Journal of Computer Science

Research Article

Improving Clustering Robustness through Fuzzy Ensemble of K-Means and Mean Shift

LNC. Prakash K.[1], Palamakula Ramesh Babu[2], Shaik Thaseentaj[3], C.V. Lakshmi Narayna[4] Ravikiranreddy Kandadi[5] and Kadiyala Ramana[6]

[1]*Department of Computer Science & Engineering, CVR College of engineering, Hyderabad, Telangana, India*
[2]*Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana, India*
[3]*Department of Computer Science & Engineering, KL University (KLEF), Green Fields, Vaddeswaram, Guntur District, Andhra Pradesh, India*
[4]*Department of Computer Science and Engineering, Annamacharya University, Rajampet, Andhra Pradesh, India*
[5]*Department of CSE-Data Science, CVR College of Engineering, Hyderabad, Telangana, India*
[6]*Department of Artificial Intelligence and Data Science, Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana, India*

Article history
Received: 20-03-2025
Revised: 05-11-2025
Accepted: 17-03-2026

Corresponding Author:
Kadiyala Ramana
Department of Artificial Intelligence and Data Science, Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana, India
Email: ramana.it01@gmail.com

Abstract: Ensemble clustering has emerged as a powerful strategy to improve the robustness and accuracy of unsupervised learning, particularly when individual algorithms struggle with noisy, heterogeneous, or high-dimensional data. This study introduces a fuzzy-based ensemble approach that integrates the complementary strengths of K-Means and Mean Shift clustering, followed by fuzzy membership assignment for data points that remain ambiguous. The inclusion of fuzzy logic provides a flexible mechanism to resolve uncertainty, ensuring that overlapping or irregularly shaped clusters are effectively managed. Experiments were conducted on three benchmark datasets-Weather History, Weather Prediction, and Dry Bean-using evaluation metrics such as the Silhouette Score and Davies–Bouldin Index. Results show that the proposed ensemble achieves consistent improvements over traditional clustering methods, with significant reductions in Davies–Bouldin Index and higher Silhouette Scores across datasets. These findings highlight the practical potential of the method for complex real-world applications and contribute to advancing ensemble clustering methodologies.

Keywords: Cluster Analysis, Artificial Intelligence, Ensemble Approaches, Precision, Patterns

Introduction

The exponential increase of data volume in the modern era has highlighted the necessity for reliable techniques to extract significant insights from large, complicated datasets. Unsupervised learning's clustering technique is a potent tool that makes it easier to find patterns, correlations, and structures in data without the requirement for labelled instances. The use of clustering is widespread, with applications ranging from image processing and social network analysis to biology and finance. It is important because it can help with data summarization, anomaly detection, and pattern recognition by providing a lens that can reveal the inbuilt organisation of data. The technique of locating similar data points alongside yet maintaining dissimilar points separate is known as clustering. The main goals of clustering are finding natural groups and improving the readability of the data. In clustering it is needed to describe the idea of distance measures, emphasising metrics like cosine similarity, Manhattan distance, and Euclidean distance as the cornerstone of methods for clustering. It is needed to highlight how distance measurements are used to establish how comparable two data points are to each other. The approaches for clustering have developed along with the number and complexity of datasets. While modern techniques like spectral clustering and affinity propagation (Dueck and Frey, 2007) exhibit the ongoing innovation in this discipline, earlier strategies like hierarchical clustering

(Pfeifer et al., 2021), set the foundation. The intent of this research is to evaluate clustering methods, emphasizing the innovations attained in modern clustering procedures and presenting a specified impression of their historical advancement. Moreover, it investigates the key advances, encounters, and trends that have modelled clustering methods over time. Clustering methods are generally treated across diverse fields due to variations in data appearances and analytical constraints. This paper investigates how clustering methods are utilized across different fields, modifying to the exclusive features of datasets in areas such as bioinformatics, genomics, marketing, consumer segmentation, and weather analysis.

Through comprehension of these multidisciplinary viewpoints, scholars can fully utilise ensemble clustering to address complicated, practical issues. To improve overall performance, cluster ensemble method is a potent machine learning technique which combine the output of several clustering algorithms. By decreasing the sensitivity to selecting a single clustering approach, cluster ensemble methods can improve the resilience of clustering findings. Combining algorithms might help reduce the drawbacks of using individual techniques because different algorithms may perform better on different parts of the data. Because clustering is responsive to the initial conditions, small changes in the input data or method settings could result in various results. By combining several results, ensemble approaches can offer clustering solutions that are more reliable and consistent. The right method for clustering for a given dataset can be difficult to find in practical applications. By combining several algorithms, ensemble approaches can be helpful in lowering the uncertainty involved in choosing a single approach. Using cluster ensembles, one can combine the output of various clustering methods to arrive at a consensus answer. The fundamental framework of the data may be more accurately represented by this consensus viewpoint. By treating the ensemble as a higher-level clustering technique, cluster ensemble approaches can be thought of as a type of meta-clustering. This may offer a more adaptable and flexible method of encoding intricate relationships within the data. Cluster ensemble methods are a useful tool in many applications because it has been demonstrated that they perform more accurately and robustly than individual algorithmic methods for clustering in many circumstances.

Ensemble clustering has a number of benefits, including improved generalisation and increased resilience, but it also has certain drawbacks. Usually, ensemble clustering bring about repeating an algorithm with a different initialization or executing several clustering algorithms. This may lead to an increase in computational complexity, particularly if the cost of the various clustering techniques is high. When compared to a single clustering technique, running multiple clustering algorithms concurrently or repeatedly may need additional computing power (CPU, memory, and storage). The selection of base clustering methods affects how well ensemble clustering performs. If the underlying algorithms are not varied or perform poorly on their own, the ensemble might not produce appreciable gains. A mixed collection of clusters that are frequently produced by ensemble clustering can be difficult to understand. Users may find it challenging to comprehend and evaluate the result if it offers unclear insights into the underlying structure of the data. The way the output of the various clustering methods is combined can have a big effect on how well the ensemble performs. Inadequate selection of the combination method could result in less-than-ideal outcomes. Ensemble clustering could be susceptible to data noise or anomalies. The ensemble outcome may suffer if individual clustering methods generate noisy or inaccurate clusters.

The selection of parameters for both the ensemble combining method and the underlying clustering algorithms can have an impact on the ensemble clustering performance. The challenge of optimizing these settings may not be simple. Combining different clustering techniques carries the risk of overfitting, particularly if the ensemble is overly customized to the particular dataset being used. Reduced generalization performance on fresh, untested data could come from this. It's crucial to remember that the accuracy and variety of the base clustering techniques, the type of data, and the ensemble mixture approach used all affect the efficacy of ensemble clustering works. Experimentation and thoughtful consideration are necessary when applying ensemble clustering approaches.

The novelty of this study lies in combining the deterministic nature of K-Means, the adaptability of Mean Shift, and the flexibility of fuzzy membership assignment into a single ensemble framework. Unlike previous ensemble approaches that rely primarily on voting or re-labeling, our method explicitly integrates fuzzy logic to resolve unassigned or ambiguous points, thus improving both accuracy and interpretability.

Related Work

Ensemble methods were first introduced and developed in supervised learning. Due to its success in classification tasks, academics have tried to apply this paradigm to unsupervised learning domains, focusing on clustering problems, for two reasons. First, unsupervised learning situations lack predetermined criteria for finding optimal data solutions and prior knowledge about the underlying structure or features to be recognized. Agárdi and Kovács (2022) proposed an automatic cluster count determination method. It presents classical clustering and improvement approaches and tests their efficacy.

Algorithm types, parameters, ensemble size, and dataset adjustments make clustering ensembles challenging to design but improve clustering. Tuysuzglu and Birant (2018) compared eight clustering ensembles utilizing k-means, expectation maximization, hierarchical, canopy, and farthest first methods. All ensemble findings are used for clustering, but ineffective solutions reduce accuracy. To achieve this, accuracy-based solution choice is intended. Results from 14 datasets show that improved ensembles locate latent designs better than single patterns. Ensemble models improve clustering performance, as shown in the conclusions. Though challenging, ensemble clustering improves clustering by combining different methods. Wang et al. (2023) proposed a divide-and-conquer parallel hierarchical clustering strategy for efficacy. A cluster consensus collecting method selects high-quality primary clusters for the final consensus. An unsupervised feature collection method removes unrelated features to improve grouping. In UCI dataset experiments, the proposed methodology improves accuracy by 6% to 24% over state-of-the-art solutions.

Clustering ensemble, or consensus clustering, integrates various clustering findings to improve resilience and stability. This notion is expanded by weighting clustering, or characteristics, by relevance and quality. Zhang (2022) discussed weighting strategies, weight value determination, and complex data applications. Practitioners can choose effective clustering task weighting mechanisms using the framework. In essence, a professional clustering ensemble should deliver more stable, accurate, and predictable outcomes than individual algorithms. However, switching from supervised to unsupervised learning is more complicated than a conceptual explanation. Clustering ensembles pose unique and difficult difficulties. Merging clusters from different ensemble clustering algorithms (members) is the first and most demanding challenge. This amalgamation requires more complex aggregating processes than categorization; voting and averaging cannot do it. Many concepts have been proposed, but no scalable and usable consensus function exists for efficient use in this setting. Scholars have chosen ways that match application features. Multiview clustering (MVC) seeks to identify shared cluster compositions within observations, however existing methods often neglect elementary divisions. A reduced Multiview ensemble clustering (M2VEC) technique increases clustering robustness using low-rank and light-staleness denoising (Tao et al., 2020). A multilayer stacked model incorporates nonlinear relationships, whereas spectral consensus graph partitioning combines several viewpoints into an optimal structure. Experimental results on eight real-world datasets demonstrate M2VEC outperforms state-of-the-art multiview and ensemble clustering methods, even with partial data. Shi et al. (2023) proposed BoostAEC, an adaptive ensemble clustering system that uses BoostBLSAE to build base partitions and adaptive weights. A consensus goal function enhances the co-association matrix, whereas a fuzzy membership function defines inter-cluster connections. The suggested strategy yields high-dimensional noise and partition optimization. Massive real-world dataset experiments show that BoostAEC outperforms state-of-the-art ensemble clustering.

Three cluster ensembles were first proposed by Strehl and Ghosh (2002). The first technique was Cluster-Based Partitioning Similarity (CSPA), which uses data point similarity. Whether data points were comparable or different. The second method, hyper graph partitioning (HGPA), re-segmented data using clusters. The final algorithm, meta-clustering (MCLA), grouped clusters and symbolised them with hyper edges. Alqurashi and Wang (2019) proposed the Adaptive Clustering Ensemble, which uses a new consensus function. A newly designed membership similarity and cluster similarity are used to compare. Its three stages are dynamic. The first stage binaryizes the clusters. The programme then creates the first few most similar clusters using the cluster similarity metric in the second stage. Repeat this adaptive iteration until the desired candidate clusters are produced. In the third stage, ambiguous objects are addressed to refine the clusters, resulting in a better clustering result with the desired number of clusters. Vega-Pons and Ruiz-Shulcloper (2011) provided an overview of clustering ensemble techniques that help the cluster analysis community choose the optimal approach to a problem. We also classify techniques and highlight essential uses for the appropriate number of clusters. The comparative research by Kalaiselvi and Karthika (2018) summarized high-dimensional data space results and their implications for clustering techniques. It also analyses a variety of clustering algorithms, including subspace approaches, model-based clustering, density-based clustering, partition-based clustering, and others. This provides a more complete overview of existing studies' pros and cons for solving problems with higher-dimensional data.

Wu et al. (2018) reviewed clustering ensemble research, focusing on consensus functions, generating mechanisms, and selective clustering ensembles. It examines twelve ensemble clustering methods to find the simplest. The best results are obtained with average-linkage agglomerative clustering as the consensus function and k-means with multiple initializations as the generating mechanism. Hashim and Muhammed (2022), recommended using three strategies (Wang and Chen, 2020; Wang et al., 2017; 2019) as an ensemble approach with the k-means algorithm to boost performance and lessen initial centroids' impact. After that, a common dataset was utilized to compare performance to typical k-means results. Ghorbanian and Razavi (2023) employed three methods. Rand's external criteria and statistical

analysis required the first stage to find the optimal algorithm configuration from the two created algorithms. Comparing this configuration to literature methodologies focused on clustering accuracy and execution time. Results showed that the proposed strategy was more accurate and faster. The article describes a three-step time series clustering method that is fast and accurate. This study by Abbasi et al. (2019) is the first to merge segmentation with ensemble clustering, improving accuracy and execution time. Lack of prior knowledge makes clustering analysis challenging. A single clustering method may not always reflect structural information well. Changing the initial values can make it harder to get good clusters, even with the same method. New ways to estimate individual clusters have been developed since conventional approaches may ignore good clusters if the partition is bad. Edited Normalized Mutual Information (ENMI) improves cluster excellence measurement (Jiang et al., 2022). A new clustering ensemble method uses Average ENMI to choose the best clusters for final findings. These selected clusters are connected using co-association, hypergraph partitioning, and feature-space clustering. Investigations show this method outperforms ensemble clustering. Jiang et al. (2022) proposed S-M3WCE, a clustering ensemble that manages data ambiguity and noise to improve robustness. Possibilistic C-Means clustering (PCM) groups objects into core, shadowed, and exclusion regions, then improves them into four multi-granularity rough set-based approximation districts. To create clustering results, shadowing sets are applied to these locations. Multi-dataset experiments reveal that S-M3WCE outperforms six clustering ensemble approaches in accuracy and efficiency.

Ensemble clustering yields more accurate, reliable, and resilient results than single clustering. Recently, ensemble clustering has received more attention, leading to numerous new approaches (Huang et al., 2016; 2018). A novel clustering ensemble paradigm based on granular computing (Xu and Ding, 2021). chooses diversified and high-quality base clustering via granularity distance to promote clustering excellence. Base clustering consistency is ensured while maximizing variations to improve accuracy. Optimization of co-association matrix construction boosts sample similarity, supporting real data forms. Structured hypergraph learning is used to create a clustering ensemble technique (Zhou et al., 2022). The hypergraph is dynamically learned from base clustering findings. A clear clustering structure eliminates the need for uncertain postprocessing like hypergraph partitioning, improving consistency. This method is more permanent and robust than hypergraph-based ensemble methods. Random Sample Partition-based Centres Ensemble (RSPCE) is used to count clusters in a large dataset (Mahmud et al., 2023). The programme selects non-overlapping random samples from the vast dataset. Each sample's clusters and initial centers are determined by the I-niceDP method. A cluster ball model is then used to merge two random samples from the same dataset cluster. Finally, utilizing the ball model, the RSPCE ensemble technique integrates all sample solutions to provide a set of initial cluster centers in the large dataset. A CFA-EDL technique for MANET intrusion detection using ensemble clustering for cluster head selection is presented in Krishnan et al. (2023). EDL Classifier enhances memory, IDS traffic, and identification. Zhao et al. (2024) suggested a multi-view ensemble clustering strategy that improves clustering by analyzing basis clusters using joint entropy. Define an uncertainty index to analyze cluster characteristics, then hypothesise a weighted co-association matrix with deleted incorrect entries. Constancy index and MvEC-DoS algorithms add candidate cluster collection to the clustering process. The recommended approach's dominance over multi-view clustering methods is determined by investigational solutions.

This research (Wang and Chen, 2020) presents a unique assertion technique for missing data and a three-way ensemble clustering approach based on imputed conclusions. Imputation clusters complete data using strong clustering and replaces missing values with cluster mean attributes, facilitating perturbation investigation. Based on clustering consensus, three-way ensemble clustering assigns objects to core or fringe regions. The three-way clustering approach (Li et al., 2024) improves clustering strength and robustness by reducing dimensionality, randomizing, and using ensemble schemes. Co-association frequency and hierarchical clustering create significant cluster illustrations of core and outlying areas. By defining lower and higher calculations, each cluster is naturally described three-way. Clustering Ensemble (CE) is noted for its great consensus capability but struggles with selection strategies, co-association matrix generation, and consensus outcome dispute resolution. Shan et al. (2023) proposed a dual-level clustering ensemble technique with three consensus strategies to address these difficulties. A backward clustering ensemble selection framework adaptively eliminates redundant members, and two updated relation matrices increase base clustering consensus. The Dempster–Shafer evidence theory is also adjusted to dynamically integrate ensemble results. According to Lakshmi et al. (2024), ensemble clustering improves cluster analysis accuracy for complicated datasets. Comparisons are made between majority voting-based ensemble clustering algorithms including k-means, affinity propagation, mean shift, and BIRCH and discrete approaches. A BIRCH-mean shift collective model improves clustering strength and pattern identification. Fuzzy-based approaches are significant in this article since they assess the possibility of a data object belonging to the base cluster. The recommended methodology is as follows.

Recent research has emphasized ensemble-based

clustering as an effective solution for high-dimensional, noisy, and heterogeneous datasets, where single algorithms often fail to deliver stable and meaningful partitions (Zhang, 2022; Tao et al., 2020; Shi et al., 2023). For example, Shi et al. (2023) introduced an adaptive ensemble clustering approach that integrates boosting with a broad learning system-based autoencoder, demonstrating significant improvements in clustering high-dimensional noisy data. Similarly, Zhang (2022) provided a comprehensive review of weighted ensemble frameworks, where partitions are assigned adaptive weights based on their quality, leading to enhanced robustness and accuracy compared to uniform ensemble strategies. Tao et al. (2020) further advanced the field through marginalized multiview ensemble clustering, which effectively captures complementary information across different feature spaces and is particularly useful for complex real-world data such as multimedia and sensor networks.

Despite these advances, most existing ensemble clustering models are grounded in deterministic consensus strategies such as majority voting, co-association matrices, or relabeling mechanisms. These methods treat cluster assignments as crisp and final, overlooking the inherent uncertainty present in ambiguous data points that could reasonably belong to multiple clusters. This limitation becomes critical in domains such as bioinformatics, text mining, and social network analysis, where overlapping clusters and soft boundaries are common. Fuzzy clustering, most notably Fuzzy C-Means (FCM), has long demonstrated the ability to address this challenge by assigning degrees of membership rather than hard labels. However, the explicit integration of fuzzy logic into ensemble clustering frameworks has been relatively limited and often restricted to the aggregation stage.

Our approach fills this gap by embedding fuzzy-based decision-making directly into the ensemble process, enabling a more nuanced treatment of overlapping and ambiguous data. By first combining the outputs of K-Means and Mean Shift to identify consensus clusters and then applying fuzzy membership allocation to unassigned or uncertain points, the proposed method introduces flexibility where deterministic consensus methods fall short. This integration ensures that the final clustering solution not only inherits the robustness of ensemble learning but also reflects the uncertainty inherent in real-world datasets, thereby achieving both accuracy and interpretability.

Methods

Clustering is rarely used alone, and it is frequently combined with other analytical techniques like feature engineering and dimensionality reduction. The symbiotic ties between clustering and complementing procedures will be clarified in this study, along with how their integration improves data analysis's overall effectiveness. We will illustrate with examples how clustering functions as a fundamental component of more comprehensive quantitative methodologies. In this research a final clustering result is obtained using ensemble clustering, which combines many clustering techniques on the same dataset and the aim is enhancing individual data clustering quality. Our technique to ensemble clustering in this study is based on majority voting. To achieve distinct outcomes, this entails using certain clustering approaches on predefined data. We contrast the outcomes of various strategies and the ensemble clustering method after implementation. Improved clustering results using ensemble clustering are shown by the results. Figure 1 poses an extensive general idea of a research analysis focused on improving the precision of cluster recognition. The initial stage involves gathering data, followed by several steps of data preparation and organization before clustering begins. During preprocessing, irrelevant elements like empty entries, unrecognized values, and insignificant attributes that have minimal impact on the outcomes are eliminated. Additionally, inconsistencies are penetrated out to create a robust clustering paradigm.

Prior to utilizing various clustering procedures to the dataset, the initial task involves governing the optimal number of groups required to efficiently segment the dataset. Before assessing the clustering models, standard algorithms such as k-means, Affinity propagation and Mean Shift clustering are employed to generate preliminary clustering results.

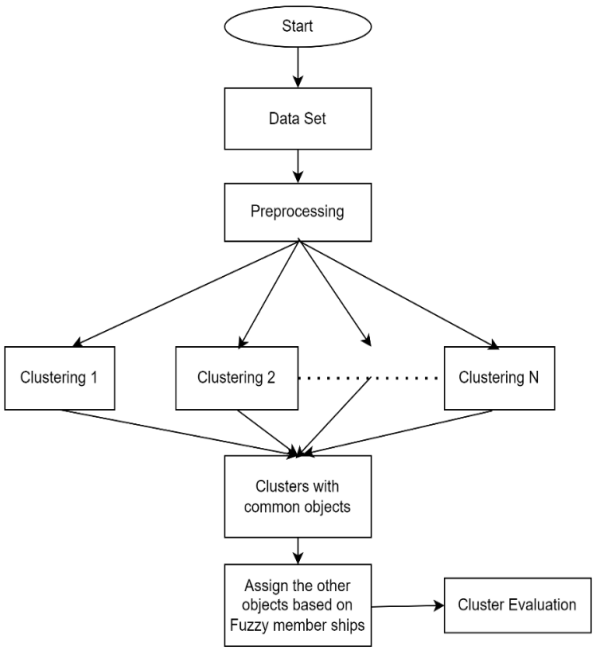

Fig. 1: Proposed methodology of ensemble clustering

DOI: 10.3844/jcssp.2026.1521.1531

These results serve as a baseline for comparison with the proposed technique. In utilizing an integrated model for this study, a hybrid model is structured by mixing the individual clustering approaches. This hybrid approach capitalizes on the métiers of each clustering technique to improve the whole toughness and accurateness of clustering.

Mean Shift Clustering

Mean shift clustering is a non-parametric type of clustering grouping approach; Mean Shift doesn't require knowledge on the number of clusters in advance. It activates by continually changing data points in the target of the establishing probability distribution's mode, or peak. The next is a described procedure for Mean Shift clustering:

1. Input: Let the data objects $Y = \{y_1, y_2 \ldots . y_n\}$, Kernel function K with bandwidth h and Convergence limit δ
2. Initialization: Initialize group midpoints $C = Y$ and shift vector $Shift = Zeros$
3. Mean Shift Iterations: Replicate till convergence
4. $for\ (every\ y_i \in \mathrm{Y})$
5. Compute the mean shift using:

$$Shift(y_i) = \frac{\sum_{j=1}^{n} K\left(\frac{y_i - y_j}{h}\right) * y_j}{\sum_{j=1}^{n} K\left(\frac{y_i - y_j}{h}\right)} \quad (1)$$

6. Update cluster centers $C = C + Shift$
7. Convergence check: $If(\|Shift\| < \delta)$, terminate the loop
8. Allocate every data instant y_i to the group whose center it converged
9. Return the clusters C

The kernel function K is generally a Gaussian kernel; though different options may be studied based on the data's properties. The bandwidth 'h' directs the extent of the neighborhood applied for approximating local density, affecting cluster shape and count. The convergence threshold ϵ signifies if the mean shift repetitions become halt. The shift vector indicates both the trend and scale of mean shift for every object. The procedure continually changes data points toward the mode of the fundamental denseness till convergence is accomplished.

K-Means Clustering

The unsupervised machine learning procedure for classifying observations into k groups is K-means segmentation. Data objects are recurrently allocated to the nearest centroid, and centroids are renewed based on the average of the objects 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:

- Initialization: Select k randomly chosen centroids of the initial groups from the data points. The clusters' primary centers will be these the centroids
- Assigning: Place each data point in relative to the closest centroid. To allocate the data object to the cluster whose centroid is nearest, this action contains calculating the distance involving each data object and each centroid
- Update Centroids: Repeat the finding of the group centroids applying the average of all the data objects distributed to every cluster. To do this, the centroid must be keep informed with the mean location of each cluster's data points
- Repetition and update: 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
- Convergence: The final cluster positions are finalized at this point in the procedure, and the centroids are the cluster centers

Embedded Methodology

This strategy combines the mean shift and K-Means techniques to create realistic groups. First, all pairs of items in the dataset are compared to see if they are part of the same cluster with respect to the base clustering techniques.

They are placed together if they consistently fall into the same cluster across all techniques; if any pair does not belong to a single the individual data objects of that pair are checked further with other combination. Any entities in the dataset that remain after classifying data items into a predefined number of clusters are analyzed. When an item in the dataset is not allocated to any cluster, it is allocated to one of the clusters that already exist grounded on its degree of similarity to the cluster the centroids, Algorithm is an illustration of this strategy.

As per the algorithm's instructions, let the dataset be D, the set of clustering methods as $M = \{M_i \backslash M_i\ is\ a\ clustering\ method\}$, the set of all base clusters as $C = \{C_{ij} \backslash C_{ij}\ is\ j^{th}\ th\ cluster\ in\ i^{th}\ method\}$, and the set of resulting optimal clusters as $K = \{K_l \backslash l = 1,2, \ldots number\ of\ clusters\}$, In this study, various clustering methods including K-Means grouping, Mean Shift grouping, Agglomerative Clustering, were employed and evaluated individually on the specified datasets. Additionally, for ensemble clustering, in this research Mean Shift Clustering and K-Means Clustering

were used as base clustering methodologies and combined to enhance clustering accuracy as demonstrated in Fig. 1.

Algorithm

Input: Database D, Clusters resulting from of all base cluster methods.

Output: optimal clusters obtained from embedded approach.

Let,

$M = \{M_i \backslash M_i \; is \; a \; clustering \; method\}$

$C = \{C_{ij} \backslash C_{ij} \; is \; j^{th} \; th \; cluster \; in \; i^{th} \; method\}$,

$K = \{K_l \backslash l = 1,2, \ldots number \; of \; clusters\}$,

1. Set $l = 1, count = 0.$ //setting initial values to variables
2. $while(l \neq |K|)$
3. $K_l = \emptyset.$
4. $for(each \; r_x \epsilon \; D)$
5. $for(each \; r_y \epsilon \; D, x \neq y)$ //for all pairs of entities.
6. $for(i = 1 \; to \; |M|)$
7. $for(j = 1 \; to \; |C|)$ // consider all clusters from all method
8. $if(\{r_x, r_y\} \epsilon \; C_{ij})$ // if each pair of objects goes to same cluster
9. $Count = count + 1$
10. $End \; of \; sep \; 8.$
11. $End \; of \; sep \; 7.$
12. $End \; of \; sep \; 6.$
13. $if(count = |M|)$
14. $K_l = K_l \cup \{r_x, r_y\}.$ // Assigning couple of tuples to same cluster if they belong to same cluster in all methods.
15. $End \; of \; sep \; 13.$
16. $End \; of \; sep \; 5.$
17. $End \; of \; sep \; 4.$
18. $count = 0.$
19. $l = l + 1.$
20. $D = D - K_l.$
21. $End \; of \; sep \; 2.$
22. $if(D \neq \emptyset)$
23. Get the centroid of every cluster K_l.
24. $for(each \; r_x \epsilon \; D)$ // Allocating the leftover objects to related clusters
25. $for(each \; k_l \epsilon K)$
26. Find the membership of r_x with respect to k_l using equation 2. $End \; of \; sep \; 25.$
27. $K_l = \{K_l \cup r_x \backslash$ membership of r_x with respect to k_l is maximum$\}$.
28. $End \; of \; sep \; 24.$
29. Resume set of ensemble clusters, $K = \{K_l \backslash l = 1,2, \ldots |C|).$

The proposed fuzzy-based ensemble clustering framework proceeds as follows:

(i) K-Means is applied to partition the dataset into k clusters
(ii) Mean Shift clustering is applied to the same dataset using bandwidth h estimated via Scott's rule
(iii) Data points consistently clustered by both algorithms are directly assigned
(iv) Ambiguous points are evaluated using a fuzzy membership function with fuzzifier m = 2
(v) Final assignments are made to the cluster with maximum membership value

In this research the partial clusters initially are formed using mean shift and K- means clustering procedures as described in the algorithm. The remaining objects that are not part of any of the initial clusters are assigned the respective clusters based on the Fuzzy based association of every object with respect to each of existing initial clusters. The Fuzzy membership μ_{ij}, is computed by using Equation 2, in which the balancing factor m is considered as 2:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} (\frac{d_{ij}}{d_{ik}})^{\frac{2}{m-1}}}$$

Where:

μ_{ij} = Member ship of i^{th} with j^{th} cluster
d_{ij}= Distance from i^{th} object to j^{th} cluster
d_{ik}= Distance from i^{th} object to k^{th} cluster
C= Number of clusters

Experimental Analysis and Performance Evaluation

This section attends details of the experimental examination overseen via a selection of selected datasets, along with an in-extensive consideration of the evaluation conclusions. To confirm a complete evaluation, the study exploits numerous datasets, including the Weather History dataset, which accounts past weather forms, the Weather Prediction dataset, which emphases on estimating future circumstances, and the Dry Bean dataset, which holds agricultural data for classification errands. These datasets deliver various testing circumstances, accepting for a vigorous confirmation of the anticipated procedure.

The proposed fuzzy-based ensemble clustering method was implemented in Python using the scikit-learn library. All experiments were executed on a system with an Intel Core i7 processor, 16 GB RAM, and Windows 11 OS. Each experiment was repeated 10 times with random initialization to minimize bias, and average results are reported.

Data Set Description

A dataset is the base of any machine learning model, presenting the data took for training, validation, and testing. It impacts model execution through feature selection, data quality, and allocation. In this research Weather History, Weather Prediction and dry bean databases were used in the extensive testing. Historical weather data for several locations may be found in the Historical Weather Database. It includes comprehensive data regarding meteorological conditions that were observed during a given time frame.

There are 96,453 recordings in all, each of which has a distinct time stamp accompanied by meteorological data. Four clusters are the ideal amount for this dataset, it has been found by looking at visual representations and drawing conclusions. This indicates that the most important and representative clustering of data points within the given dataset and context is supplied by separating the data into four clusters according to the specified criteria. The 18 distinct European sites from which the weather data was gathered are included in the Weather Forecast dataset. This dataset, which spans the years 2000–2010, has 3654 daily recordings. It includes a number of variables, including the highest, lowest, and mean temperatures, among others. For this dataset, two clusters is the optimal number. This indicates that, given the dataset and context, the data is most accurately and representatively grouped into two clusters based on the chosen criteria. 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 procedure was established to recognize between these seven documented kinds of dry beans that have comparable attributes. Utilizing an extreme-end camera, the arrangement took pictures of 13,611 individual beans from these seven recognized kinds. After segmenting and extracting features from the pictures acquired by the computer vision system, a total of 16 characteristics 12 dimensions and 4 form categories were identified from the beans.

For K-Means, the number of clusters k was set according to the known class distributions in each dataset. Mean Shift employed bandwidth h estimated via Scott's rule of thumb. The fuzzifier parameter was fixed at m=2, consistent with standard fuzzy clustering practice.

The ensemble method was compared against four established clustering algorithms: K-Means, Mean Shift, Agglomerative Clustering, and DBSCAN. Performance was evaluated using the Silhouette Score and Davies–Bouldin Index, ensuring both internal compactness and separation were measured.

To validate improvements, statistical significance was assessed using the Wilcoxon signed-rank test with a confidence level of 95% ($p<0.05$). This ensured that observed performance differences were not due to random variation

Exploration of Performance in Investigation

The experimental inquiry, a few chosen datasets, and the evaluation's findings are all covered in depth in this part. Weather History, Weather Prediction and dry bean databases were used in the extensive testing. The vast tests were conducted primarily in two domains. First, an analysis was carried out to determine how accurate the traditional techniques of clustering like K-Means, Mean Shift, and BIRCH clustering. An acceptable distance measure was used in this assessment to extract data for grouping from the dataset. To emphasize the importance of the anticipated methodology, succeeding investigations incorporated clustering as a crucial stage. An ensemble approach merging K-Means and Mean Shift clustering techniques was exercised to improve the clustering performance. To judge the efficiency of this method, metrics such as the Silhouette Score and the Davies-Bouldin Score were evaluated in the methodology section, evaluating the results of the ensemble methodology with those of traditional clustering methods.

From Table 1, for the Weather History dataset, the Davies-Bouldin Score and the Silhouette Score were used to evaluate the clustering methods based on how expertly they divided the dataset into 4 groups. With the lowest Davies-Bouldin Score of 0.184 and a respectable Silhouette Score of 0.873, the Ensemble method stood out as the best performance. This implies that the clustering was effective and well-separated, with little intersection across clusters.

Using the Davies-Bouldin Score and the Silhouette Score, the clustering procedures were assessed overall based on how appropriately they divided the dataset into 3 groups as shown in Table 2 and Figs. 2 and 3.

Table 1: Ensemble vs. Traditional Clustering on Weather Data

Clustering Algorithm	Number of clusters	Davis Bouldin Score	Silhoutte Score
K-Means	4	0.401	0.608
Mean Shift	4	0.435	0.857
Agglomerative	4	0.405	0.588
Ensemble	4	0.184	0.873

Table 2: Comparing Ensemble Clustering to Traditional Models on Weather Prediction data

Clustering algorithm	Number of Clusters	Davis Bouldin Score	Silhoutte Score
KMeans	3	0.937	0.414
Mean Shift	3	0.955	-0.002
Agglomerative	3	1.021	0.354
Ensemble	3	0.683	0.427

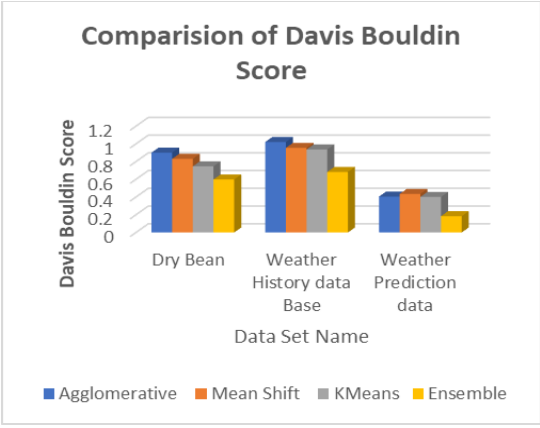

Fig. 2: Comparison of Davis Bouldin Score for different datasets

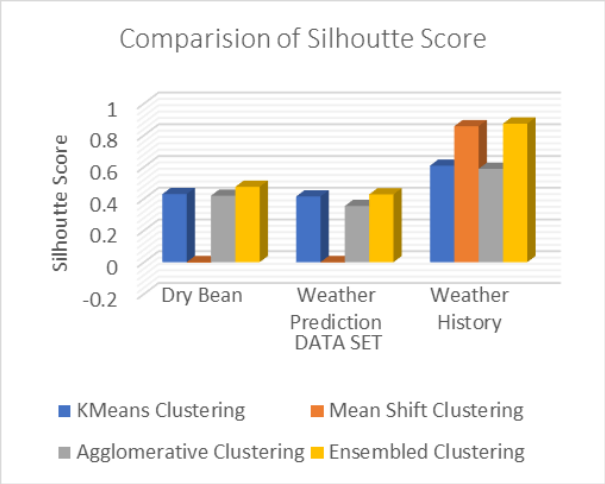

Fig. 3: Comparison of Silhoutte Score for different datasets

With a low Davies-Bouldin Score of 0.683 and a reasonable Silhouette Score of 0.427, the Ensemble approach clearly performed best, showing efficient and well-separated clustering with little common ground among clusters.

From Table 3, for the dataset Dry bean, the Davies-Bouldin Score and the Silhouette Score were used to evaluate the clustering methods based on how well they divided the dataset into 7 groups. With the lowest Davies-Bouldin Score of 0.598 and a respectable Silhouette Score of 0.475, the Ensemble technique stood out as the best performance. This indicates that the clustering was efficient and well-separated, with little overlap across clusters. For better and easy understanding of the results the graphs are represented by picture 1 and picture 2 are drawn. As seen visually by the bar graph the Ensemble method outperformed other algorithms in both Davies-Bouldin and Silhouette Scores. Additionally, it would highlight how poorly the Mean Shift method performed, particularly when the Silhouette Score was negative and showed overlapping clusters and inaccurate clustering assignments. It would be clear that other techniques performed moderately, with only minor differences in clustering quality as shown by their individual scores. Overall, the performance of the clustering procedures established on the assessment measures could be quickly and easily compared by the graphical depiction.

Table 3: Comparing Ensemble Clustering to Traditional Models on Dry bean data

Clustering Algorithm	Number of Clusters	Davis Bouldin Score	Silhoutte Score
KMeans	7	0.746	0.429
Mean Shift	7	0.831	-0.003
Agglomerative	7	0.901	0.419
Ensemble	7	0.598	0.475

The findings imply that the ensemble model integrates the outputs of Mean Shift and K-Means, two well-known clustering methods. Using a common object process along with the member ship mechanism, this method produces the final outcomes. These techniques are good at detecting real clusters and work especially well with a variety of sized and shaped clusters, which makes them appropriate for collecting intricate data structures. Mean Shift excels in catching complex patterns, whereas K-Means provides a good partition and computing efficiency along with scalability. Because of this combination, the model can handle huge datasets with less computing effort, which makes it a viable option in situations when computational resources are limited.

Across all datasets, the proposed fuzzy-based ensemble achieved the highest Silhouette Scores and lowest DBI values, indicating more coherent and well-separated clusters than baseline algorithms. The improvement was particularly notable on the Dry Bean dataset, where cluster structures are complex and overlapping. Here, the fuzzy membership assignment played a critical role in properly allocating ambiguous data points.

Conclusion

This study introduced a fuzzy-based ensemble clustering approach that combines the complementary strengths of K-Means and Mean Shift with fuzzy membership assignment to handle ambiguous data points. The experimental evaluation on three benchmark datasets-Weather History, Weather Prediction, and Dry Bean-demonstrated that the proposed ensemble consistently outperformed traditional clustering methods by achieving lower Davies–Bouldin Index values and higher Silhouette Scores. These results confirm that the integration of deterministic and fuzzy decision-making not only improves robustness but also enhances the interpretability of clustering outcomes. Despite these promising findings, the method requires higher computational resources due to the dual execution of clustering algorithms and the fuzzy allocation process, which may limit scalability in very large datasets. Future research will focus on developing automated parameter tuning strategies, extending the approach to distributed and parallel computing frameworks for big data applications, and exploring integration with deep learning–based clustering methods. Additionally, applying the proposed model to domain-specific problems such as bioinformatics, image segmentation, and social network analysis will further demonstrate its adaptability and real-world relevance.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

LNC Prakash K.: Conceptualization, data curation, formal analysis, methodology, software, write original draft.

Palamakula Ramesh Babu: Supervision, write review and edited, Project administration, Visualization.

Shaik Thaseentaj: Supervision, Project administration, write review and edited.

C.V. Lakshmi Narayna and Ravikiranreddy Kandadi: Write review and edited, supervision, project administration, formal analysis.

Kadiyala Ramana: Visualization, investigation, formal analysis, software.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abbasi, S., Nejatian, S., Parvin, H., Rezaie, V., & Bagherifard, K. (2019). Clustering ensemble selection considering quality and diversity. *Artificial Intelligence Review*, *52*(2), 1311–1340. https://doi.org/10.1007/s10462-018-9642-2

Agárdi, A., & Kovács, L. (2022). Clustering algorithms with prediction the optimal number of clusters. *Journal of Applied Research and Technology*, *20*(6), 638–651. https://doi.org/10.22201/icat.24486736e.2022.20.6.1077

Alqurashi, T., & Wang, W. (2019). Clustering ensemble method. *International Journal of Machine Learning and Cybernetics*, *10*(6), 1227–1246. https://doi.org/10.1007/s13042-017-0756-7

Dueck, D., & Frey, B. J. (2007). Non-metric affinity propagation for unsupervised image categorization. *2007 IEEE 11th International Conference on Computer Vision*, 1–8. https://doi.org/10.1109/iccv.2007.4408853

Ghorbanian, A., & Razavi, H. (2023). A new method based on ensemble time series for fast and accurate clustering. *Data Technologies and Applications*, *57*(5), 756–779. https://doi.org/10.1108/dta-08-2022-0300

Hashim, D. K., & Muhammed, L. A. N. (2022). Performance of K-means algorithm based an ensemble learning. *Bulletin of Electrical Engineering and Informatics*, *11*(1), 575–580. https://doi.org/10.11591/eei.v11i1.3550

Huang, D., Lai, J., & Wang, C.-D. (2016). Ensemble clustering using factor graph. *Pattern Recognition*, *50*, 131–142. https://doi.org/10.1016/j.patcog.2015.08.015

Huang, D., Wang, C.-D., & Lai, J.-H. (2018). Locally Weighted Ensemble Clustering. *IEEE Transactions on Cybernetics*, *48*(5), 1460–1473. https://doi.org/10.1109/tcyb.2017.2702343

Jiang, C., Li, Z., & Yao, J. (2022). A shadowed set-based three-way clustering ensemble approach. *International Journal of Machine Learning and Cybernetics*, *13*(9), 2545–2558. https://doi.org/10.1007/s13042-022-01543-5

Kalaiselvi, K., & Karthika, D. (2018). Review of Traditional and Ensemble Clustering Algorithms for High Dimensional Data. *SSRN Electronic Journal*, *6*, 1–8. https://doi.org/10.2139/ssrn.3170321

Krishnan, V. G., Dr, a, P., Abdul, Kirubakaran, N., Sankaradass, V., Kumar, A., Jehan, C., Deepa, J., & Dhanalakshmi, G. (2023). Ensemble Deep Learning Classifier with Optimized Cluster Head Selection for NIDS in MANET. *Journal of Information Science and Engineering*, *36*(6), 1233–1246.

Lakshmi, H. N., Ramana, T. V., Prakash K, L., Reddy, L. K. K., & Raju, K. B. (2024). A novel comprehensive investigation for enhancing cluster analysis accuracy through ensemble learning methods. *International Journal of Electrical and Computer Engineering (IJECE)*, *14*(5), 5802–5812. https://doi.org/10.11591/ijece.v14i5.pp5802-5812

Li, A., Meng, Y., & Wang, P. (2024). Similarity-Based Three-Way Clustering by Using Dimensionality Reduction. *Mathematics*, *12*(13), 1951. https://doi.org/10.3390/math12131951

Mahmud, M. S., Huang, J. Z., Ruby, R., & Wu, K. (2023). An ensemble method for estimating the number of clusters in a big data set using multiple random samples. *Journal of Big Data*, *10*(1), 1–29. https://doi.org/10.1186/s40537-023-00709-4

Pfeifer, B., Voicu-Spineanu, A., Schimek, M. G., & Alachiotis, N. (2021). Integrative hierarchical ensemble clustering for improved disease subtype discovery. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 337–342. https://doi.org/10.1109/bibm52615.2021.9669608

Shan, Y., Li, S., Li, F., Cui, Y., & Chen, M. (2023). Dual-level clustering ensemble algorithm with three consensus strategies. *Scientific Reports*, *13*(1), 22756. https://doi.org/10.1038/s41598-023-49947-9

Shi, Y., Yang, K., Yu, Z., Chen, C. L. P., & Zeng, H. (2023). Adaptive Ensemble Clustering With Boosting BLS-Based Autoencoder. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12369–12383. https://doi.org/10.1109/tkde.2023.3271120

Strehl, A., & Ghosh, J. (2002). Cluster ensembles a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*(Dec), 583–617. https://doi.org/153244303321897735

Tao, Z., Liu, H., Li, S., Ding, Z., & Fu, Y. (2020). Marginalized Multiview Ensemble Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(2), 600–611. https://doi.org/10.1109/tnnls.2019.2906867

Tuysuzglu, G., & Birant, D. (2018). Comparison of Different Clustering Ensembles by Solution Selection Strategy. *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, 201–206. https://doi.org/10.1109/ubmk.2018.8566542

Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A Survey of Clustering Ensemble Algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, *25*(03), 337–372. https://doi.org/10.1142/s0218001411008683

Wang, P., & Chen, X. (2020). Three-Way Ensemble Clustering for Incomplete Data. *IEEE Access*, *8*, 91855–91864. https://doi.org/10.1109/access.2020.2994380

Wang, P., Liu, Q., Yang, X., & Xu, F. (2017). Ensemble Re-clustering: Refinement of Hard Clustering by Three-Way Strategy. *Rough Sets: International Joint Conference, IJCRS 2017, Olsztyn, Poland, July 3–7, 2017, Proceedings*, *10313*, 423–430. https://doi.org/10.1007/978-3-319-67777-4_37

Wang, P., Shi, H., Yang, X., & Mi, J. (2019). Three-way k-means: integrating k-means and three-way decision. *International Journal of Machine Learning and Cybernetics*, *10*(10), 2767–2777. https://doi.org/10.1007/s13042-018-0901-y

Wang, Y., Krishna Saraswat, S., & Elyasi Komari, I. (2023). Big data analysis using a parallel ensemble clustering architecture and an unsupervised feature selection approach. *Journal of King Saud University - Computer and Information Sciences*, *35*(1), 270–282. https://doi.org/10.1016/j.jksuci.2022.11.016

Wu, X., Ma, T., Cao, J., Tian, Y., & Alabdulkarim, A. (2018). A comparative study of clustering ensemble algorithms. *Computers & Electrical Engineering*, *68*, 603–615. https://doi.org/10.1016/j.compeleceng.2018.05.005

Xu, L., & Ding, S. (2021). A novel clustering ensemble model based on granular computing. *Applied Intelligence*, *51*(8), 5474–5488. https://doi.org/10.1007/s10489-020-01979-8

Zhang, M. (2022). Weighted clustering ensemble: A review. *Pattern Recognition*, *124*, 108428. https://doi.org/10.1016/j.patcog.2021.108428

Zhao, X., Niu, X., Ma, Y., & Zhang, J. (2024). A multi-view ensemble clustering approach using joint entropy. *Expert Systems with Applications*, *255*, 124683. https://doi.org/10.1016/j.eswa.2024.124683

Zhou, P., Wang, X., Du, L., & Li, X. (2022). Clustering ensemble via structured hypergraph learning. *Information Fusion*, *78*, 171–179. https://doi.org/10.1016/j.inffus.2021.09.003