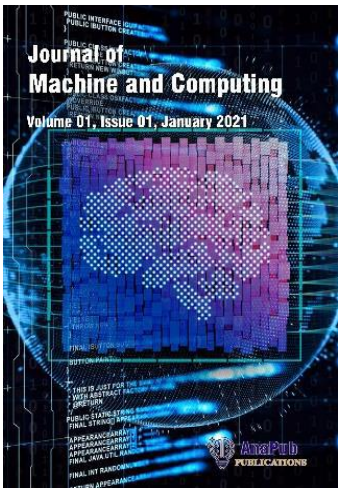
Journal Pre-proof

Optımısıng Clusterıng Accuracy Usıng Mahalanobıs Dıstance-Based Ensemble Methods: A Novel Data Analysıs Paradıgm



Kalimuthu Marimuthu, LNC. Prakash K, Durgadevi P, Prashanthi M and Sunil P

DOI: 10.53759/7669/jmc202505168

Reference: JMC202505168

Journal: Journal of Machine and Computing.

Received 02 March 2025

Revised from 16 May 2025

Accepted 21 July 2025

Please cite this article as: Kalimuthu Marimuthu, LNC.Prakash K, Durgadevi P, Prashanthi M and Sunil P, ‘‘Optımısıng Clusterıng Accuracy Usıng Mahalanobıs Dıstance-Based Ensemble Methods: A Novel Data Analysıs Paradıgm”, Journal of Machine and Computing. (2025). Doi: https:// doi.org/10.53759/7669/jmc202505168.

AnaPub
PUBLICATIONS

OPTIMISING CLUSTERING ACCURACY USING MAHALANOBIS DISTANCE-BASED ENSEMBLE METHODS: A NOVEL DATA ANALYSIS PARADIGM

[1]Dr.Kalimuthu Marimuthu, [2]LNC.Prakash K*, [3]Dr.Durgadevi P, [4]M.Prashanthi, [5]Dr.P.Sunil
[1]Associate Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh 522302.
[*2]Associate Professor, CSE-DS, CVR College of Engineering, Mangalpalle, Hyderabad, Telangana-501510.
[3]Assistant Professor Senior Scale, School of Computer Science and Engineering, Presidency University, ItgalpurRajanakunte, Bengaluru- 560064, Karnataka.
[4]Assistant Professor, Computer Science Engineering, CMR ENGINEERING COLLEGE, Medchal, Kandlakoya, Telangana-501401.
[5]Assistant Professor, Department of CSE, N.B.K.R.I.S.T, Vidyanagar-524413

mkmuthu73@gmail.com, klncprakash.research@gmail.com, durgadeviswamy@gmail.com, prashanthi.m@cmrec.ac.in, . sunilsami541@gmail.com

***Corresponding Author:** LNC.Prakash K

Abstract:

A conspicuous paradigm in Machine Learning involves leveraging the collaborative integration of multiple models through Ensemble Learning to surpass the performance, accuracy, and generalization capabilities of individual models. This study aims to investigate the deployment of Ensemble Learning as an approach to refine the precision of Cluster Analysis. Conventional clustering algorithms normally tackle with the complexity of datasets portrayed by multifaceted attributes, many of which demonstrate obscure or non-linear interdependencies. By employing advanced Ensemble Learning approaches, this research hopes to elaborate clustering efficacy and perceive latent patterns that remain imperceptible through traditional approaches. The proposed investigation underscores Ensemble Learning as a transformative approach in the domain of cluster analysis within data science. Its ability to augment the accurateness of clustering techniques and extract hidden structures from data not promptly apparent through conventional algorithms accentuates its indispensability. Accordingly, this research endeavours to revolutionize the precision and applicability of clustering methodologies in solving real-world data analysis challenges. To substantiate the efficacy of the proposed framework, a comprehensive comparative evaluation is performed, benchmarking its outcomes against those obtained from varied datasets and established clustering algorithms.

Keywords: Machine Learning, Ensemble Techniques, Clustering Accuracy, Hidden Patterns, ,Mahalanobis distance, Optimization.

1. Introduction:

Clustering is an unsupervised machine learning approach that divides a set of data objects into clusters. The idea is to group comparable data points in the same cluster while retaining dissimilar data points in separate clusters. Clustering seeks to find underlying patterns and structures in data based on shared or distinct characteristics between data points. An exploratory data analysis technique called clustering can shed light on the underlying patterns

or natural groups in the data. It is frequently used for a variety of activities across several domains, including data compression, anomaly detection, customer segmentation, and data summarization. Generally, clustering algorithms fall into one of many categories: partitional(K-Means, for example), [1], hierarchical [2,3], density-based (DBSCAN, for example) [4.5], or model-based (Gaussian Mixture Models, for example) [6]. The features of the data, the intended clustering attributes, and the available processing power all influence the method selection. In data mining, machine learning, and exploratory data analysis, clustering is a crucial approach that helps reveal latent structures and patterns in datasets without the need for labelled data, [7].

Even though there are many different clustering techniques, clustering analysis is particularly difficult when there is no previous knowledge. It is commonly known that no one clustering technique can accurately and consistently capture structural information. Furthermore, different starting settings might cause even the identical clustering technique to fail in producing clusters that are suitable and specific clustering techniques may be affected by the dataset's properties, beginning circumstances, noise level, and outliers. Ensemble approaches [8, 9], seek to address these problems by combining several clustering and utilising the advantages of various techniques. To provide a more reliable and accurate clustering result, ensemble clustering is a technique that combines many clustering solutions. When utilising ensemble clustering techniques, as opposed to solo clustering approaches, more stable, accurate, and resilient results are produced. Many novel ensemble clustering approaches have been developed recently due of the increased interest in ensemble clustering [10-12]. The previously described research has shown promising results when tackling an ensemble problem in clustering. However, most existing ensemble clustering results use hard clustering, in which each item is assigned to one or more groups, with clear borders between each cluster. When sample information is scarce, hard clustering techniques often result in increased judgement uncertainty. To clarify the ambiguity in the data, the idea of a three-way choice [13-16], presented as a solution to this problem. Given its capacity to improve the accuracy, stability, scaling adaptability, variability, and interpretability of clustering analysis, ensemble clustering is a crucial tool in machine learning and data mining. Ensemble approaches, which reduce individual algorithm biases and mistakes and produce more dependable clustering results, are especially useful for managing large, heterogeneous, or high-dimensional datasets. They do this by merging numerous techniques for clustering or runs of just a single algorithm. Ensemble clustering is also very useful in bioinformatics, image processing, text mining, and social network analysis. It does this by utilising several viewpoints on data structures and interactions, which improves stability, scalability, and comprehension. Modern data analytics paradigms value it even more because of its ability to incorporate new algorithms and adjust to changing requirements for data analysis.

The three main steps of ensemble clustering are usually creating multiple base clusterings with various parameter settings or algorithms, merging or balancing the base clusterings with voting schemes or consensus functions, and creating an ultimate consensus clustering that symbolises the ensemble solution. There have been several approaches to ensemble clustering developed, including evidence accumulation clustering (EAC), cluster-based similarity partitioning

algorithm (CSPA), and hyper-cube-based ensemble clustering. These techniques help lessen the curse of dimensionality in high-dimensional data settings, manage complicated data distributions, provide better cluster quality, and capture different points of view. More reliable, accurate, and stable clustering solutions may be obtained with ensemble clustering by utilising the advantages of many clustering techniques while minimising the drawbacks of each one separately.

The improved two-way clustering technique presented in this article uses an ensemble approach to solve problems that result from judgements being made incorrectly because of incomplete or faulty data. Our suggested method carefully samples feature subsets and uses a standard clustering algorithm to create different foundational clustering outcomes, in contrast to traditional clustering ensemble strategies that use several clustering algorithms to obtain foundational clustering outcomes. We develop a two-way clustering strategy with a voting mechanism by using these fundamental clustering results. Our suggested method's main process is divided into two stages. First, we employ base clustering techniques to provide baseline clustering results. Then, we use the common tuple approach to identify two stage clustering and label alignment to arrange all clustering results in a predetermined order.

This paper follows the following structure. We mainly explain the literature and concepts of grouping and ensemble clustering in Section 2. The approach of the proposed algorithm is described in Section 3. Using a variety of datasets, Section 4 shows how effective the suggested ensemble clustering approach is. Conclusions and future research directions are outlined in Section 5.

2. Literature review:

In this section, we discuss certain concepts and related works of and ensemble clustering. Each clustering technique takes a different approach to finding patterns in the data. diverse clustering approaches might provide diverse results even when they are used to the same data. There isn't a single clustering technique that works for every data structure. Due to the lack of available previous class knowledge, selecting a particular clustering technique is difficult. Consequently, ensemble clustering the process of merging several clustering results into a single, cohesive resultbecomes the focus of research. When it comes to outcomes, ensemble clustering is more resilient, stable, and high-quality than individual clustering techniques.

Strehl and Ghosh [17], first proposed the concept of ensemble clustering by merging cluster labels without directly accessing the original attributes. An ensemble clustering method that considers factors such as cluster magnitude, sample size, and density was developed by Wang et al. [18]. Punera and Ghosh [19] built on the more inflexible clustering procedures and proposed several consensuses approaches appropriate for soft clustering. An ensemble clustering technique based on sample stability was created by Li et al. [20]. components of a cluster. Initial cluster generation and cluster merging are the two basic steps of ensemble clustering. The first stage is the initial cluster creation, during which new clusters are produced. These clusters can be created in a number of ways, either by varying the parameters of one algorithm or by using distinct algorithms. We focus on the cluster merging procedure and the conversion of different set of clusters to real clustering in our work. Figure 1 shows the technique of ensemble clustering.

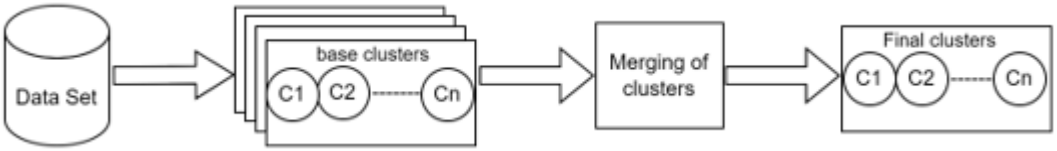


Figure 1: general approach of ensemble clustering

An ensemble clustering method should combine several clustering results into a single partition that is like the original clustering without using the original dataset. Some researchers use the original characteristics and the different clustering results as inputs to further improve the clustering accuracy [21, 22].

Although several approaches for ensemble clustering have been developed recently, they usually suffer from two fundamental issues. Firstly, their handling of ambiguous links is poor, which might lead to inaccurate conclusions. Secondly, they are unable to enhance local relationships using general information. The study [23], provide a novel approach to clustering by employing probability analysis and sparse graphs. To identify ambiguous connections, researchers employ a unique technique that yields a network that contains only the most trustworthy relationships. Using this method yields better results than using every link. They also apply a random walk method to have a better understanding of the entire graph. This study [24] explores a novel method of ensemble clustering for time series analysis in finance. Improving the resilience, accuracy, and consistency of financial related data clustering is the goal. The underlying complexity and volatility of financial time series are addressed by ensemble clustering, which combines several clustering results into a single, cohesive conclusion. authors addressed over the main obstacles and limitations of the existing approaches, as well as the theoretical foundations, algorithmic strategies, computational features, and real-world applications of ensemble clustering.Wireless sensor networks (WSNs) are collections of several tiny sensors that collect environmental data. The energy used by these gadgets is finite and not rechargeable. Thus, energy conservation is critical for WSNs. Using clusters and data routing via designated cluster leaders, or Cluster Heads (CH), is one method for doing this. This technique increases the network's longevity and helps distribute energy more evenly. [25], provides a novel energy-efficient routing technique in this work that combines ODMA, Genetic Algorithm, and K-means.

A cluster ensemble is created by combining various data grouping techniques to provide superior results. It employs many methods rather than just one, combining the outcomes. This contributes to the accuracy of the groups. Given the success of recently developed strategies, it is important to examine and comprehend them. To assist readers in selecting the cluster ensemble approach that best suits their objectives, [26], covers an overview of different techniques. It discusses their types, properties, and practical applications.The focus of recent research has mostly been on four areas. First, the process of selecting ensemble members [27-29]. Secondly, before to joining them, choose the finest members of the ensemble [30, 31]. Third, how to combine these individuals [32, 33]. Fourth, applications of clustering ensemble in practice [34–36].

Using a variety of feature groups while creating components is useful for data sets with numerous dimensions. The authors of [29] propose a novel clustering ensemble approach that combines random projection with fuzzy c-means clustering. The authors of [37] compare three methods for lowering dimensions: random sampling, principal component analysis, and

random projection. Anomalies and noise might affect the clustering ensemble's results. Many ensemble components are often produced via a clustering ensemble approach. However, it is not a good idea to combine every single component that is accessible. As such, it is imperative to carefully choose suitable ensemble members. A strategy called selective clustering ensemble combines only some of the ensemble's elements instead of combining them altogether. The researchers examine the variation among ensemble components in reference [38]. They concluded that, even when the latter includes more precise components, combining components with large diversity is better than combining those with little variability.The authors cited in [39] investigate several methods for using relative clustering validity measures to evaluate and choose ensemble members. These measures identify high-quality ensemble members appropriate for clustering ensembles by measuring the correlation between clusters and partitions. Through the combination of these relative criteria, they create an assessment criterion that is decisive enough to choose only the best individuals for participation, as opposed to the entire ensemble.

The authors of citation [31], present a progressive semi-supervised clustering ensemble technique that, using two different cost criteria, removes unneeded ensemble components. The first cost measure assesses the similarity between two subspaces as well as the cost of ensemble elements. In the meantime, the second cost metre calculates the total cost incurred during the process of incorporating the selected members into the final partition. The graph technique also used as consensus in this research, authors use the normalised cut method to address the problem of combining members by viewing it as a graph partitioning problem. The authors of reference [40], present a clustering ensemble approach based on the evidence theory of Dempster-Shafer. This approach considers the contextual knowledge of the data's cluster structure and uses neighbouring data to characterise it. For each piece of data, they first determine its neighbours and calculate the label probability among all members of the ensemble. The result is then obtained by combining these label probabilities using Dempster-Shafer theory.

3. Proposed two-step ensemble methodology:

This section introduces our suggested ensemble clustering technique, which uses several clustering partitions to improve the accuracy and resilience of clustering algorithms. Motivated by the effectiveness of ensemble learning in classification problems, our method seeks to maximize clustering performance by combining the advantages of several clustering techniques. Using the concept of an outer border region, a two-stage clustering method is presented to illustrate the uncertainty details in the dataset. Even while several two-way ensemble clustering approaches have shown encouraging results, there is still a great deal of room for improvement. This section presents an improved two stage clustering technique in which base cluster formation and the proposed ensemble are the first and second stages respectively.

3.1. Base clusters formation:

Unlike existing methods, our suggested method uses traditional clustering techniques to provide a variety of basic clustering results after randomly selecting a subset of characteristics from the data. Algorithm 1 represents the procedure of procuring base clustering results. Let $D = \{X_1, X_2, \ldots X_n\}$, be the dataset with n number of tuples and k be the number of class labels.

Consider $M = \{M_i / M_i \; is \; a \; Clustering \; method \; and \; i = 1 \; to \; m\}$, be the set of clustering methodologies and $C^i = \{C_1^i, C_2^i, \ldots \ldots \; C_k^i\}$, be the set of clusters formed by the clustering method M_i.

ALGORITHM 1: PROCURING BASE CLUSTERS.

Input: Dataset D, Number of Clusters k.
Output: Base clusters.
1. *For (i =1 to m) do*
2. *Find the clusters using the clustering method* m_i.
3. *Return the clusters* $C_1^i, C_2^i, \ldots \ldots \; C_k^i$
4. *End*

In this explorationK-Means Clustering and Mean Shift Clustering are used as base clustering methodologies. Initially the primitive task is to determine the standard number of clustersrequired to separate the dataset into significant groups before utilizing different clustering procedures to train the dataset [30]. The dataset is exposed to widespread clustering methods, such as BIRCH Clustering, Mean Shift, Affinity Propagation, and k-means Clustering. Main clustering findings are supplied using these methods and competed to the suggested method. As stated, two distinct clustering techniques K-Means Clustering and Mean Shift Clusteringare combined to create a collective model for this study's embedded model. This groupingapproach advances the whole strength and accurateness of clustering by utilising the gains of every grouping procedure.

3.2.Proposed ensemble approach:

In this approach, true combinations are determined through the amalgamation of the K-Means and mean shift techniques. In the beginning, the correlation between every pair of objects within the dataset is scrutinized to ascertain if they pertain to the same cluster through all engagedapproaches. They should constantly align with the identical cluster across each method, they are designated to that specific cluster; conversely, an evaluation for an alternative cluster is subsequently undertaken. Upon the distribution of data records to a predetermined number of groups, a thorough inspection is conducted for any residual entities within the dataset. For the objects that not yet allocated to any of existing cluster, they are assigned to one of the existent clusters grounded on the similarity of their objects to the partial ensemble clusters. The delineated approach is expounded in Algorithm 1. The likely ness of the tuples that are not part of any partial clusters is determined by mahalanobis distance.

3.2.1. Mahalanobis distance:

The distance between a point and a distribution in an N-dimensional space is computed using the Mahalanobis distance. It is a useful tool for finding anomalies, but it may also be used for point classification when there is a lack of available data. If it needed to compare just two points, p and q, in a space of N dimensions, it becomes necessary to take into consideration the variation of these locations along each axis to calculate the overall distance between them. Thus, the N-dimensional distance equation, also known as the Euclidean distance or any other distance measure is used. When it comes to statistical and machine learning techniques, the Euclidean distance is highly valued and often employed. However, its use is restricted to point-by-point comparisons. There are a few things to keep in mind while trying to compare a point

to a group of points. Usually, we start by using mean computation to compress the spread into a single point to evaluate the span from a spread to a certain point, [41]. The span to the point can then be evaluated in respect to departures from this mean. This method works well in one-dimensional situations, but it cannot perform well in multidimensional settings with a group of points.To address this issue mahalanobis is instituted to find the similarity between each leftover object and initial groups formed, equation 1 is used to compute the mahalanobis distance.

$Dist = \sqrt{(x-m)^T * Cov^{-1} * (x-m)}$ ..... (1)

Where,

$Dist$, is the Mahalanobis distance,

x, is the vector of an object which needs to be allocated to the existing groups.

m, Is the mean value of the objects that belongs to an existing cluster.

Cov^{-1}, is the inverse covariance matrix of the objects that belongs to an existing cluster.

3.2.2. Mean Shift Clustering:

Mean shift grouping is a non-parametric assemblage approach; Mean Shift doesn't need a perception on the number of clusters beforehand. It initiates by repeatedly switching data points in the recipient of the confirming probability distribution type. The next is anillustratedtechnique for Mean Shift grouping:

1. Input: Let the data objects $Y = \{y_1, y_2 \ldots . y_n\}$, Kernel function K with bandwidth h and Convergence threshold δ.
2. Initialize group midpoints $C = Y$ and shift vector $Shift = Zeros.$
3. Replicate till convergence:
   $for\ (every\ y_i \in Y)$**:**

   Compute the mean shift using $Shift(y_i) = \frac{\sum_{j=1}^{n} K\left(\frac{y_i - y_j}{h}\right) * y_j}{\sum_{j=1}^{n} K\left(\frac{y_i - y_j}{h}\right)}$. ..... (2)

   Update cluster centres $C = C + Shift$
4. $If(\|Shift\| < \delta)$, end the for loop.
5. Assign each data point y_i to the group whose centre it converged.
6. Return the clusters C

The kernel function K is commonly a Gaussian kernel, though several options may be exploredconstructed on the data's possessions. The bandwidth h directs the extent of the neighborhood employed for estimating local density, concerning cluster shape and count. The convergence threshold ϵ indicates when mean shift replications should terminate. The shift vector indicates both the trend and scale of mean shift for every data object. The procedure continually changes data points toward the mode of the fundamental density until convergence is undertaken.

3.2.3.K-Means Clustering:

A well-liked unsupervised machine learning method for classifying observations into k groups is k-means segmentation. Data points are continuallyallocated to the nearest centroid, and centroids are reorganizedmatching to the common of the tuples within every cluster. The within-cluster variance is the target of the method. Because of its effectiveness and straightforwardness, it is frequently used for tasks like picture reduction and consumer segmentation. Comprehending its procedures is essential for efficiently dividing datasets and obtaining significant insights. The following is the procedure for K-means clustering.

- Select k randomly chosen centroids of the initial groups from the data points. The clusters' primary centres will be these the centroids.
- Place each data point in relative to the nearest centroid. To assign the data point to the cluster whose centroid is nearest, this action involves calculating the distance involving each data point and each centroid.
- Recalculate the group centroids applying the average of all the data points distributed to every cluster. To do this, the centroid must be keep informed with the mean location of each cluster's data points.
- Repeat above two steps until the convergence conditions are convinced, then replicate allocation and updating. Convergence is usually obtained when the centroids do not vary considerably between repetitions.
- The final cluster positions are finalized at this point in the procedure, and the centroids are the cluster centres.

3.4.**Embedded methodology:**

This method combines the Mean Shift and K-Means algorithms to generate coherent clusters. The process begins by comparing every pair of items in the dataset to assess whether they belong to the same cluster based on all the applied techniques. If they consistently appear in the same cluster across these methods, they are grouped together. Otherwise, they are assigned to separate clusters in subsequent steps. Any unclassified items left after organizing the data into the predefined number of clusters are further examined to determine their appropriate grouping to one of the clusters that already exist based on the degree of similarity to the existing

cluster, the degree of similarity is determined by the mahalanobis distance.The defined methodology is illustrated in Algorithm 2. Corresponding to the procedure Let D be the database, $P = \{P_i \backslash P_i \; is \; a \; clustering \; procedure\}$, which is the set of clustering procedures, the set of cluster groups is denoted by G and is defined as $G = \{G_{ij} \backslash G_{ij} \; is \; j^{th} \; Group \; in \; i^{th} \; clustering \; procedure\}$,and $N = \{N_k \; \backslash k = 1,2, \ldots number \; of \; cluster \; groups\}$, be the set of resultant cluster groupings. In this exploration K-Means Clustering, Mean Shift Clustering, Agglomerative Clustering, BIRCH Clustering are exercised and attainedsolutionsseparately on the defined databases. For the ensemble clustering Mean Shift Clustering and K-Means Clustering procedures are

Algorithm 2:
Input: Data base D, base clusters of all clustering procedures.
Output: optimal clusters resulted from proposedembedded method.
Consider,

$$P = \{P_i \backslash P_i is a clustering method\}$$

$G = \{G_{ij} \backslash G_{ij} is j^{th} th \; cluster group in i^{th} method\}$,
$N = \{N_k \; \backslash k = 1,2, \ldots number of cluster groupings\}$,

1. Consider$k = 1, step = 0$.//Initialization of variable
2. $while(k \neq |N|)$
3. $N_k = \emptyset$.
4. $for(every \; object_x \epsilon \; D)$
5. $for(every \; object_y \epsilon \; D, x \neq y)$ // Consider every couple of tuples
6. $for(i = 1 \; to \; |P|)$
7. $for(j = 1 \; to \; |G|)$ // For each cluster from very technique
8. $if(\{object_x, object_y\} \epsilon \; C_{ij})$ //for each pair of the objects in same group
9. $Count += 1$
10. $Sep \; 8 \; terminated.$
11. $Sep \; 7 \; terminated.$
12. $Sep \; 6 \; terminated.$
13. $if(step = |P|)$
14. $N_k = N_k \cup \{object_x, object_y\}$. // Assigning couple of tuples to same group
15. $Sep \; 13 \; terminated.$
16. $Sep \; 5 \; terminated.$
17. $Sep \; 4 \; terminated.$
18. $step = 0.$
19. $k = k + 1.$
20. $D = D - N_k.$
21. $Sep \; 2 \; terminated.$
22. $if(D \neq \emptyset)$
23. Find the mahalanobis distance from each object of D to each cluster group N_k using equation 1.
24. $for(every \; object_x \epsilon \; D \; and \; every \; group \; N_k)$
25. $if(Mahalanobis \; distance \; from \; object_x \; to \; N_k \; is \; minimum)$
26. $N_k = \{N_k\} \cup \{object_x\}$.
27. $Sep \; 24 \; terminated.$
28. $Sep \; 22 \; terminated.$
29. Return the optimal set of ensemble clusters, $N = \{N_k \backslash l = 1,2, \ldots |N|)$.

employedto improve the clustering truthfulness. The architecture of the proposed methodology is represented in the figure 1.

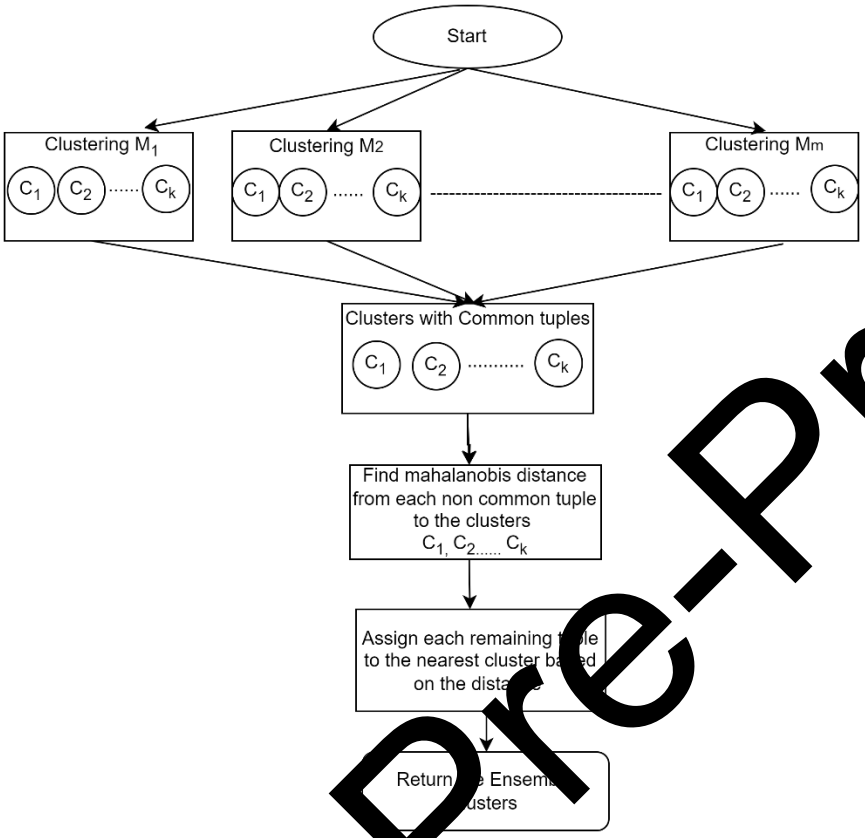


Figure 1: architecture of the proposed t step ensembled clustering methodology

4. **Experimental analysis and performance evaluation:**

This section delves into the experimental investigations, the datasets utilized, and the insights derived from their evaluation. Comprehensive experiments were carried out using the Weather History dataset and the Weather Prediction dataset.

4.1 Database Description

The Historical Weather Archive provides past weather data for various locations, containing general information about weather conditions documented over specific intervals. The dataset includes a total of 96,453 entries, each demonstrating a unique timestamp matching with associated weather parameters. After thorough analysis using visualizations and perceptions, the most suitable number of clusters for this dataset was recognized as four. This reveals that partitioning the dataset into four clusters supplies the most appropriate and significant grouping of data points based on the defined conditions and the problem circumstances.The Weather Prediction dataset involves of meteorological data collected from 18 different European locations between the years 2000 and 2010. The dataset comprises 3,654 daily records and includes variables such as average temperature, maximum temperature, minimum temperature, and more. The optimal number of clusters for this dataset was established to be two, meaning that dividing the data into two clusters delivers the most meaningful and descriptive classification of data points for the given dataset and problem obligations.In dry bean dataset

seven different types of dried beans were used in this study, which considered variables including appearance, morphology, category, and content depending on the state of the market. To achieve consistent seed classification, a sophisticated computer vision system was developed to distinguish between these seven recorded varieties of dry beans that have comparable attributes. Using a high-end camera, the system took pictures of 13,611 individual beans from these seven recognised kinds. After segmenting and extracting features from the pictures acquired by the computer vision system, a total of 16 characteristics 12 dimension and 4 form categories were identified from the beans.

4.2. Analysis of Performance:

This section presents a complete exploration of the investigational results and calculates the performance of the anticipated model in comparison to sophisticated clustering methods. The study was directed in two key components: measuring the clustering accurateness of determined methods and showcasing the returns of the anticipated ensemble methodology. Traditional methods such as K-Means Clustering, Affinity Propagation, Mean Shift Clustering, and BIRCH Clustering were evaluated using applicableproximity measures to extract significantperceptions from the datasets. Each technique showed unique strengths, with K-Means excelling in simplicity, Affinity Propagation identifying exemplars, Mean Shift adapting to non-linear clusters, and BIRCH efficiently handling large datasets. The proposed ensemble method combined the adaptability of Mean Shift with the excelling in simplicity of K-Means, advancing a strong and computationally effectiveway out. Execution was measured using the Davies-Bouldin and Silhouette scores, which highlighted the ensemble method's superior ability to form well-defined clusters. For the Weather History dataset, the Elbow Method identified four optimal clusters, effectively capturing the dataset's structure. Similarly, for the Weather Prediction dataset, the Elbow Method and Silhouette Score determined two clusters as the most significant configuration. These solutionshighlight the consequence of choosing clustering methods that associate with dataset features, with the proposed model determining its potential to produce discerning and demonstrativecategories for both datasets. Based on the findings presented in Table 1, it is evident that the ensemble clustering method achieved a lower Davies-Bouldin score [31] and a higher Silhouette score [32] compared to all other traditional clustering techniques applied to the Weather History dataset. These results highlight that the ensemble model provides better clustering performance, excelling in both the separation and compactness of clusters as evaluated by these metrics.

Algorithm	Number of clusters	Davis Bouldin Score	Silhoutte Score
K-Means Clustering	4	0.401	0.608
Mean Shift Clustering	4	0.435	0.867
Agglomerative Clustering	4	0.405	0.588
BIRCH Clustering	4	0.405	0.628
Ensembled Clustering	**4**	**0.124**	**0.896**

Table 1: Comparison of Ensemble and Traditional Models on Weather Data

According to the data presented in Table 2, the ensemble clustering methodology attained a substantially lower Davies-Bouldin score when compared to a range of conventional clustering

algorithms applied to the Weather Prediction dataset. These conclusions indicate that the ensemble model outperforms traditional approaches, representative superior clustering excellence by succeeding better partition between clusters and impressive cohesion within clusters, as assessed by these metrics.

Clustering algorithm	**Number of Clusters**	**Davis Bouldin Score**	**Silhoutte Score**
K-Means	2	0.937	0.414
Mean Shift	2	0.955	-0.002
Agglomerative	2	1.021	0.354
BIRCH	2	0.97	0.378
Ensemble	**2**	**0.562**	**0.452**

Table 2: Comparison of Ensemble and Traditional Models on Weather prediction dataset.

From the table 3, on the Dry Bean Dataset, the ensemble clustering method bests other algorithms, attaining the lowest Davies-Bouldin Score (0.517) and the highest Silhouette Score (0.497), demonstrating advanced cluster partition and consistency. K-Means operates reasonably well with a Davies-Bouldin Score of 0.746 and a Silhouette Score of 0.429, suggesting decent cluster efficiency. Mean Shift struggles with a higher Davies-Bouldin Score (0.831) and a negative Silhouette Score (-0.003), highlighting poor clustering. Agglomerative Clustering shows the weakest separation with the highest Davies-Bouldin Score (0.901) but upholds some efficiency (Silhouette Score 0.419). Overall, the ensemble method confirms most efficient for this dataset.

Clustering algorithm	**Number of Clusters**	**Davis Bouldin Score**	**Silhoutte Score**
K-Means	7	0.746	0.429
Mean Shift	7	0.831	-0.003
Agglomerative	7	0.901	0.419
BIRCH	2	0.653	0.431
Ensemble	**7**	**0.517**	**0.497**

Table 3: Comparison of Ensemble and Traditional Models on Dry bean data.

The analysis reveals that the ensemble model integrates the outcomes of two obvious clustering procedures, Mean Shift and k-Means. The mahalanobis distance measure places a vital role in allocating the left-over objects from the database which are not allocated in any of the base clusterings through a voting mechanism to derive the optimal ensemble clustering results. These models are selected due to their capability to recognize dense regions in the data successfully. Mean Shift excels in finding clusters of varying shapes and sizes, making it ideal for capturing intricate patterns in the dataset. Meanwhile, The K-Means algorithm is efficient, scalable, and easy to implement, making it ideal for large datasets. It performs well with distinct, spherical clusters and provides clear centroids for easy assignment. Requiring minimal tuning. Despite limitations with complex cluster shapes, its simplicity and speed make it highly popular. For reliableinsight of comparison of ensemble and traditional models on Weather Data it is depicted in figure 2.

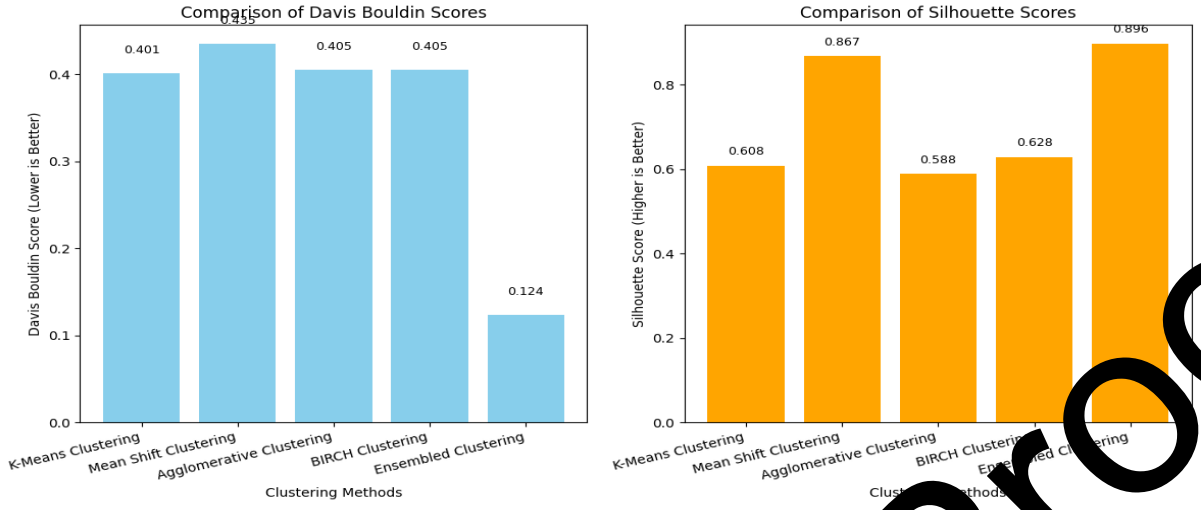


Figure 2: Comparison of Ensemble and Traditional Models on Weather Data

For better identification of comparison of ensemble and traditional models on Weather Data it is depicted in figure 3.

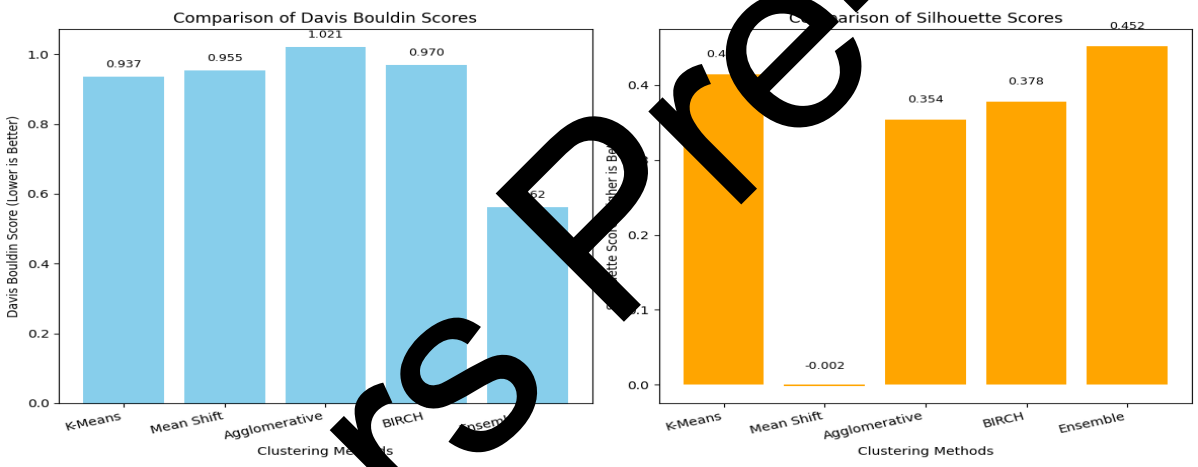


Table 2: Comparison of Ensemble and Traditional Models on Weather prediction dataset.

For capable understanding of comparison of ensemble and traditional models on Dry bean data it is depicted in figure 4. Based on the analysis of the three datasets with different cluster counts (2, 4, and 7), Ensemble Clustering consistently outperforms other methods, achieving the lowest Davis Bouldin Scores and the highest Silhouette Scores across all scenarios. This indicates that Ensemble Clustering produces compact, well-separated, and high-quality clusters regardless of the number of clusters. In contrast, Mean Shift Clustering generally performs poorly, particularly with negative Silhouette Scores for datasets with 2 and 7 clusters, suggesting its inability to handle compact and well-defined clusters effectively. The results in table 4 exhibit the finer performance of the recommended method over the reference [41] in both Davis Bouldin (DB) and Silhouette Scores throughout Weather History and Prediction datasets. A substantial reduction in DB Scores focuses improved cluster trimness, while marginally higher Silhouette Scores imply better-defined clusters.

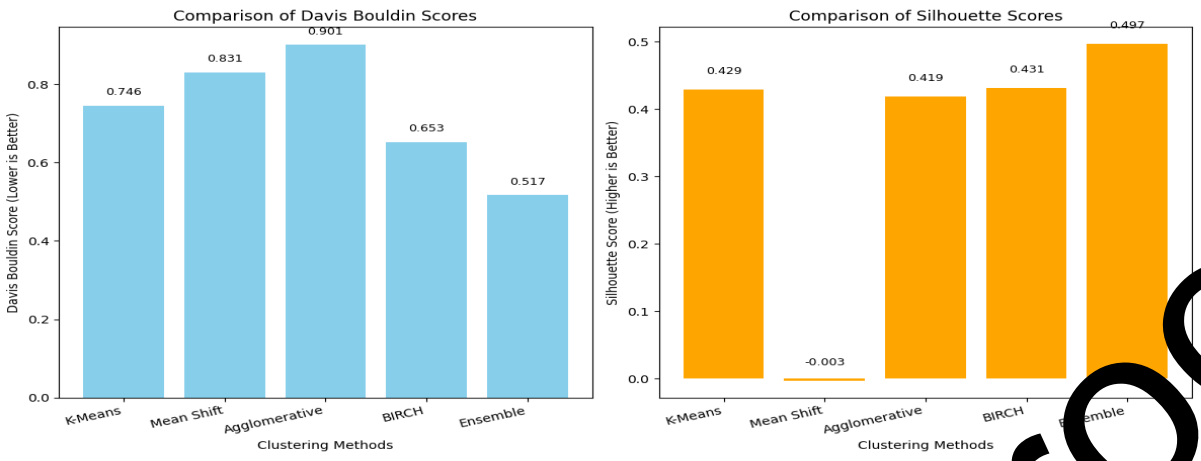


Figure 3: Comparison of Ensemble and Traditional Models on Synthetic data.

K-Means Clustering and BIRCH Clustering show moderate performance, with decent clustering quality in most cases, but they are outperformed by Ensemble Clustering. Agglomerative Clustering, however, often exhibits the worst performance with the highest Davis Bouldin Scores, reflecting poor separation of clusters. Overall, Ensemble Clustering is the most robust and reliable algorithm across different datasets, while Mean Shift and Agglomerative Clustering struggle to deliver consistent results.

	Davis Bouldin Score		Silhoutte Score	
	Reference [41]	Proposed method	Reference [41]	Proposed method
Weather History Data	**0.184**	**0.124**	**0.873**	**0.896**
Weather Prediction data	**0.683**	**0.562**	**0.427**	**0.452**

Table 4: Comparison of the proposed ensemble with the ensemble presented in reference [41].

The proposed method shows greater advancement on the Weather History Data, suggesting its strength lies in handling structured datasets with clear borders. For noisier datasets like Weather Prediction Data, its efficiency is moderate, leaving scope for further enhancement.

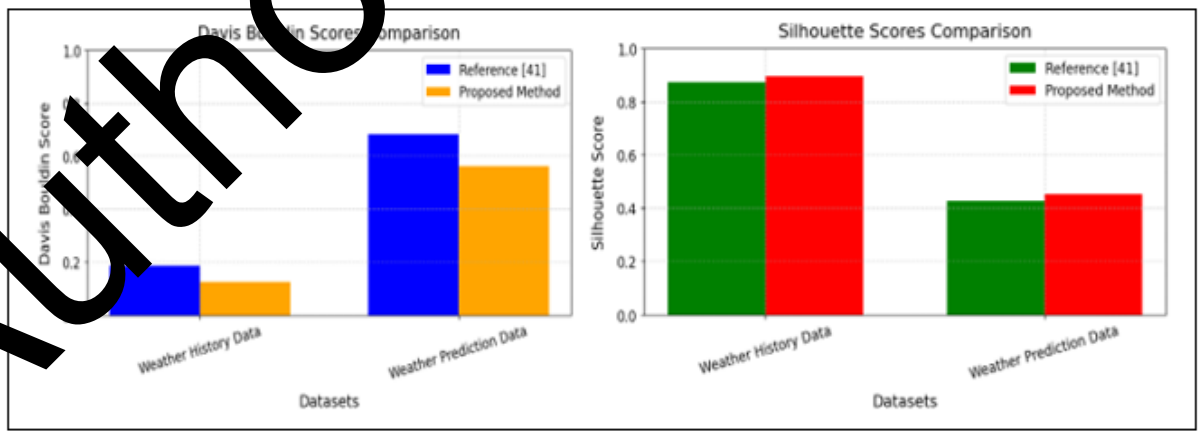


Figure 4: Comparison of the proposed ensemble with the ensemble presented in reference [41].

Figure 4 illustrates the comparison between the proposed ensemble method and the one introduced in reference [41]. It highlights the performance differences, showcasing the

improvements in accuracy, efficiency, or robustness achieved by the proposed approach over the existing ensemble technique.To further enhance the accuracy of these clusters, optimization techniques, as proposed in references [42, 43], are suggested for future application.

5. Conclusion:

This research has addressed the enduring limitations of traditional clustering algorithms by mounting a novel ensemble model that embodies a harmonious blend of precision, scalability, and adaptability. Through the strategic combination of the Mean Shift and K-Means algorithms via a robust voting mechanism, this research introduced a method capable of delivering superior performance compared to conventional clustering approaches. Extensive comparative evaluations accentuated the efficacy of this ensemble model, highlighting its remarkable ability to discern elaborate patterns and reveal the latent structures within complex datasets. The consequence of this study extends beyond its empirical findings. By highlighting adaptability and flexibility, the proposed model begins as a transformative tool for data analysis, capable of seamlessly transitioning across diverse datasets and domains. It serves not only as a mechanism for uncovering significant insights but also as a catalyst for enhancing decision-making processes by offering a more nuanced and reliable understanding of data-driven phenomena. Looking to the future, the potential of this ensemble model is vast and promising. Advancing this work will involve extending its applicability to accommodate a broader spectrum of data types, scaling its capabilities to manage increasingly large datasets, and purifying its accessibility through the integration of advanced automation and user-centric features. By addressing these avenues, this research sets the foundation for an advanced clustering framework that is as versatile as it is powerful. This contribution not only elevates the field of clustering methodologies but also concretes the way for innovative solutions that can meet the evolving demands of data science and analytics with elegance and efficacy.

References:

[1] MacQueen, J. Some methods for classification and analysis of multivariate observations. *Berkeley Symp. Math. Stat. Probab.* 1967, *5*, 281–297.

[2] Gurrutxaga, I.; Albisua, I.; Arbelaitz, O.; Martín, J.I.; Muguerza, J.; Pérez, J.M.; Perona, I. An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index. *Pattern Recognit.* 2010, *43*, 3364–3373.

[3] Azzag, H.; Lebbah, M. *A New Way for Hierarchical and Topological Clustering*; Guillet, F., Pinaud, B., Venturini, G., Zighed, D., Eds.; Advances in Knowledge Discovery and Management; Springer: Berlin/Heidelberg, Germany, 2013; pp. 85–97.

[4] Lotfi, A.; Moradi, P.; Beigy, H. Density peaks clustering based on density backbone and fuzzy neighborhood. *Pattern Recognit.* 2020, *107*, 107449.

[5] Birant, D.; Kut, A. ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data Knowl. Eng.* 2007, *60*, 208–221.

[6] Govaert, G.; Nadif, M. An EM algorithm for the block mixture model. *IEEE Trans. Pattern Anal. Mach. Intell.* 2005, *27*, 643–647.

[7] Suhaas, K. P., Deepa, B. G., Shashank, D., & Narender, M. (2024). Millets Industry Dynamics: Leveraging Sales Projection and Customer Segmentation. *SN Computer Science*, *5*(8), 1063.

[8] Strehl, A.; Ghosh, J. Cluster ensembles-a knowledge reuse framework for combing multiple partitions. *J. Mach. Learn. Res.* 2003, *3,* 583–617.

[9] Fred, A.L.N.; Jain, A.K. Combining multiple clusterings using evidence accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.* 2005, *27*, 835–850.
[10] Huang, D.; Wang, C.D.; Lai, J.H. Locally weighted ensemble clustering. *IEEE Trans. Cybern.* 2018, *48*, 1460–1473.
[11] Xu, L.; Ding, S.F. A novel clustering ensemble model based on granular computing. *Appl. Intell.* 2021, *51*, 5474–5488.
[12] Zhou, P.; Wang, X.; Du, L.; Li, X.J. Clustering ensemble via structured hypergraph learning. *Inf. Fusion* 2022, *78*, 171–178.
[13] Afridi, M.K.; Azam, N.; Yao, J.T. A three-way clustering approach for handling missing data using gtrs. *Int. J. Approx. Reason.* 2018, *98*, 11–24.
[14] Wang, P.X.; Yao, Y.Y. CE3: A three-way clustering method based on mathematical morphology. *Knowl.-Based Syst.* 2018, *155*, 54–65.
[15] Wang, P.X.; Yang, X.B. Three-way clustering method based on stability theory. *IEEE Access* 2021, *9*, 33944–33953.
[16] Lavanya, K., Reddy, Y.S., Varsha, D.C., Sai, N.V., Meghana, K.L. (2024). IDS-PSO-BAE: The Ensemble Method for Intrusion Detection System Using Bagging–Autoencoder and PSO. In: Hassanien, A.E., Castillo, O., Anand, S., Jaiswal, A. (eds) International Conference on Innovative Computing and Communications. ICICC 2023. Lecture Notes in Networks and Systems, vol 731. Springer, Singapore. https://doi.org/10.1007/978-981-99-4071-4_61.
[17] Strehl, A.; Ghosh, J. Cluster ensembles-a knowledge reuse framework for combing multiple partitions. *J. Mach. Learn. Res.* 2003, *3*, 583–617.
[18] Wang, X.; Yang, C.Y.; Zhou, J. Clustering aggregation by probability accumulation. *Pattern Recognit.* 2009, *42*, 668–675.
[19] Punera, K.; Ghosh, J. Consensus-based ensembles of Soft clusterings. *Appl. ArtificalIntell.* 2008, *22*, 780–810.
[20] Li, F.J.; Qian, Y.H.; Wang, J.T.; Dang, C.Y.; Li, L.P. Clustering ensemble based on sample's stability. *Artif. Intell.* 2019, *273*, 37–55.
[21] S. Vegapons, J. Correa-Morris, J. Ruizshulcloper Weighted partition consensus via kernels Pattern Recognit., 43 (8) (2010), pp. 2712-2724
[22] Z. Yu, H. Wong, J. You, G. Yu, G. Han Hybrid cluster ensemble framework based on the random combination of data transformation operators Pattern Recognit., 45 (5) (2012), pp. 1826-1837.
[23] Huang D, Lai JH, Wang CD. Robust ensemble clustering using probability trajectories. *IEEE Trans Knowl Data Eng*. 2015; 28(5): 1312-1326.
[24] Bagg, Carlo. "Ensemble Clustering of Financial Time Series", The 2024 RCEA International Conference on Economics, Econometrics and Finance At: Brunel University, London (UK), (2024).
[25] Rezaeipanah, A., Amiri, P., Nazari, H. *et al.* An Energy-Aware Hybrid Approach for Wireless Sensor Networks Using Re-clustering-Based Multi-hop Routing. *Wireless Pers Commun* 120, 3293–3314 (2021). https://doi.org/10.1007/s11277-021-08614-w.
[26] Vega-Pons S, Ruiz-Shulcloper J. A survey of clustering ensemble algorithms. *Int J Pattern RecognitArtifIntell*. 2011; 25(3): 337-372.
[27] Ma TH, Zhang YL, Cao J, Shen J, Tang ML, Tian Y, Al-Dhelaan A, Al-Rodhaan M. KDVEM: a k-degree anonymity with vertex and edge modification algorithm. Computing 2015;97(12):1165–84.
[28] Alizadeh H, Minaei-Bidgoli B, Parvin H. Cluster ensemble selection based on a new cluster stability measure. Intell Data Anal 2014;18(3):389–408.
[29] Ye M, Liu W, Wei J, Wei J, Hu X. Fuzzy c-means and cluster ensemble with random projection for big data clustering. Math Probl Eng 2016:1–13.

[30] Ma TH, Wang Y, Tang ML, Cao J, Tian Y, Al-Dhelaan A, Al-Rodhaan M. LED: A fast overlapping communities detection algorithm based on structural clustering. Neurocomputing 2016;207:488–500.
[31] Yu Z, Luo P, You J, Wong HS, Leung H, Wu S, Zhang J, Han G. Incremental semi-supervised clustering ensemble for high dimensional data clustering. IEEE Trans Knowl Data Eng 2016;28(3):701–14.
[32] Ma TH, Rong H, Ying CH, Tian Y, Al-Dhelaan A, Al-Rodhaan M. Detect structural-connected communities based on BSCHEF in c-DBLP. ConcurrenComputPract Ex 2016;28(2):311–30.
[33] Anandhi RJ, Natarajan S. Privacy protected mining using heuristic based inherent voting spatial cluster ensembles. Springer 2014;236:1183–93.
[34] Mahrooghy M, Younan NH, Anantharaj VG, Aanstoos J, Yarahmadian S. On the use of a cluster ensemble cloud classification technique in satellite precipitation estimation. IEEE J Sel Topics Appl Earth Observ Remote Sensing 2012;5(5):1356–63.
[35] Ahmadian S, Norouzi-Fard A, Svensson O, Ward J. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. Found Comput Sci 2017:61–72.
[36] Zhang L, Lu W, Liu X, Pedrycz W, Zhong C. Fuzzy c-means clustering of incomplete data based on probabilistic information granules of missing values. Knowl Based Syst 2016;99(C):51–70.
[37] Fern XZ, Brodley CE. Cluster ensembles for high dimensional clustering: an empirical study. Corvallis Or Oregon State University Dept of Computer Science; 2004. p. 1–26.
[38] Kuncheva LI, Hadjitodorov ST. Using diversity in cluster ensembles. IEEE Int Conf Syst, Man Cybern 2004;2:1214–9.
[39] Naldi MC, Carvalho AC, Campello RJ. Cluster ensemble selection based on relative validity indexes. Data Mining Know Discovery 2013;27(2):259–89.
[40] Li F, Qian Y, Wang J, Liang J. Multigranulation information fusion: a dempster-shafer evidence theory-based clustering ensemble method. Inf Sci 2016;1:58–63.
[41] Lakshmi, H. N., Thaduri Venkata Ramana, LNC Prakash K, L. Kiran Kumar Reddy, &Kachapuram Basava Raju. "A novel comprehensive investigation for enhancing cluster analysis accuracy through ensemble learning methods." International Journal of Electrical and Computer Engineering (IJECE) [Online], 14.5 (2024): 5802-5812. Web. 16 Dec. 2024
[42] K., L. P., Suryanarayana, G. ., Swapna, N. ., Bhaskar, T. ., & Kiran, A. . (2023). Optimizing K-Means Clustering using the Artificial Firefly Algorithm. International Journal of Intelligent Systems and Applications in Engineering, 11(9s), 461–468. Retrieved from https://www.ijisae.org/index.php/IJISAE/article/view/3154
[43] V. Gopala Krishnan, Dr & Sankar, K. &Saradhi, M. & Priya, K. &Vijayaraja, V.. (2023). Tom and Jerry Based Multipath Routing with Optimal K-medoids for choosing Best Cluster head in MANET. International Journal of Communication Networks and Information Security (IJCNIS). 15. 70-82. 10.17762/ijcnis.v15i1.5707.