance and the Euclidean distance.

When k is equal to 1, the Minkowski distance coincides with the

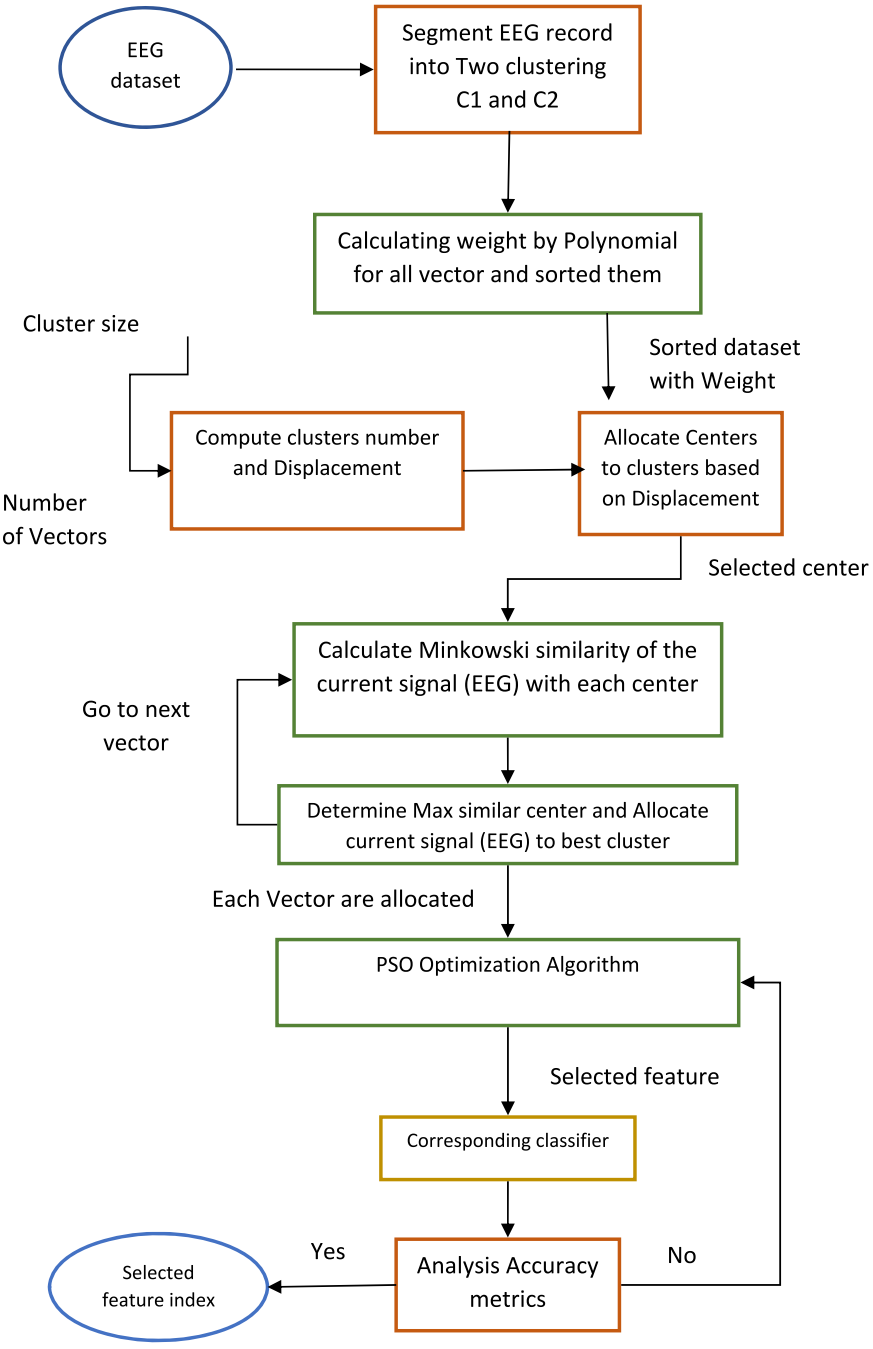

Fig. 2. Illustrates the main operations of the proposed Minkowski PSO clustering.

Manhattan distance, given by

$$\mathbf{D}_{\mathbf{M}=}\sum_{i=1}^{n}\left|\boldsymbol{P}_i-\boldsymbol{Q}_i\right| \tag{2}$$

Alternatively, when k is equal to 2, the Minkowski distance aligns with the Euclidean distance, represented by

$$\mathbf{D}_{\mathbf{E}=}\left[\sum_{i=0}^{n}\left(\boldsymbol{P}_i-\boldsymbol{Q}_i\right)^2\right]^{1/2} \tag{3}$$

The Minkowski distance is widely utilized in the field of machine learning, particularly when seeking to determine optimal correlations or classifications of data [14].

4. Proposed method

The Particle Swarm Optimization (PSO) algorithm is widely recognized as one of the most commonly used metaheuristic algorithms in feature selection for high-dimensional datasets due to its effectiveness and ease of implementation [15]. PSO has demonstrated the capability to efficiently identify optimal features compared to other approaches. However, the traditional PSO algorithm often encounters the issue of stagnation at local optimum, resulting from deficiencies in solution search and exploitation [7].

This paper introduces a novel clustering approach based on Minkowski for (PSO), specifically for selecting highly effective features in EEG datasets. Minkowski similarity measurements are applied to the static clustering of the EEG dataset based on the maximum similarity values. To achieve potentially high accuracy, divide the EEG dataset

Table 1
Explains the Accuracy metric results of classifiers SVM, KNN, DT, NB, and RF With PSO Minkowski clustering optimizer based on EEG signal length 15s (2604 features)test size 10%.

Dataset	SVM	KNN	DT	NB	RF
S, Z	60	60	85	100	90
S, Z, O	76.6	76.6	90	100	96.6
F, N, S	76.6	50	80	66.6	83.3
N, O, S, Z	50	62.5	75	85	90
F, O, S, Z	50	60	72.5	72.5	92.5
F, N, O, S	50	70	72.5	70	82.5
F, N, S, Z	52.5	62.5	62.5	75	87.5
F, N, O, S, Z	32	58	58	70	90

Table 2
Explains the Accuracy metric results of classifiers SVM, KNN, DT, NB, and RF With PSO Minkowski clustering optimizer based on EEG signal length 15s (2604 features)test size 30%.

Dataset	SVM	KNN	DT	NB	RF
S, Z	65	51.6	81.6	98.3	96.6
S,Z,O	70	73.3	86.6	100	96.6
F, N, S	70	58.8	74.4	71.1	83.3
N,O,S,Z	48.3	58.3	72.5	85	90.8
F, O, S, Z	48.3	56.6	72.5	80.8	94.1
F, N, O, S	48.3	63.3	75	79.1	85.8
F, N, S, Z	52.5	61.6	69.1	74.1	79.1
F, N, O, S, Z	37.3	55.3	58.6	69.3	81.3

Table 3
Describe the Accuracy metric results of classifiers SVM, KNN, DT, NB, and RF With PSO Minkowski clustering optimizer based on EEG signal length 23.6s (4097 features)test size 30%.

Dataset	SVM	KNN	DT	NB	RF
S,Z	60	53	75	100	96.6
S,Z,O	70	73	77.7	100	96.6
F, N, S	70	56	75.5	71.1	82.2
N,O,S,Z	48	55	65.8	82.5	85.8
F, O, S, Z	48	55	69.1	75	87.5
F, N, O, S	48	60.8	69.1	79.1	83.3
F, N, S, Z	48	66	72.5	80	85
F, N, O, S, Z	37	52.6	58.6	72	80.66

Table 4
Explains the precision metric results of classifiers SVM, KNN, DT, NB, and RF With PSO clustering based on EEG signal length 15s (2604 features)test size 10%.

Dataset	SVM	KNN	DT	NB	RF
S, Z	30	30	87.3	100	92.7
S,Z,O	38.3	38.3	98.0	100	98.0
F, N, S	38.3	60.7	53.2	70	67.8
N,O,S,Z	25	78.5	75.9	86.9	90.6
F, O, S, Z	25	77.8	72.5	70.0	92.6
F, N, O, S	25	79.5	75	67.8	83.1
F, N, S, Z	22.0	61.0	56.0	66.1	88.8
F, N, O, S, Z	10.6	62.2	47.6	52.7	93.6

records into two clustering to obtain better accuracy by selecting the optimal and more similar features in the classification process. Experiments have shown that EEG dataset can have necessary features for the diagnosis of brain diseases more effectively when using the Minkowski clustering model compared to other methods. The selected features are then classified using machine learning algorithms such as Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes (NB), and Decision Tree (DT).

The proposed system aims to address the challenges associated with

Table 5
Explains the recall metric results of classifiers SVM, KNN, DT, NB, and RF With PSO clustering based on EEG signal length 15s and test size 10%.

Dataset	SVM	KNN	DT	NB	RF
S, Z	50	50	78.3	100	86.6
S,Z,O	50	50	91.6	100	91.6
F, N, S	50	59.8	72.7	69.3	72.7
N,O,S,Z	50	62.5	72.5	85	90
F, O, S, Z	50	60	72.5	72.5	92.5
F, N, O, S	50	70	75	70	82.5
F, N, S, Z	41.6	54.4	57.1	65.8	82.0
F, N, O, S, Z	33.3	43.4	34.2	63.5	85.4

Table 6
Explains the F1-score metric results of classifiers SVM, KNN, DT, NB, and RF With PSO clustering based on EEG signal length 15s and test size 10%.

Dataset	SVM	KNN	DT	NB	RF
S, Z	37.2	37.2	89.5	100	89.5
S,Z,O	43.3	43.3	94.6	100	94.6
F, N, S	43.3	60.2	58.5	69.6	70
N,O,S,Z	33.3	69.6	72.6	85.9	90.3
F, O, S, Z	33.3	67.7	72.5	71.2	92.5
F, N, O, S	33.3	74.1	67.9	68.8	82.8
F, N, S, Z	28.8	57.4	56.5	66.0	85.2
F, N, O, S, Z	15.9	51.2	41.2	57.6	89.1

Table 7
Describes the Accuracy metric results of classifiers SVM, KNN, DT, NB, and RF With PSO optimization based on EEG signal length 15s (2604 features) test size 30%.

Dataset	SVM	KNN	DT	NB	RF
S, Z	48	61	90	100	100
S, Z, O	30	43	61	78	76
F, N, S	30	53	62	72	91
N, O, S, Z	21	35	56	60	67
F, O, S, Z	21	38	51	62	75
F, N, O, S	21	45	39	57	72
F, N, S, Z	21	50	53	56	70
F, N, O, S, Z	18.0	32	38	58	67

Table 8
Describes the Accuracy metric results of classifiers SVM, KNN, DT, NB, and RF With PSO optimizer based on EEG signal length 15s (2604 features) test size 10%.

Dataset	SVM	KNN	DT	NB	RF
S, Z	50	65	80	100	100
S, Z, O	20	50	56.6	73.3	66.6
F, N, S	20	50	73.3	76.6	93.3
N,O,S,Z	20	32.5	57.5	67.5	67.5
F, O, S, Z	20	37.5	55	62.5	77.5
F, N, O, S	20	47.5	62.5	60	77.5
F, N, S, Z	20	50	35	50	82.5
F, N, O, S, Z	18	32	54	60	76

the classification of high-dimensional data by partitioning the dataset into two clusters with high feature harmony. This approach promotes exploration and reduces the likelihood of stagnation. Mathematical models and statistics are employed in machine learning to identify patterns within high-dimensional datasets. Specifically, Minkowski similarity measurements are utilized to cluster EEG datasets based on the maximum similarity values. Fig. 2 provides a detailed overview of the main steps of the proposed method.

In this model, vector weights are calculated using a polynomial equation (Eq. 1), and they are subsequently sorted in ascending order to determine the center of the point. Additionally, the displacement be-

Table 9
Describes the Accuracy metric results of classifiers SVM, KNN, DT, NB, and RF With PSO optimizer based on an EEG signal length of 23.6s and test size of 30%.

Dataset	SVM	KNN	DT	NB	RF
S, Z	48.3	61.6	76.6	100	100
S,Z,O	30	43.3	60	78.8	75.5
F, N, S	30	55.5	56.6	71.1	91.1
N,O,S,Z	21.6	35	46.6	65	72.5
F, O, S, Z	21.6	36.6	55	61.6	75
F, N, O, S	21.6	43.3	49.1	56.6	75.8
F, N, S, Z	21.6	46.6	56.6	57.5	74.1
F, N, O, S, Z	18	28.6	36	59.3	66

Table 10
Describes the precision metric results of classifiers SVM, KNN, DT, NB, and RF With PSO optimizer based on EEG signal length 15s (2604 features) test size 10%.

Dataset	SVM	KNN	DT	NB	RF
S, Z	25	79.4	85.3	100	100
S,Z,O	6.6	55.8	53.	70.0	65.5
F, N, S	6.6	66.8	77.2	69.7	91.4
N,O,S,Z	5	48.8	59.3	70.6	67.1
F, O, S, Z	5	62.8	56.2	61.3	80.9
F, N, O, S	5	55.4	63.0	65.0	84.4
F, N, S, Z	5	55.5	40.0	52.9	82.1
F, N, O, S, Z	3.6	52.7	59.5	55.0	76.0

Table 11
Describes the recall metric results of classifiers SVM, KNN, DT, NB, and RF With PSO optimizer based on EEG signal length 15s (2604 features) test size 10%.

Dataset	SVM	KNN	DT	NB	RF
S, Z	50	65	85	100	100
S, Z, O	33.3	43.3	67.9	70.5	65.3
F, N, S	50	65	80	100	100
N, O, S, Z	33.3	43.3	51.2	70.5	65.3
F, O, S, Z	33.3	57.6	72.1	69.1	91.4
F, N, O, S	25	34.9	52.0	65.8	65.1
F, N, S, Z	25	41.1	46.3	61.7	75.8
F, N, O, S, Z	25	43.8	52.2	64.0	79.4

Table 12
Describes the F1-score metric results of classifiers SVM, KNN, DT, NB, and RF With PSO optimizer based on EEG signal length 15s (2604 features) test size 10%.

Dataset	SVM	KNN	DT	NB	RF
S, Z	33.3	71.4	75.1	100	100
S,Z,O	11.1	48.8	60.3	70.2	65.4
F, N, S	11.1	61.8	69.1	69.4	91.4
N,O,S,Z	8.3	40.7	48.3	68.1	66.1
F, O, S, Z	8.3	49.7	43.4	61.5	78.3
F, N, O, S	8.3	48.9	49.8	64.5	81.8
F, N, S, Z	8.3	50.8	31.9	54.2	83.2
F, N, O, S, Z	6.1	41.8	48.5	56.4	75.7

tween centers is computed to achieve better clustering by obtaining optimal spacing. The remaining EEG dataset is then allocated to clusters based on Minkowski measurements. For each vector, Minkowski similarity measurements are computed with all centers, and the signal is allocated to the cluster with the maximum similarity value. This process ensures that highly harmonious features are selected through optimization techniques like PSO, which enhances the system's performance and improves accuracy. Finally, the average accuracy for the two clusters is determined.

$$W = \sum_{k=1}^{N} \mathrm{k} * x_k \tag{4}$$

Where: N: - number of vector items.
x_k: - vector items
k: - vector index.

5. Discussion and experimental results

In order to assess the effectiveness of the proposed model, various performance metrics including accuracy, recall, F1-score, and precision are calculated. These metrics provide a comprehensive analysis and clear evaluation of the system's performance [16]. In order to evaluate the proposed model, a high-dimensional and widely used Bonn University EEG dataset is employed. This dataset is commonly utilized for diagnosing and identifying epileptic seizures [17]. Furthermore, PSO empowered with Minkowski clustering model is compared with the traditional PSO approach. Experimental results indicate that the utilization of the Minkowski clustering model yields superior effectiveness compared to alternative methods.

The main limitation for the proposed model is forcing records allocation to the closest cluster although it may not highly similar. Technically, this is clearly caused by the static clustering behavior as all records have to be allocated to one of the two clusters.

5.1. Dataset description

The electroencephalogram (EEG) is a valuable and cost-effective diagnostic tool used to examine the electrical activity in the brain. It serves as the most prevalent method for diagnosing changes in brain functions. EEG data is obtained by placing electrodes on the scalp, enabling the measurement of brain activity. This technique is particularly effective in identifying and monitoring neurological conditions such as epilepsy and sleep disturbances [18]. In addition to its diagnostic applications, EEG signals are used in many investigations and research projects, including as applications for gaming and lie detection. However, these signals are characterized by complexity, noise, nonlinearity, and instability. Consequently, extracting meaningful information related to the brain from EEG signals is a challenging task. As a result, many researchers have proposed a range of feature selection and optimization techniques to accurately analyze and classify EEG signals, to safeguard individuals' health, and facilitate early detection of brain diseases. The dataset employed in this study consists of 500 segments of single-channel EEG recordings, which have been evenly divided into five sets. Each set comprises 100 files representing five distinct classes. All sets adhere to identical patient conditions. Each EEG signal within the dataset lasts 23.6 s and is associated with 4097 features.

5.2. Results for clustering approach based on Minkowski

The superiority of a feature selection and clustering algorithm lies not in its ability to select fewer features, but rather in its capability to identify the most effective features that significantly impact the accuracy of the objective function. In this study, the implementation of Minkowski clustering aims to group the dataset records into two clusters with high feature harmony, thereby enhancing accuracy through the use of Particle Swarm Optimization (PSO) to select highly optimal features. Various data test sizes (10%, 20%, and 30%) have been utilized in order to evaluate the system's performance and assess the impact of training accuracy on the developed patterns. Additionally, different signal lengths (1s, 5s, 15s, and 23.6s) have been employed during the evaluation. No preprocessing techniques have been applied to the EEG signals. Remarkably, even with a small record of EEG signals with a length of 1 s, the Minkowski clustering approach has demonstrated promising achievements. Experimental results indicate that the highest accuracy

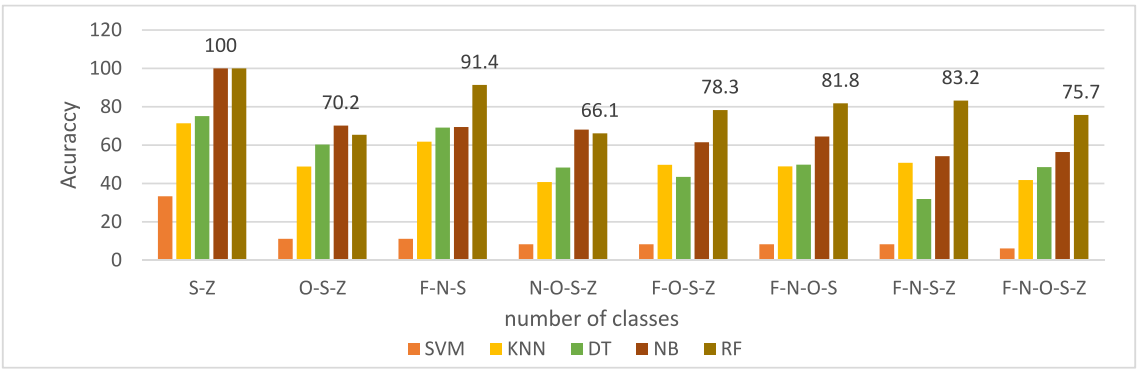

Fig. 3. Describes the Accuracy metric results of classifiers SVM, KNN, DT, NB, and RF With PSO optimizer based on EEG signal length 15s (2604 features) test size 30%.

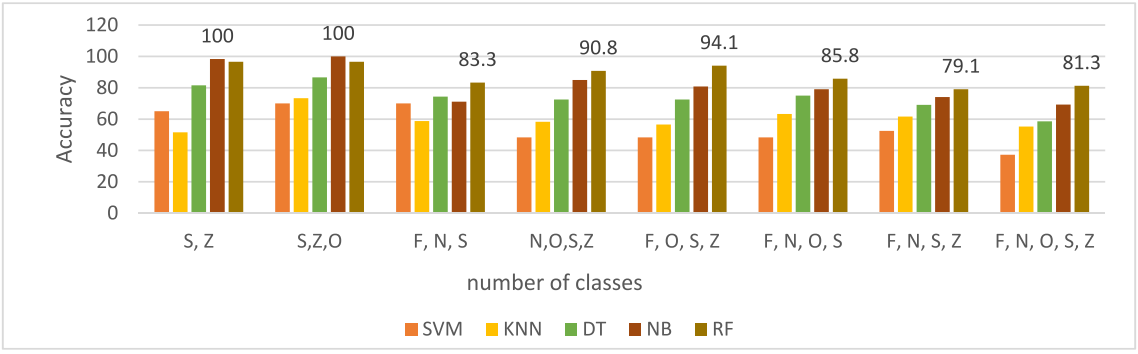

Fig. 4. Explains the Accuracy metric results of classifiers SVM, KNN, DT, NB, and RF With PSO Minkowski clustering optimizer based on EEG signal length 15s (2604 features) test size 30%.

Fig. 5. Explains the Accuracy metric results of classifiers SVM, KNN, DT, NB, and RF With PSO Minkowski clustering optimizer based on EEG signal length 15s (2604 features) test size 10%.

achieved for two classes was 100% for classifiers Naive Bayes (NB) and Random Forest (RF), except for the K-Nearest Neighbors (KNN) classifier, which exhibited a decrease in accuracy. The performance tests presented in Tables 1–6 provide detailed insights into the accuracy metric results of classifiers Support Vector Machine (SVM), KNN, Decision Tree (DT), and RF with Minkowski feature selection clustering. Notably, the clustering model required less execution time to select effective features in the five classes for the NB classifier, with an average of 0.00028 s, compared to the standard PSO approach, which had an average execution time of 0.00036 s.

6. PSO traditional evaluation

The Particle Swarm Optimization (PSO) algorithm, Stochastic optimization is a technique first presented by Eberhart and Kennedy in 1995. PSO is inspired by the collective behavior of various species such as insects, herds, birds, and fish [19]. These species exhibit cooperative behavior in acquiring food, with each member continuously adapting their search strategy based on its own experiences and the experiences of other members.

PSO is a computational technique that iteratively improves candidate solutions to optimize a given quality measure. It operates by maintaining a swarm of particles, where each particle represents a potential solution to the problem at hand. These particles navigate through a multidimensional search space, adjusting their positions based on their own experiences and the experiences of their neighboring particles [20]. Recently, PSO has been successfully applied in various research and application domains. It has demonstrated superior performance

Table 13
Comparison of Minkowski selection features with other previous models.

Author	Method	Dataset	Best accuracy
Jamali-Dinan et al. [8]	PSO and MWK methods	The EEG	93.6%
Suraj et al. [9]	a hybrid GA-PSO-based *K*-means clustering	classifying two class MI tasks	60.42 %.
Miroslav Bursa et al. [10]	Clustering techniques using the ACO-Decision Tree method	The EEG	71%
Satapathy et al. [11]	of Artificial Bee Colony (ABC) and Radial-Based Function Networks (RBFNNs)	University of Bonn	82.3
Wang et al. [12]	KNN method employed the Minkowski Distance metric	University of Bonn	100%
Proposed method	Minkowski selection features and clustering	The EEG	100%

compared to other methods, offering faster and more cost-effective outcomes. Another notable advantage of PSO is its limited number of adjustable parameters [15]. A single version of the algorithm can be readily applied to different applications with minimal modifications. The local best optimum (pbest) designates the PSO's particle's optimal solution. Following a full optimization cycle, the global best solution (gbest), which is the best among the pbests, is updated. Equation (5) calculates the velocity of particles [21].

$$Vi\,d(t+1) = w(t)Vid(t) + c1r1(pbest\,id - xi\,d(t) + c2r2(gbest\,d - xid(t))\ldots\ldots \quad (5)$$

where: $vi\,d\,(t)$,$xi\,d\,(t)$ indicates the velocity and position of $i\,th$ particle at iteration t in $d\ th$ dimension respectively, $c1$ and $c2$ are positive coefficients. $r1$ and $r2$ are random variables in the range [0, 1]. w is the inertia weight. Equation (6) is applied to find the new value of particle position (candidate solution).

$$xi\,t+1 = xi\,t + vi\,t + 1 \ldots\ldots \quad (6)$$

Where: $xi\ t{+}1$ is the new particle position and $xi\ t$ is the old particle position.

6.1. Results for traditional PSO algorithm

The following tables present the results of the PSO optimizer for various evaluation metrics, including accuracy, precision, recall, and F1-score. These results are obtained using a signal duration of 15 s and a testing size of 30% with a corresponding training size of 70%. Furthermore, the proposed Minkowski feature selection technique is compared with the traditional PSO approach. Experimental observations reveal that the accuracy tends to decrease as the number of classes increases. Additionally, the highest accuracy is achieved for the Naive Bayes (NB) and Random Forest (RF) classifiers when dealing with two classes. Evidently, PSO optimizers typically rely on random feature selection, necessitating multiple iterations to attain the best feature subset during the system training stage [19,20]. Furthermore, these optimizers often require a larger number of features and more execution time, leading to increased complexity. Detailed performance test results can be found in Tables 7–12.

Bar chart figures serve as a visual display, With the goal of obtaining a lucid visual evaluation of all resulting Accuracy, shown in Table 7. (Fig. 3), by different classes for the PSO optimizer for 15 s of signal length. Although Minkowski has been applied to a small dataset testing size, potential Accuracy has been achieved. SVM has shown the least accurate achievement when Minkowski is applied. However, KNN keeps outperforming SVM with the Minkowski feature selection. Moreover, the performance increases with the height of the sample length signals. Shown in Tables 3 and 4 (Figs. 4 and 5).

Finally, several different methods have been devised to identify seizures caused by epilepsy. Using accuracy metrics, the suggested strategy is contrasted with other previously established techniques. This comparison only includes strategies tested within the same dataset, allowing Results from groups of the same classes will be compared. The comparison results in Table 13 show that Minkowski feature selection and clustering models outperformed most of the earlier techniques. Most alternative EEG signal categorization methods have employed up to two classes to evaluate the performance of their classifiers. In contrast to our way of classifying various EEG data, we have discovered more states.

7. Conclusion

In summary, the progress of feature selection algorithms often requires additional support to overcome the challenge of stagnation and improve exploration efficiency. This paper presents a novel clustering approach based on Minkowski's mathematical similarity, which demonstrates high efficiency in machine learning applications, characterized by fast detection and high accuracy. The proposed method effectively mitigates the issue of stagnation commonly encountered in clustering algorithms. Generally, clustering is a widely used machine learning technique, assigns data points to groups based on their similarities, without prior knowledge of the data point labels. The accuracy of the Minkowski clustering approach surpasses that of alternative methods, offering improved classification results. Moreover, compared to other models, Minkowski exhibits lower complexity as it does not require an iterative mode like Particle Swarm Optimization (PSO). One limitation of the proposed system is that the accuracy decreases with an increasing number of classes. However, empirical findings demonstrate the potential of Minkowski feature selection, even with a small dataset of EEG signals. Minkowski enables traditional classifiers to achieve high accuracy levels, close to or equal to 100%, with the exception of the K-Nearest Neighbors (KNN) classifier, which exhibits reduced accuracy as the number of features increases. Although the evaluation is conducted on a small dataset testing size, the potential for achieving high accuracy is evident. Support Vector Machine (SVM) achieves the lowest accuracy when combined with Minkowski optimization, while KNN consistently outperforms SVM in this regard.

8. Future work

In order to overcome the limitation of the static Minkowski clustering when all records would be allocated to clusters although they may not quietly similar, a new dynamic Minkowski clustering model would be proposed for future work. This would assure allocating records to their best similar clusters.

Using various methods for processing EEG data, such as the wavelet transform, Fourier transform, and others before the Minkowski measure Similarity.

The system proposed used only PSO algorithms for clustering data. Can be developed system through applying many optimization algorithms such as GWO, ABC, ACO, Firefly, and other approaches to increase efficiency.

Ethical statement

We wish to confirm that all experiments involved in this research are applied on a publicly available dataset as no human or animal lives included in any experiments.

CRediT authorship contribution statement

Dhiah Al-Shammary: Writing – review & editing, Supervision, Project administration, Methodology. **Ekram Hakem:** Writing – original draft, Methodology, Conceptualization. **Ahmed M. Mahdi:** Formal analysis, Conceptualization. **Ayman Ibaida:** Writing – review & editing, Investigation, Formal analysis. **Khandakar Ahmed:** Writing – review &

editing, Supervision, Investigation, Formal analysis.

Declaration of competing interest

No conflict of interest exists.

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgements

This research was supported/partially both the University of Al-Qadisiyah (Iraq) and Victoria University (Australia). We are thankful to both Universities who provided labs, devices, and expertise that greatly assisted the research.

References

[1] Qaisar SM, Hussain SF. Effective epileptic seizure detection by using level-crossing EEG sampling sub-bands statistical features selection and machine learning for mobile healthcare. Comput Methods Progr Biomed May 2021;203. https://doi.org/10.1016/j.cmpb.2021.106034.

[2] Al-hamzawi AA, Al-Shammary D, Hammadi AH. A survey on healthcare EEG classification-based ML methods. In: Shakya S, Ntalianis K, Kamel KA, editors. Mobile computing and sustainable informatics. Lecture notes on data engineering and communications technologies, vol. 126. Singapore: Springer; 2022. https://doi.org/10.1007/978-981-19-2069-1_64.

[3] Nakisa B, Rastgoo MN, Tjondronegoro D, Chandran V. Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors. Elsevier Ltd Expert Syst Appl 2018;93:143–55. https://doi.org/10.1016/j.eswa.2017.09.062. Mar. 01.

[4] Al-hamzawi AA, Al-Shammary D, Hammadi AH. Health electroencephalogram epileptic classification based on Hilbert probability similarity. Int J Electr Comput Eng 2023;13(Issue 3). https://doi.org/10.11591/ijece.v13i3.pp3339-3347.

[5] Zhu Z, Ong YS, Dash M. Wrapper-filter feature selection algorithm using a memetic framework. IEEE Trans Syst Man Cybern B Cybern Feb 2007;37(1):70–6. https://doi.org/10.1109/TSMCB.2006.883267.

[6] Aayesha, Qureshi MB, Afzaal M, Qureshi MS, Fayaz M. Machine learning-based EEG signals classification model for epileptic seizure detection. Multimed Tool Appl May 2021;80(12):17849–77. https://doi.org/10.1007/s11042-021-10597-6.

[7] Al-Shammary D, Albukhnefis AL, Alsaeedi AH, Al-Asfoor M. Extended particle swarm optimization for feature selection of high-dimensional biomedical data. Concurr Comput Nov 2022;34(10):e6776. https://doi.org/10.1002/CPE.6776.

[8] Jamali-Dinan SS, et al. A combination of particle swarm optimization and minkowski weighted K-means clustering: application in lateralization of temporal lobe epilepsy. Brain Topogr Jul 2020;33(4):519–32. https://doi.org/10.1007/s10548-020-00770-9.

[9] Suraj, Tiwari P, Ghosh S, Sinha RK. Classification of two class motor imagery tasks using hybrid GA-PSO based K-means clustering. Comput Intell Neurosci 2015;2015. https://doi.org/10.1155/2015/945729.

[10] Bursa M, Lhotska L. Artificial intelligence methods in electrocardiogram and electroencephalogram data clustering. 2009 [Online], www.worldscientific.com.

[11] Satapathy SK, Dehuri S, Jagadev AK. ABC optimized RBF network for classification of EEG signal for epileptic seizure identification. Egypt Inform J Mar 2017;18(1):55–66. https://doi.org/10.1016/j.eij.2016.05.001.

[12] Wang S, Jiang Y. Exploration of smart medical technology based on intelligent computing methods. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Springer Science and Business Media Deutschland GmbH; 2021. p. 284–93. https://doi.org/10.1007/978-3-030-84529-2_24.

[13] Xu H, Zeng W, Zeng X, Yen GG. An evolutionary algorithm based on Minkowski distance for many-objective optimization. IEEE Trans Cybern Nov 2019;49(11):3968–79. https://doi.org/10.1109/TCYB.2018.2856208.

[14] Vera JF, Heiser WJ, Murillo A. Global optimization in any minkowski metric: a permutation-translation simulated annealing algorithm for multidimensional scaling. J Classif Sep 2007;24(2):277–301. https://doi.org/10.1007/s00357-007-0020-1.

[15] George ST, Subathra MSP, Sairamya NJ, Susmitha L, Joel Premkumar M. Classification of epileptic EEG signals using PSO based artificial neural network and tunable-Q wavelet transform. Biocybern Biomed Eng Apr 2020;40(2):709–28. https://doi.org/10.1016/j.bbe.2020.02.001.

[16] Caelen O. A Bayesian interpretation of the confusion matrix. Ann Math Artif Intell Dec 2017;81(3–4):429–50. https://doi.org/10.1007/s10472-017-9564-8.

[17] Omidvar M, Zahedi A, Bakhshi H. EEG signal processing for epilepsy seizure detection using 5-level Db4 discrete wavelet transform, GA-based feature selection and ANN/SVM classifiers. J Ambient Intell Hum Comput Nov 2021;12(11):10395–403. https://doi.org/10.1007/s12652-020-02837-8.

[18] Jiao Y, Zhou T, Yao L, Zhou G, Wang X, Zhang Y. Multi-view multi-scale optimization of feature representation for EEG classification improvement. IEEE Trans Neural Syst Rehabil Eng Dec 2020;28(12):2589–97. https://doi.org/10.1109/TNSRE.2020.3040984.

[19] Zhang Y, Wang S, Ji G. A comprehensive survey on particle swarm optimization algorithm and its applications. Math Probl Eng 2015;2015. https://doi.org/10.1155/2015/931256. Hindawi Limited.

[20] Wang D, Tan D, Liu L. Particle swarm optimization algorithm: an overview. Soft Comput Jan 2018;22(2):387–408. https://doi.org/10.1007/s00500-016-2474-6.

[21] Sun W, Su Y, Wu X, Wu X, Zhang Y. EEG denoising through a wide and deep echo state network optimized by UPSO algorithm. Appl Soft Comput Jul 2021;105. https://doi.org/10.1016/j.asoc.2021.107149.

[22] Hakem Ekram, Al-Shammary Dhiah, Ahmed M. Mahdi, Survey analysis for optimization algorithms applied to electroencephalogram. Int J Electr Comput Eng 2023;13(6). https://doi.org/10.11591/ijece.v13i6.pp6891-6903.

[23] Hakem Ekram, Al-Shammary Dhiah, Ahmed M. Mahdi, Survey Analysis on smart features selection for machine learning techniques mainly applied to EEG. J Al-Qadisiyah Comput Sci Math 2023;15(3). https://doi.org/10.29304/jqcm.2023.15.3.1266.