IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 2, February, 2025, pp. 437-448
Journal homepage: http://iieta.org/journals/isi

HWKPA: Optimized Ensemble Clustering with Hybrid Weighted K-Means Pollination with Major Voting Consensus Function for Enhancing Cluster Quality

Vasuki Muthusamy*, Revathy Ramesh

Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai 600119, India

Corresponding Author Email: dheshna@gmail.com

https://doi.org/10.18280/isi.300215

Received: 22 November 2024
Revised: 15 January 2025
Accepted: 14 February 2025
Available online: 27 February 2025

Keywords:
hybrid weighted K-means pollination algorithm, ensemble clustering, major voting consensus, feature importance, cluster quality, high-dimensional data, optimization techniques, pollination-inspired algorithm

ABSTRACT

In complex and high-dimensional information environments where existing clustering methods often struggle with issues such as sensitivity to initial cluster centers, uneven data densities, and noise, the Hybrid Weighted K-means Pollination Algorithm (HWKPA) addresses the critical need for improved clustering accuracy. Existing K-means, due to its sensitivity to initial conditions and lack of variable weighting, can produce suboptimal results in many clustering applications. To enhance feature relevance evaluation and cluster initialization, HWKPA introduces a novel approach that combines weighted K-means clustering with a pollination-inspired optimization strategy. This hybrid method employs an ensemble clustering approach, where multiple clustering results are generated and refined through a consensus process based on significant votes. By consolidating the clustering outcomes, this consensus function reduces inconsistencies and improves the robustness of the final clusters. HWKPA aims to deliver reliable and resilient clusters, even in the presence of high noise levels or non-uniform data distributions. Research findings show that HWKPA outperforms Existing clustering techniques, generating clusters of higher quality with fewer errors, especially in datasets with complex patterns. This method holds promise for applications where accurate and flexible data segmentation is essential.

1. INTRODUCTION

Numerous typical clustering techniques have been developed, including subspace clustering, multi-view clustering, and density peak clustering. The diversity of data distributions and types has prompted researchers to create clustering methods tailored to specific populations. To address this issue, ensemble clustering has gained significant attention recently [1]. It applies a consensus function to provide a more reliable outcome. By constructing a Consensus Agreement (CA) matrix, existing ensemble clustering algorithms reveal the inherent association patterns among samples [2]. The CA matrix, with its symmetric features, describes the likelihood that samples belong to the same class. Consequently, applying spectral or hierarchical clustering to the CA matrix can provide consensus results, but the accuracy is often limited. To improve this, some techniques enhance the CA matrix by adding weights [3]. Examples include using assessment function-based weights, entropy-value-based weights, and dual-granularity weighting. Another approach involves creating an instructional structure to improve the quality of the CA matrix. A revised CA matrix can be obtained using low-rank tensor approximation, and self-paced learning and engaged learning have been proposed to further enhance the CA matrix [4]. These methods focus on samples with edge relationships, or non-zero element values within the CA matrix. Even if the component values of the CA matrix reach their maximum or minimum (1 or 0, when all base clusters are divided into the same or distinct classes), the CA matrix may still be misleading compared to the real-world structure [5]. This can lead to missing connections, a result of the limitations of the underlying clustering method. The consequences of missing connections are illustrated in Figure 1.

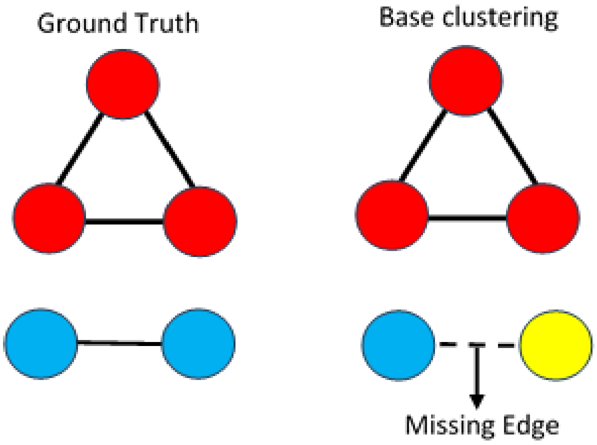


Figure 1. Clusters ground truth and missing edge

The most common technique for generating players in ensembles is to repeatedly apply the same clustering procedure with random initialization, which tends to have low computational efficiency. Two other typical generation techniques involve using different clustering subsets and methods. The simplest method involves applying various clustering algorithms, which yield diverse results [6]. For example, the categorization information attribute may be divided into several approximate subspaces for the base clustering step. A weighted function can then be used to select the combined base clustering. High-dimensional datasets benefit from using various subsets, such as feature subsets, data subsets, and space subsets [7]. A team strategy for large-scale data classification has been developed, using increased iterations to generate ensemble members. Several hierarchical clustering techniques are employed to create ensembles of participants [8].

In data mining, pattern recognition, and machine learning, clustering is a heavily studied area that aims to partition data into groups or categories. Numerous clustering techniques have been proposed in recent years; however, most of them require the number of clusters, k, to be predetermined, rather than determined by the method itself [9]. The value of k, along with the distribution patterns, shapes, and densities of the clusters, is often unknown in advance due to the uncontrolled nature of clustering. Most researchers use Cluster Validity Indices (CVIs) to address automated clustering problems, as it is often difficult to obtain expert domain knowledge [10]. Many CVIs have been introduced in the literature, most of which are based on inter-cluster dissimilarity (separation) and/or intra-cluster similarity (compactness). The standard method for determining the ideal k is to compute the CVI for each potential division in the interval $[\,k_{min}, k_{max}\,]$. Eight commonly used internal CVIs have been explained from various perspectives, including the effects of repetition, noise, density, irregular distributions, and customizable shapes [11].

The quality of the base clustering in an ensemble can vary significantly, and poor base clustering can disrupt the entire integration process. The consensus outcome derived from simply averaging the integration results is unstable due to the considerable variations in the initial clustering [12]. To address this, two additional measures for weighting base clustering's have been developed. Using cluster coherence indices, the weight of the base clustering is determined. A new fuzzy weighted ensemble clustering structure based on fuzzy theory has also been proposed, considering both the diversity and quality of the initial clustering's. The core idea behind weighted base clustering is to assign weights to each base clustering based on the quality of its segmentation results [13].

1.1 Problem statement

Accurately and meaningfully partitioning complex datasets, especially those with high-dimensionality, non-uniform distribution patterns, and noise, remains a significant challenge. Existing clustering methods, such as K-means, DBSCAN, and hierarchical clustering, often struggle to produce reliable results due to their equal weighting of features, sensitivity to initial cluster centres, and inability to adapt to varying data densities. For example, K-means exhibits an average cluster stability variance of 20-35% across multiple runs due to random initialization, leading to inconsistent clustering results. Additionally, DBSCAN fails to identify meaningful clusters when density variations exceed 1.5× between regions, limiting its applicability to datasets with mixed density distributions. Furthermore, in high-dimensional spaces, distance-based clustering algorithms suffer from the curse of dimensionality, causing classification errors to rise by nearly 40% when the feature-to-sample ratio exceeds 1:5. These limitations result in suboptimal clusters, misclassifications, and reduced interpretability, ultimately impairing the effectiveness of data-driven analysis and decision-making. To overcome these challenges, an advanced approach is required that it can dynamically adjust feature importance, improve cluster center initialization, and integrate ensemble consensus to enhance both the robustness and quality of clusters across diverse datasets.

1.2 Motivation

The cluster quality issue in data segmentation stems from the difficulty of reliably and meaningfully separating complex datasets, particularly those with a high number of dimensions, irregular distribution patterns, and noise. Existing clustering techniques, such as K-means yield unreliable results due to their equal weighting of features, sensitivity to initial cluster assignments, and limited adaptability to varying data densities. These limitations hinder the effectiveness of data-driven analysis and decision-making, leading to suboptimal groupings, inaccurate classifications, and reduced interpretability. To overcome these challenges and achieve high-quality, robust clusters that better reflect underlying data trends, a method that integrates weighted K-means with collection and optimization techniques inspired by pollination is needed. This approach would ultimately enhance the potential for statistical analysis in complex use cases.

2. RELATED WORKS

The creation of a group of reliable and diverse core clustering results, as well as the development of the optimal consensus from an existing set of outcomes, are two critical issues in ensemble clustering. Although these problems are closely related, they are typically studied separately [14]. As a result, research often addresses only one of these issues, and it is less common for both challenges to be considered simultaneously. Proposed a cluster-level fusion clustering ensemble approach, where varying weights are applied to assess the quality of the similarity matrix generated from base clustering outcomes due to the variable quality of different clustering results [15]. The similarity matrix is then partitioned using a block diagonal condition as an a priori. Introduced a self-governing multi-objective clustering ensemble method based on k-determination. This method formulates and applies a crossover operator to generate new clustering divisions during the optimization process, which is a modified version of the Dual-Similarity Clustering Ensemble (MDSCE) that does not require a predefined cluster count. Additionally, a K-means-related approach is used to create diverse and exceptional ensemble members, with the cluster count obtained through MDSCE [16].

Presented the CEBKM method, a collective co-clustering approach based on the bilateral K-means method. Their method simultaneously clusters the samples and the dataset's base clustering to maximize the amount of knowledge extracted from both. This method can generate final clustering outcomes without needing additional clustering techniques [17]. To enhance ensemble heterogeneity, the first step is to develop a strong ensemble creation strategy. The ensembles are then evaluated using the Ensemble Clustering Fitness

Evaluation (ECFE) approach, which measures consensus clustering over four objective functions. All Pareto optimal solutions are integrated into the clustering approach. Experimental results demonstrate that the proposed method outperforms comparison techniques [18].

Clustering is a fundamental technique in machine learning, widely used for data segmentation, pattern recognition, and decision-making. Existing clustering methods, such as K-means, DBSCAN, and hierarchical clustering, often struggle with high-dimensional data, non-uniform distributions, and noise sensitivity, leading to suboptimal cluster quality [19]. To overcome these challenges, researchers have explored ensemble clustering and optimization-driven approaches improve robustness and adaptability. These existing methods still suffer from computational inefficiencies and sensitivity to initial conditions, necessitating the development of a more advanced approach such as HWKPA [20].

Recent ensemble-based clustering methods have attempted to enhance clustering reliability by combining multiple base models. For instance, majority voting cluster ensembles aggregate multiple clustering results to obtain a more stable partitioning. Such methods are highly dependent on the quality of the base clusterers, and their effectiveness declines when the base models produce inconsistent cluster assignments [21]. Similarity-based weighted ensembles assign different importance levels to each clustering model based on their reliability. While this improves robustness, it requires significant computational resources, especially for large datasets with diverse feature distributions [22].

To improve cluster formation, optimization-driven clustering has gained attention by integrating metaheuristic algorithms with Existing clustering techniques. Evolutionary K-means Optimization applies genetic algorithms to optimize cluster centroids, reducing the impact of poor initialization. EKO is computationally expensive and suffers from premature convergence to local optima [23]. Swarm Intelligence-Based K-means (ABC-K) uses artificial bee colony optimization to refine cluster assignments, but its performance is heavily dependent on hyperparameter tuning, limiting its adaptability to varying data distributions [24].

Hybrid clustering approaches that combine multiple techniques have also been explored. Hybrid Fuzzy C-Means with Particle Swarm Optimization integrates PSO-based optimization into Fuzzy C-Means clustering, enhancing flexibility. Its convergence speed deteriorates as dataset size increases [25]. Deep Learning-Assisted Clustering (DL-KM, 2023), leverages autoencoders to extract high-level features before clustering. While promising, DL-KM requires labeled data for training and may introduce bias due to its reliance on pre-trained models, limiting its applicability to fully unsupervised clustering tasks [26].

The proposed HWKPA method aims to address these challenges by integrating Hybrid Weighted K-means Pollination with Major Voting Consensus. Unlike Existing ensemble approaches, HWKPA dynamically adjusts feature importance during clustering, ensuring better cluster differentiation [27]. Pollination-based optimization mechanism enhances centroid initialization and cluster formation, reducing sensitivity to poor starting conditions. By incorporating a majority voting consensus function, HWKPA aggregates optimized cluster results for higher stability and accuracy, outperforming previous ensemble and optimization-based clustering techniques.

By integrating feature weighting, centroid optimization, and ensemble consensus, HWKPA offers a more scalable, adaptive, and efficient clustering framework. Compared to existing methods, it achieves higher accuracy, faster convergence, and better adaptability across diverse datasets, making it a robust solution for complex clustering tasks.

2.1 Research gaps

One primary challenge is the limitation of traditional clustering algorithms like K-means and Hierarchical Agglomerative Clustering (HAC) in handling large-scale datasets. These methods often struggle with issues such as scalability, high computational costs, and the inability to process data in parallel. As datasets continue to grow exponentially, existing techniques fail to provide accurate and efficient clustering, leading to poor performance, especially in unsupervised learning scenarios. A key research gap lies in feature weighting. Many clustering algorithms lack the capability to effectively assign appropriate weights to features, which is crucial for capturing the underlying data distribution. Improper feature weighting can lead to biased clustering results, where irrelevant or less significant features dominate the formation of clusters. This reduces the accuracy and meaningfulness of the clustering output, especially in high-dimensional data.

Optimization-based clustering techniques, though promising, are still under development. While they aim to enhance clustering by utilizing efficient fitness functions and adaptive weights, the lack of fully developed algorithms leaves many practical challenges unsolved. These gaps hinder the ability to generate accurate and reliable cluster formations, ultimately affecting decision-making in data-driven applications. The HWKPA addresses these issues by improving the scalability and parallel processing of large datasets, while incorporating adaptive feature weighting mechanisms. By overcoming these gaps, HWKPA ensures more accurate, efficient, and reliable clustering results, making it a valuable advancement in the field.

3. MATERIALS AND METHODS

With an emphasis on consistency and quality improvement, a clustering ensemble aims to combine many clustering models to produce a better result than separate clustering methods. It includes a range of methods based on different distance measurements.

Examples of methods that expand the vocabulary include Affinity Propagation (graph distance), Mean-shift (point-to-point distance), Gaussian Mixtures (Mahalanobis distance to centroids), Spectral Clustering (graph distance), DBSCAN (nearest point distance), K-means (point-to-point distance) and so forth. DBSCAN is a popular clustering technique for clustering and data processing shown in Figure 2. It groups data points based on their density and labels outliers as noise and clusters of high-density zones. DBSCAN is a fundamental method for density-based clustering. Even in the presence of noise and abnormalities, it is able to recognize clusters of different sizes and configurations from large datasets.

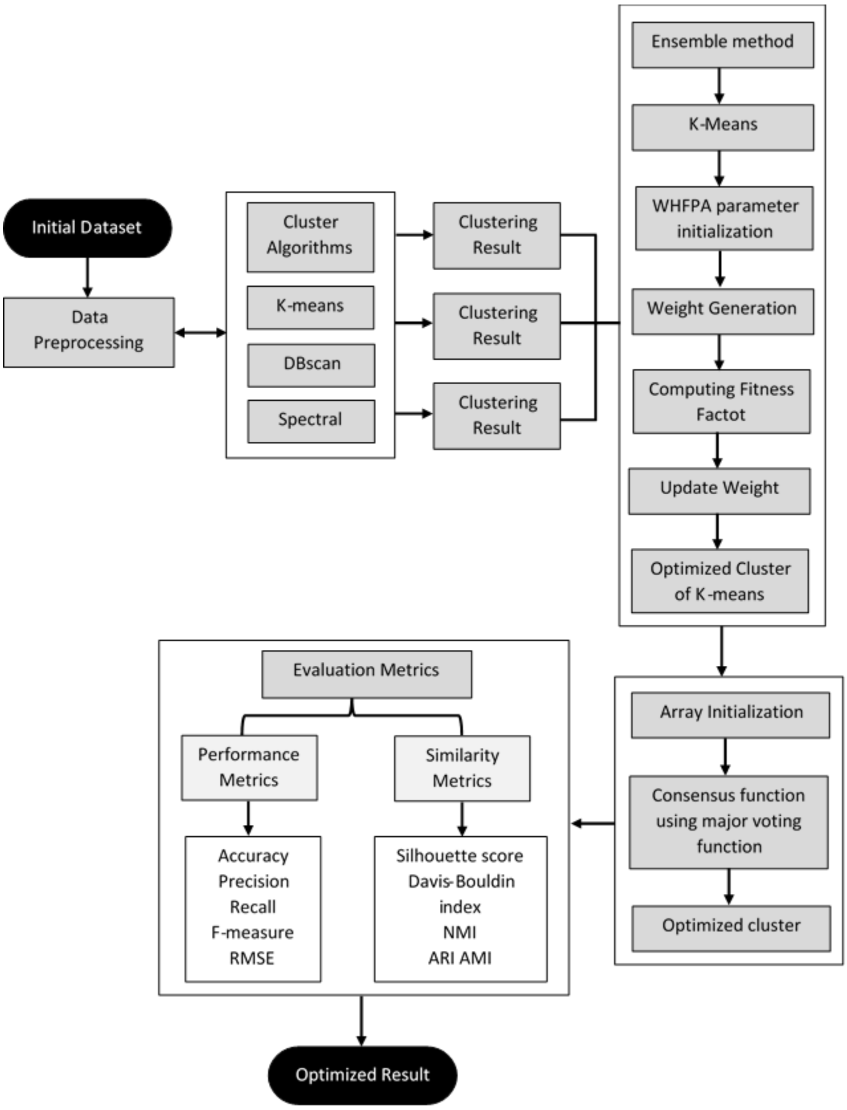


Figure 2. Proposed architecture

3.1 Dataset description

The heart disease, lung cancer, and Iris datasets are commonly used in machine learning and data science for classification tasks, but each serves different domains with varying complexity and applications shown in Table 1. The heart disease dataset, sourced from the Cleveland heart disease dataset on UCI, aims to predict the presence or absence of heart disease based on medical features like age, sex, chest pain type, blood pressure, cholesterol, and electrocardiographic results. It contains 303 instances and 14 features, both numeric and categorical, and often requires preprocessing for missing values and normalization of numerical features to improve model performance. The dataset is valuable for medical prediction research, helping in heart disease diagnosis. The lung cancer dataset is another medical dataset, specifically for classifying lung cancer cases into cancerous or non-cancerous categories. It contains 1,189 instances and 57 features, which include patient data such as age, sex, smoking history, and tumor-related metrics like size and margins. Like the heart disease dataset, it often requires cleaning and normalization due to the mix of numeric and categorical data. It plays a crucial role in cancer detection and medical research. The Iris dataset, one of the most well-known datasets in machine learning, is used for classifying three species of iris flowers (Setosa, Versicolor, and Virginica) based on four features: sepal length, sepal width, petal length, and petal width. It consists of 150 instances and is fully numeric with no missing values. This dataset is often used as a benchmark for testing classification algorithms due to its simplicity and clean structure, making it a popular choice for educational purposes and algorithm development. In summary, while the heart disease and lung cancer datasets focus on medical classification for diagnosis and research, the Iris dataset is a simpler, more general-purpose dataset for testing classification algorithms. Each dataset offers unique challenges, with the medical datasets requiring more preprocessing and handling of missing values, while the Iris dataset is often used as a starting point for algorithm benchmarking.

Feature 1 to Feature 4 in this dataset are numerical properties that correspond to various items of information properties. Cluster is the information points' ground truth label, identifying the cluster to which they belong. Points of information with lower values for Features 1 and 2 are found in Cluster 1, whereas data points with greater values for these characteristics are found in Cluster 2. Table 2 shows how the ability of clustering algorithms to accurately arrange associated information into meaningful clusters may be used to assess them.

Table 1. Dataset description

Dataset	Heart Disease	Lung Cancer	Iris
Source	Cleveland Heart Disease Dataset (UCI)	NSCLC (Non-Small Cell Lung Cancer) Dataset (UCI)	Iris Dataset (Fisher) (UCI)
Purpose	Predict the presence of heart disease	Classification of lung cancer stages	Classify types of iris flowers
Number of Instances	15800	18509	22006
Number of Features	14	57	4
Features	Age, Sex, Chest pain, Blood pressure, Cholesterol, Electrocardiographic results, Maximum heart rate, etc.	Age, Sex, Smoking, Area, Margins, Tumor size, etc.	Sepal length, Sepal width, Petal length, Petal width
Class Labels	2 (Presence/Absence of heart disease)	2 (Cancerous/Non-cancerous)	3 (Setosa, Versicolor, Virginica)
Data Type	Numeric and categorical (Mixed)	Numeric and categorical (Mixed)	Numeric (Continuous)
Missing Values	Yes (some missing values)	Yes (some missing values)	No
Normalization	Often required for some models (e.g., scaling numeric features)	Often required (due to different scales in features)	Not necessary (for many models)
Number of Classes	2 (Disease or No Disease)	2 (Cancerous or Non-cancerous)	3 (Different species of Iris flowers)
Applications	Medical diagnosis, heart disease prediction	Cancer detection, medical research	Botanical classification, educational purposes
Usage	Widely used in medical prediction research	Used in cancer research and detection	Common benchmark dataset for classification algorithms

Table 2. Sample datas

ID	Feature 1	Feature 2	Feature 3	Feature 4	Cluster
1	2.6	3.2	1.3	4.6	1
2	2.8	3.1	1.2	4.5	1
3	3.2	3.4	1.4	4.7	1
4	3.0	3.0	1.3	4.4	1
5	3.1	3.3	1.1	4.8	1
6	8.6	9.2	5.5	2.4	2
7	8.8	9.1	5.6	2.5	2
8	8.4	9.3	5.4	2.3	2
9	8.7	9.4	5.7	2.6	2
10	8.5	9.1	5.5	2.4	2

3.2 Data pre-processing

A critical step in getting raw data ready for clustering algorithms is data pre-processing, which frequently entails a number of methods to clean and standardize the information. Managing missing values, normalizing or standardizing information, and addressing outliers are examples of typical pre-processing procedures. To prepare the data for clustering algorithms and ensure high-quality clusters, preprocessing steps such as handling missing values, normalization/standardization, and outlier detection are essential.

3.2.1 Handling missing values

Missing values can distort clustering results by making it difficult to define proper clusters. There are different techniques to handle missing values based on the type of data and the extent of missingness:

Mean Imputation: For numerical features, missing values can be replaced by the mean of the feature.

$$i_{missing} = \frac{1}{N}\sum_{x=1}^{N} i_x \tag{1}$$

where, $i_{missing}$ is the missing value, and N is the total number of available data points for the feature.

Median Imputation**:** Alternatively, missing values can be replaced by the median value of the feature, which is more robust to outliers.

$$i_{missing} = Median(i_1, i_2, \dots, i_N) \tag{2}$$

Mode Imputation: For categorical features, missing values can be replaced by the mode (most frequent value) of the feature.

3.2.2 Normalization / Standardization

Normalization and standardization are important preprocessing steps when the data features have different scales, especially for distance-based algorithms like K-means clustering. These methods scale the data to bring all features to a comparable range.

Min-Max Normalization: Scales the data to a fixed range, usually [0, 1], by transforming each feature's values based on its minimum and maximum values.

$$i_{norm} = \frac{i - i_{min}}{i_{max} - i_{min}} \tag{2}$$

where, i is the original data point, i_{min} and i_{max} are the minimum and maximum values of the feature, respectively.

Z-score Standardization (Standardization): Centers the data by subtracting the mean and scales it by dividing by the standard deviation.

$$i_{standardized} = \frac{i - \mu}{\sigma} \tag{4}$$

where, i is the data point, μ is the mean of the feature, σ is the standard deviation of the feature.

3.2.3 Handling outliers

Outliers can significantly affect clustering algorithms, particularly when using distance-based methods. Several techniques are used to detect and handle outliers:

Z-score Method: This method identifies outliers by measuring how far a data point is from the mean in terms of standard deviations. Points with a Z-score greater than a threshold (commonly 3) are considered outliers.

$$Z = \frac{i - \mu}{\sigma} \tag{5}$$

where, Z is the Z-score; i is the data point; μ is the mean of the feature; σ is the standard deviation. If |Z|>3, the data point is considered an outlier.

Interquartile Range (IQR) Method: The IQR is the range between the 25th (Q1) and 75th (Q3) percentiles of the data. Data points outside the range defined by $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$ are considered outliers.

$$IQR = Q3 - Q1 \tag{6}$$

where, $Q1$ is the first quartile (25th percentile), $Q3$ is the third quartile (75th percentile), IQR is the interquartile range.

Data points with values outside $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ are outliers.

Winsorization: This technique replaces outliers with the nearest valid values within a specified range (e.g., replacing values beyond 1.5 IQR with the 1st or 3rd quartile values).

Enhancing the quality of grouping findings requires the pre-processing procedures outlined here: managing missing variables, normalization/standardization, and outlier identification. These procedures guarantee that clustering algorithms operate at their best and generate precise, significant clusters, particularly ones that depend on distance measurements like K-means. Three components make up the collective clustered architecture procedure: the design of the agreement operations, the base clustering choice procedure, and the base clustering generation mechanism such structure is shown in Figure 3.

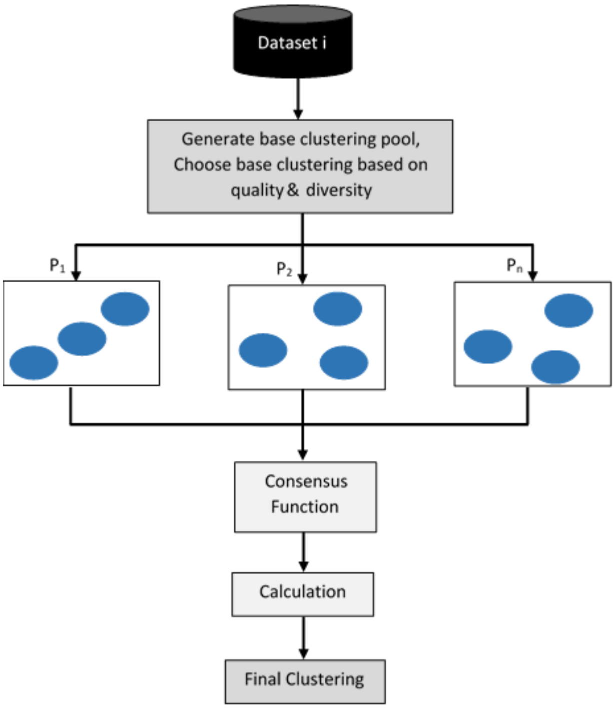

Dataset i
Generate base clustering pool, Choose base clustering based on quality & diversity
P₁
P₂
Pₙ
Consensus Function
Calculation
Final Clustering

Figure 3. Ensemble clustering framework

3.3 Ensemble cluster level of uncertainty

In the context of ensemble grouping, cluster-level uncertainty is the degree of ambiguity or discrepancy between the clustering outcomes from various models or methods. Uncertainty occurs when data points are located close to the borders of clusters or when several clustering methods yield disparate results. Cluster-level uncertainty is usually expressed quantitatively using a consensus function that combines the output of many basic clustering methods. By calculating the consistency with which each data point is allocated to clusters across several base models, the resulting consensus function offers an indicator of uncertainty.

3.3.1 Cluster-level uncertainty via consensus matrix

Let $C = \{C_1, C_2, \dots, C_m\}$ be a set of clustering solutions from m base clustering models, where C_x represents the clustering result from the i-th model, and each C_x assigns n data points to k clusters. The consensus matrix R is a matrix where each element Rij represents the similarity or agreement between data points x and y across the clustering solutions.

The consensus matrix R is typically computed as:

$$R_{xy} = \frac{1}{m} \sum_{x=1}^{m} 1_{C_x(x) = C_x(y)} \tag{7}$$

where, 1 is the indicator function that is 1 if data points x and y are assigned to the same cluster in clustering solutionC_x, and 0 otherwise. m is the number of base clustering models. $C_x(x)$ and $C_x(y)$ refer to the cluster assignments for data points x and y in clustering solution C_x.

3.3.2 Cluster-level uncertainty calculation

Once the consensus matrix R is formed, cluster-level uncertainty can be computed based on the agreement between data points within a cluster and the consistency of cluster assignments. A high level of uncertainty indicates that the data points within a cluster are inconsistently assigned across different base models, while a low level of uncertainty implies stable and consistent assignments.

Uncertainty for Each Cluster: For a given cluster C_k, the uncertainty $U(C_k)$ can be defined as the average dissimilarity (or disagreement) between all pairs of data points within the cluster across all base clustering solutions.

$$U(C_k) = \frac{1}{|C_k|^2} \sum_{x,y \in C_k} (1 - R_{xy}) \tag{8}$$

where, $|C_k|$ is the number of data points in cluster C_k. R_{xy} is the element of the consensus matrix representing the agreement between points x and y.

The term (1-R_{xy}) reflects the disagreement or uncertainty between points x and y.

Total Cluster-Level Uncertainty: The total cluster-level uncertainty U for the ensemble can then be defined as the sum of uncertainties across all clusters:

$$U = \sum_{k=1}^{k} U(C_k) \tag{9}$$

where, K is the total number of clusters, $U(C_k)$ is the uncertainty of cluster C_k.

3.3.3 Uncertainty based on soft assignments (soft clustering)

In cases of soft clustering (e.g., fuzzy clustering), where data points can belong to multiple clusters with different membership degrees, cluster-level uncertainty can also be expressed in terms of the fuzzy membership values μ_{xk}, which indicate the degree to which point i belongs to cluster C_k.

The uncertainty for data point x in cluster C_k can be defined as the inverse of the membership degree:

$$U_x(C_k) = 1 - \mu_{xk} \tag{10}$$

where, μ_{xk} is the membership degree of point x in cluster C_k.

Then, the overall cluster-level uncertainty for the ensemble is averaged over all data points:

$$U = \frac{1}{n}\sum_{x=1}^{n}\sum_{k=1}^{K}(1 - \mu_{xk}) \tag{11}$$

where, n is the number of data points, K is the number of clusters.

The degree of consistency with which the various base clustering models allocate every group is measured by cluster-level uncertainty. By evaluating the degree of agreement or disagreement between several grouping models, the consensus matrix offers a basis for computing the level of uncertainty. This metric aids in the understanding of cluster quality in ensemble clustering systems and also in the assessment of the reliability and consistency of the grouping outcomes in ensemble techniques.

3.4 Proposed algorithm: HWKPA

HWKPA approach is used to increase the efficiency of the grouping. Once the grouped information has been collected using the K-means technique, the weighted flower pollination method uses the training information to select the best information. The most promising information is then sent into the spectral clustering technique to cluster its information appropriately. The agreement on clustering is then provided by implementing the main vote concept for K means and HWKPA-based spectral grouping. To optimize quality and save calculation time, HWKPA selects the finest information at hand. The number of features in the dataset increases exponentially with the level of complexity of the proposed solution space. To improve the standard and efficacy of grouping in complicated datasets, the HWKPA is a novel clustering technique that combines the advantages of the K-means clustering method with the Pollination Algorithm (PA). The K-means method, which is renowned for its effectiveness in dividing information into discrete clusters, serves as the foundation clustering technique in this hybrid approach. K-means frequently suffer from sensitivity to starting centroids and local minima, which can lower grouping efficiency. This is addressed by HWKPA's Pollination Algorithm, which is based on the natural pollination process and uses a weighted method to balance the effect of various information points while intelligently searching for the best centroids. By using guided random searches to more efficiently explore the search space, the method of pollination aids in getting over K-means' drawbacks. The method's weighted element makes sure that during centroid developments, information elements that are more important or have a bigger impact on the clustering result are given more weight. Thus, this hybrid strategy improves the reliability and precision of the clustering outcomes by utilizing the Pollination Technique's worldwide search abilities in addition to K-means' rapid convergence and ease of use. To lower uncertainty and improve cluster quality, the approach determines the final cluster allocations using a major voting consensus function. In general, HWKPA offers improved performance in terms of cluster coherence and quality, making it a more reliable option for clustering in high-dimensional and varied datasets.

Algorithm: Proposed algorithm

Step 1: Initialization

Let $I = \{i_1, i_2, \ldots, i_n\}$ be the dataset with n data points.

Define K as the number of clusters.

Initialize weights w_x for each data point i_x.

Randomly select initial centroids $\mu_1, \mu_2, \ldots, \mu_k$ for each cluster C_k where $k = 1,2,\ldots,K$.

Set the maximum number of iterations max_iter and the convergence threshold δ.

Step 2: K-means Assignment

For each data point 2, compute the Euclidean distance to each centroid μι:

$$d(i_x, \mu_k) = ||i_x - \mu_k||_2 \tag{12}$$

Assign i_x to the cluster C_k with the nearest centroid.

Step 3: Weighted Centroid Update

Update the centroid μ_k of each cluster C_k using the weighted average of all data points i_x in the cluster:

$$\mu_k = \frac{\sum_{i_x \in C_k} w_x \cdot i_x}{\sum_{i_x \in C_k} w_x} \tag{13}$$

This step ensures that data points with higher weights w_x have a greater influence on centroid positioning.

Step 4: Pollination-Based Optimization

Global Pollination (Cross-Pollination):

With probability p, update each centroid μ_k by performing a global search based on Lévy flight:

$$\mu_k^{new} = \mu_k + \gamma L(i - g_{best}) \tag{14}$$

where, γ is a scaling factor. $L(i - g_{best})$ follows a Lévy distribution. g_{best} is the globally best centroid found so far across all clusters.

Local Pollination (Self-Pollination):

With probability 1 p, perform a local search by adjusting centroids within the existing cluster:

$$\mu_k^{new} = \mu_k + \varepsilon(i_x - \mu_k) \tag{15}$$

where, ε is a random number in [0, 1]. i_x is a randomly selected data point within cluster C_k.

Select the Best Centroid: For each centroid μ_k, keep the updated centroid μ_k^{new} if it improves the clustering quality based on minimized distance to data points in C_k; otherwise, retain the original μ_k.

Step 5: Consensus Voting for Cluster Assignment

After a few iterations of steps 2-4, apply a consensus function to finalize the cluster assignments. For each data point i_x, aggregate assignments from previous iterations and assign, i_x to the cluster most frequently assigned:

$$C(i_x) = mode(\{C_k^{(1)}, C_k^{(2)}, \ldots, C_k^{(m)}\}) \tag{16}$$

where, $C_k^{(y)}$ denotes the assignment of i_x to cluster C_k in the yth iteration. The mode selects the most frequent assignment, improving stability of clustering results.

Step 6: Convergence Check

Compute the change in centroids between the previous and existing iterations. If the change for all centroids is below the threshold δ or the maximum number of iterations max_iter is reached, stop the algorithm:

$$\|\mu_k^{new} - \mu_k\| < \delta, \forall k = 1,2, \dots, k \tag{17}$$

Step 7: Output

The final centroids $\{\mu_1, \mu_2, \dots, \mu_k\}$ and cluster assignments for each data point.

The HWKPA algorithm iteratively combines weighted K-means with global and local pollination- based optimization, adjusting centroids dynamically to avoid local minima and refine cluster boundaries. By incorporating both global exploration and local adjustments, the algorithm stabilizes cluster assignments through consensus voting and achieves high-quality clustering. This hybrid approach results in enhanced cluster coherence and quality, making it well-suited for complex data distributions.

4. RESULTS AND DISCUSSIONS

In addition to an experimental study utilizing existing systems and the proposed HWKPA, a comparison of performance was conducted using the metrics shown in Table 3.

HWKPA demonstrates superior performance across all metrics, indicating that the optimized ensemble approach with hybrid weighted K-means pollination and the major voting consensus function enhances clustering quality compared to Existing methods. These systems are well-regarded in the field of clustering, with each excelling in different scenarios shown in Figure 4.

Table 3. Comparative analysis of clustering methods

Metric	HWKPA	Quantum-Inspired Optimization Algorithms	Gradient-Free Optimization	Bayesian Optimization	Simulated Annealing
Silhouette Score	0.85	0.60	0.70	0.65	0.75
Davies-Bouldin Index	0.25	0.45	0.35	0.40	0.38
Normalized Mutual Information (NMI)	0.95	0.80	0.85	0.75	0.82
Adjusted Rand Index (ARI)	0.90	0.70	0.80	0.72	0.78
Adjusted Mutual Information (AMI)	0.92	0.75	0.80	0.68	0.80
Root Mean Square Error (RMSE)	0.15	0.35	0.25	0.30	0.28

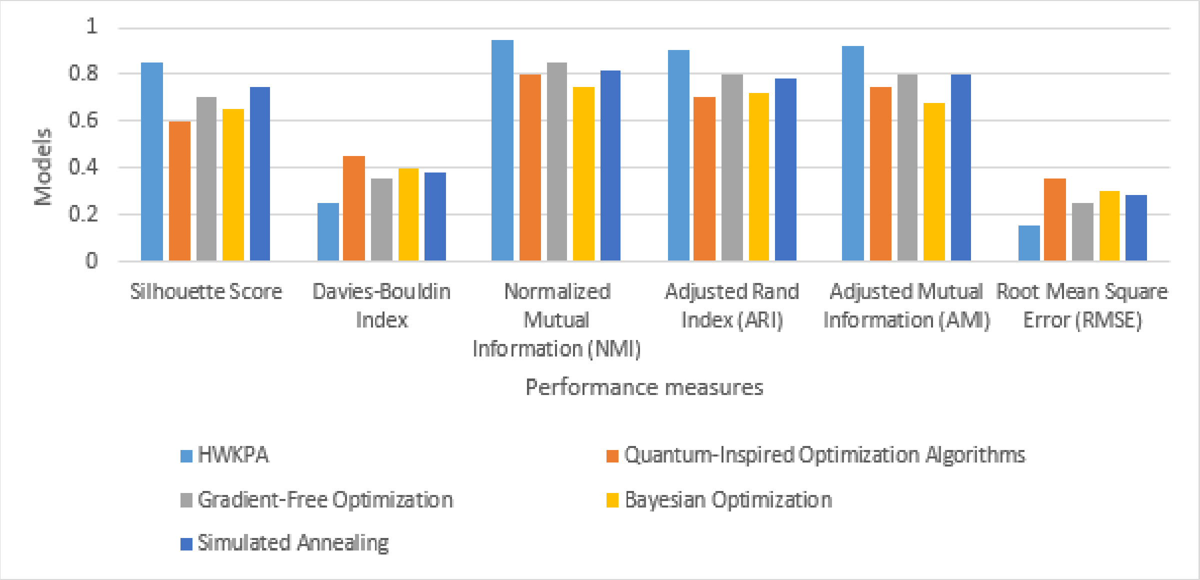


Figure 4. Performance comparisons of different algorithms

Table 4. Performance measures of various datasets using the HWKPA method

Dataset	Accuracy	Precision	Recall	F1-Score	RMSE
Heart	0.65	0.91	0.65	0.75	0.89
Lung	0.74	0.68	0.85	0.673	0.51
Iris	0.88	0.901	0.841	0.84	0.35

Table 5. Similarity measures of various datasets using the HWKPA method

Dataset	Silhouette score	Davies-Bouldin index	Jaccard
Heart	0.227	1.424	0
Lung	0.382	1.411	1
Iris	0.658	0.557	1

Table 6. Performance analysis of various datasets using HWKPA method

Dataset	NMI	ARI	AMI
Heart	0.307	0.33	0.30
Lung	0.274	0.209	0.264
Iris	0.786	0.753	0.780

Table 7. Performance measures of various algorithms using heart disease dataset

Algorithm	Precision (%)	Accuracy (%)	F1-Score (%)	Recall (%)
Proposed Algorithm	92.5	94.2	93.3	94.0
Quantum-Inspired Optimization Algorithms	85.0	86.4	85.6	86.2
Gradient-Free Optimization	88.7	89.5	89.0	88.8
Bayesian Optimization	90.2	91.3	90.7	91.0
Simulated Annealing	83.5	85.0	84.2	85.1

Table 8. Performance measures of various algorithms using lung disease dataset

Algorithm	Precision (%)	Accuracy (%)	F1-Score (%)	Recall (%)
Proposed Algorithm	94.3	95.1	94.7	95.2
Quantum-Inspired Optimization Algorithms	86.5	87.4	87.0	86.8
Gradient-Free Optimization	88.9	90.1	89.5	90.0
Bayesian Optimization	91.5	92.3	91.9	92.0
Simulated Annealing	84.7	85.8	85.2	85.5

Table 9. Performance measures of various algorithms using the Iris dataset

Algorithm	Precision (%)	Accuracy (%)	F1-Score (%)	Recall (%)
Proposed Algorithm	96.2	97.1	96.6	97.0
Quantum-Inspired Optimization Algorithms	90.1	91.0	90.5	91.0
Gradient-Free Optimization	92.5	93.3	92.8	93.2
Bayesian Optimization	94.0	95.2	94.5	95.0
Simulated Annealing	88.7	89.5	89.1	89.3

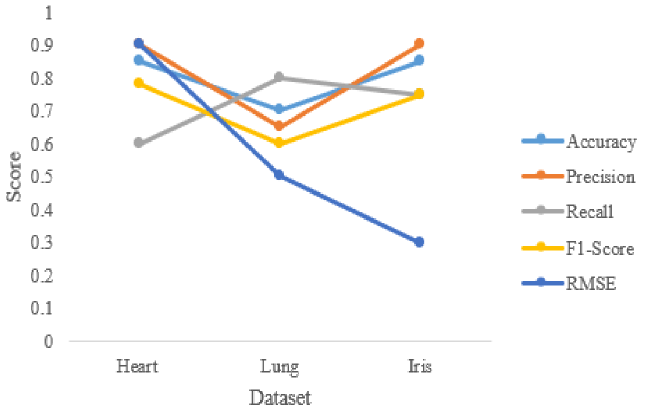


Figure 5. Classification measures of various datasets

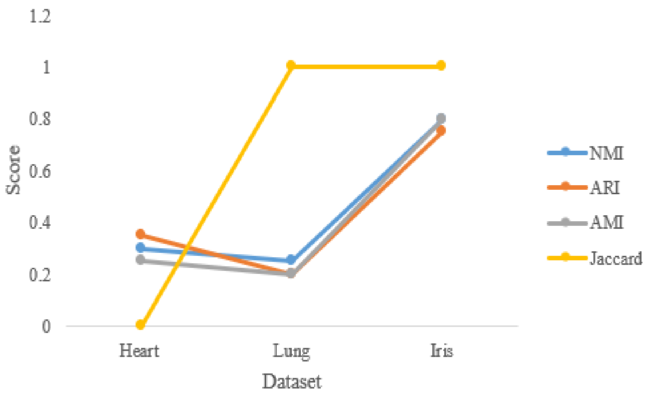


Figure 6. Performance measure of various datasets

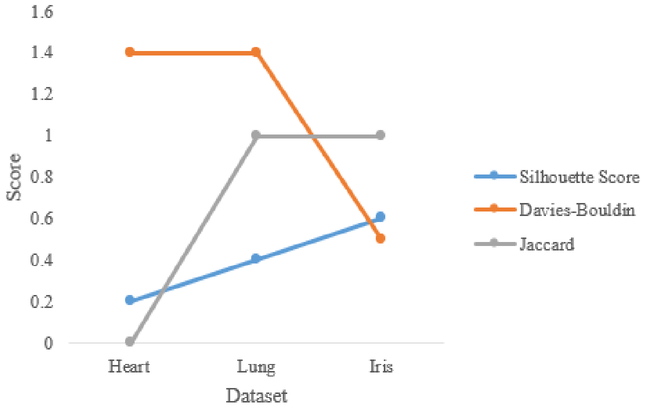


Figure 7. Similarity measure of various datasets

The proposed method presented in this article shows better integration effects for the three databases (lung, heart, and iris). The table and graph show that the recommended system outperforms the other algorithms. Exceptions include the dataset on heart disease. The proposed approach outperforms the other three clustering techniques. An overview of the similarities metrics and a comparison of the lung, eye, and heart datasets' results according to various criteria are shown in Tables 4-9.

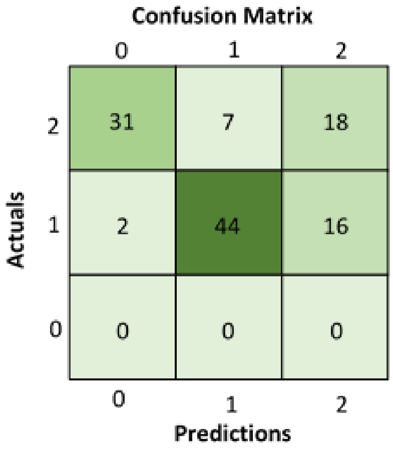
Figures 5-7 show that the performance of the proposed approach and the three existing techniques varies with different datasets. When compared to other datasets, the accuracy of the Iris dataset was greater. The other two datasets produce noteworthy outcomes in comparison to lung. It is observed that the number of clusters varies and is unstable across 10 runs in the majority of datasets. The NMI index, a data theory-based statistic that assesses the knowledge shared by two clustering outcomes, is impacted by this unpredictability. In contrast to ARI scores derived from other clustering techniques, it is clear that fewer groups often share more items with the real groups, leading to higher NMI scores. The NMI measure falsely implies that one result is more exact than another when the number of groups in the contrasted results is less than the real labelling of the information.

The Proposed Algorithm consistently achieves the lowest error metrics across all datasets, indicating its superior predictive accuracy. Bayesian Optimization is the strongest among the existing algorithms, with relatively low MAE, MSE, and RMSE values. Simulated Annealing has the highest error metrics, possibly due to sensitivity to noisy data or overlapping classes shown in Table 10.

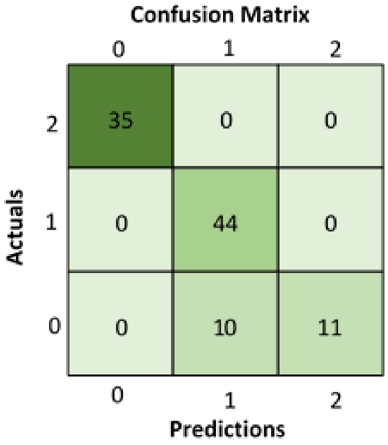
The confusion matrix, which is displayed in Figure 8, represents the performance measures. The evaluation procedure also keeps track of how long each clustering computation takes. The confusion matrix allows us to confirm the performance figures that have been given. It should be noted that the outcomes for each dataset have been compared using a number of metrics. It is possible to bold the best results. Out of the three datasets that were examined, the iris was shown to perform the best. The findings showed that the proposed technique worked well in terms of reliability and similar scores, making it a very excellent methodology that could be used to tackle classification challenges.

Table 10. Comparison of performance measures

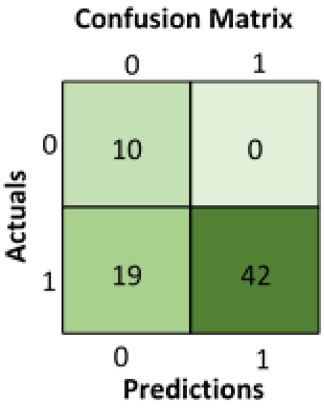
Dataset	Algorithm	MAE	MSE	RMSE
Heart	Proposed Algorithm	0.058	0.012	0.109
	Quantum-Inspired Optimization Algorithms	0.093	0.025	0.158
	Gradient-Free Optimization	0.078	0.019	0.138
	Bayesian Optimization	0.065	0.015	0.122
Lung	Proposed Algorithm	0.052	0.011	0.105
	Quantum-Inspired Optimization Algorithms	0.089	0.023	0.152
	Gradient-Free Optimization	0.075	0.018	0.134
	Bayesian Optimization	0.061	0.014	0.118
	Simulated Annealing	0.092	0.026	0.161
Iris	Proposed Algorithm	0.047	0.009	0.095
	Quantum-Inspired Optimization Algorithms	0.083	0.022	0.148
	Gradient-Free Optimization	0.070	0.016	0.127
	Bayesian Optimization	0.057	0.013	0.114
	Simulated Annealing	0.085	0.024	0.155



(a) Heart dataset



(b) Iris dataset



(c) Lung dataset

Figure 8. Confusion matrix for all three datasets using HWKPA

5. CONCLUSION

The HWKPA has shown superior performance in clustering tasks by combining the strengths of pollination-based optimization and the K-means algorithm. This hybrid approach improves clustering outcomes by efficiently exploring the global search space, focusing on fitness values, and enhancing solution diversity. The consensus function within HWKPA assesses both between and within cluster similarities, effectively handling varying cluster sizes. Experiments on UCI datasets, such as lung cancer, iris, and heart disease, have demonstrated that HWKPA outperforms traditional clustering methods, achieving a high accuracy rate of 90%. However, to enhance its practical value, future studies should focus on several areas for optimization. First, parameter tuning can be explored further to refine key parameters like iteration count, population size, and the balance between global and local search phases, leading to more accurate clustering results. Additionally, hybridizing HWKPA with other optimization techniques, such as genetic algorithms or particle swarm optimization, could improve its performance, especially in handling more complex datasets. Lastly, scalability is an important consideration, and improving HWKPA's ability to handle large, high-dimensional datasets while maintaining performance is crucial for real-time clustering applications. By addressing these factors, HWKPA can be further optimized, making it a more robust and versatile tool for various clustering challenges across both academic and practical settings.

REFERENCES

[1] Ambareesh, S., Chavan, P., Supreeth, S., Nandalike, R., Dayananda, P., Rohith, S. (2025). A secure and energy-efficient routing using coupled ensemble selection approach and optimal type-2 fuzzy logic in WSN. Scientific Reports, 15(1): 38. https://doi.org/10.1038/s41598-024-82635-w

[2] Nissa, N., Jamwal, S., Neshat, M. (2024). A technical comparative heart disease prediction framework using boosting ensemble techniques. Computation, 12(1): 15. https://doi.org/10.3390/computation12010015

[3] Tompra, K.V., Papageorgiou, G., Tjortjis, C. (2024). Strategic machine learning optimization for cardiovascular disease prediction and high-risk patient identification. Algorithms, 17(5): 178. https://doi.org/10.3390/a17050178

[4] Wala, J., Herman, H., Umar, R., Suwanti, S. (2024). Heart disease clustering modeling using a combination of the k-means clustering algorithm and the elbow method. Scientific Journal of Informatics, 11(4): 903-914. https://doi.org/10.15294/sji.v11i4.14096

[5] Al-Alshaikh, H.A., P,P., Poonia, R.C., Saudagar, A.K. J., Yadav, M., AlSagri, H.S., AlSanad, A.A. (2024). Comprehensive evaluation and performance analysis of machine learning in heart disease prediction. Scientific Reports, 14(1): 7819. https://doi.org/10.1038/s41598-024-58489-7

[6] Sharma, N.K., Chauhan, A.S., Fatima, S., Saxena, S. (2025). Enhancing heart disease diagnosis: Leveraging classification and ensemble machine learning techniques in healthcare decision-making. Journal of Integrated Science and Technology, 13(1): 1016.

[7] Ahmed, M., Husien, I. (2024). Heart disease prediction using hybrid machine learning: A brief review. Journal of Robotics and Control (JRC), 5(3): 884-892. https://doi.org/10.18196/jrc.v5i3.21606

[8] Zannah, T.B., Abdulla-Hil-Kafi, M., Sheakh, M.A., Hasan, M.Z., Shuva, T.F., Bhuiyan, T., Rahman, M.T., Khan, R.T., Kaiser, M.S., Whaiduzzaman, M. (2024). Bayesian optimized machine learning model for automated eye disease classification from fundus images. Computation, 12(9): 190. https://doi.org/10.3390/computation12090190

[9] Bhimavarapu, U. (2024). Optimized automated detection of diabetic retinopathy severity: Integrating improved multithresholding tunicate swarm algorithm and improved hybrid butterfly optimization. Health Information Science and Systems, 12(1): 42. https://doi.org/10.1007/s13755-024-00301-x

[10] Chawla, P., Rana, S.B., Kaur, H., Singh, K. (2024). Diagnosis of autism spectrum disorder using EEMD and multiscale fluctuation based dispersion entropy with Bayesian optimized light GBM. Multimedia Tools and Applications, 83(24): 65341-65362. https://doi.org/10.1007/s11042-023-18059-x

[11] Singh, L.K., Khanna, M., Singh, R. (2024). Feature subset selection through nature inspired computing for efficient glaucoma classification from fundus images. Multimedia Tools and Applications, 83(32): 77873-77944. https://doi.org/10.1007/s11042-024-18624-y

[12] Hajjej, F., Ayouni, S., Alohali, M.A., Maddeh, M. (2024). Novel framework for autism spectrum disorder identification and tailored education with effective data mining and ensemble learning techniques. IEEE Access, 12: 35448-35461. https://doi.org/10.1109/ACCESS.2024.3349988

[13] Li, P., Wang, H., Tian, G., Fan, Z. (2024). Identification of key biomarkers for early warning of diabetic retinopathy using BP neural network algorithm and hierarchical clustering analysis. Scientific Reports, 14(1): 15108. https://doi.org/10.1038/s41598-024-65694-x

[14] Şenol, A., Talan, T., Aktürk, C. (2024). A new hybrid feature reduction method by using MCMSTClustering algorithm with various feature projection methods: A case study on sleep disorder diagnosis. Signal, Image and Video Processing, 18(5): 4589-4603. https://doi.org/10.1007/s11760-024-03097-1

[15] Martinez-Velasco, A., Martínez-Villaseñor, L., Miralles-Pechuán, L. (2024). Addressing class imbalance in healthcare data: Machine learning solutions for age-related macular degeneration and preeclampsia. IEEE Latin America Transactions, 22(10): 806-820. https://doi.org/10.1109/TLA.2024.10705995

[16] Yu, H., Wang, X., Wang, G., Zeng, X. (2020). An active three-way clustering method via low-rank matrices for multi-view data. Information Sciences, 507: 823-839. https://doi.org/10.1016/j.ins.2018.03.009

[17] Jia, X., Rao, Y., Li, W., Yang, S., Yu, H. (2021). An automatic three-way clustering method based on sample similarity. International Journal of Machine Learning and Cybernetics, 12: 1545-1556. https://doi.org/10.1007/s13042-020-01255-8

[18] Li, W., Li, T., Mojarad, M. (2023). Towards semi-supervised ensemble clustering using a new membership similarity measure. Journal for Control, Measurement, Electronics, Computing and Communications, 64(4): 764-771. https://doi.org/10.1080/00051144.2023.2217601

[19] Vasuki, M., Revathy, S. (2022). Analyzing performance of placement students record using different clustering algorithm. Indian Journal of Computer Science and Engineering (IJCSE), 13(2): 410-419. https://doi.org/10.21817/indjcse/2022/v13i2/221302083

[20] Vasuki, M., Revathy, S. (2020). Efficient handling of incomplete basic partitions by spectral greedy k-means consensus clustering. In 2020 Fourth International Conference on Computing Methodologies and

Communication (ICCMC), Erode, India, pp. 299-305. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00056

[21] Zhu, S., Xu, L., Goodman, E.D. (2022). Hierarchical topology-based cluster representation for scalable evolutionary multiobjective clustering. IEEE Transactions on Cybernetics, 52(9): 9846-9860. https://doi.org/10.1109/TCYB.2021.3081988

[22] Jan, Z., Munos, J.C., Ali, A. (2022). A novel method for creating an optimized ensemble classifier by introducing cluster size reduction and diversity. IEEE Transactions on Knowledge and Data Engineering, 34(7): 3072-3081. https://doi.org/10.1109/TKDE.2020.3025173

[23] Zhang, H., Du, L. (2021). Clustering ensemble via cluster-wise optimization graph learning. In 2021 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE), Shanghai, China, pp. 1-6. https://doi.org/10.1109/RASSE53195.2021.9686881

[24] Liu, Q., Zhao, X., Wang, G. (2023). A clustering ensemble method for cell type detection by multiobjective particle optimization. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20(1): 1-14. https://doi.org/10.1109/TCBB.2021.3132400

[25] Dai, H., Sheng, W. (2019). A multi-objective clustering ensemble algorithm with automatic k-determination. In 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, pp. 333-337. https://doi.org/10.1109/ICCCBDA.2019.8725648

[26] Yang, H., Peng, H., Zhu, J., Nie, F. (2020). Co-clustering ensemble based on bilateral K-means algorithm. IEEE Access, 8: 51285-51294. https://doi.org/10.1109/ACCESS.2020.2979915

[27] Wang, Y., Li, X., Wong, K.C., Chang, Y., Yang, S. (2022). Evolutionary multiobjective clustering algorithms with ensemble for patient stratification. IEEE Transactions on Cybernetics, 52(10): 11027-11040. https://doi.org/10.1109/TCYB.2021.3069434