

## WEIGHT OPTIMIZATION FOR ENSEMBLE CLUSTERING USING THE MARINE PREDATOR ALGORITHM FOR SPECTRAL DATA

Anwiti Jain<sup>1</sup>, Dinesh Kumar Sahu<sup>2</sup>

<sup>1</sup>Department of Computer Science & Engg., SRK University, Bhopal, India.

[anwitijain@gmail.com](mailto:anwitijain@gmail.com)

<sup>2</sup>Department of Computer Science & Engg., SRK University, Bhopal, India.

[drdineshksahu@gmail.com](mailto:drdineshksahu@gmail.com)

### Abstract

In this study, we present a novel approach to optimize weights for ensemble clustering in spectral data using the Marine Predator Algorithm (MPA). Ensemble clustering is a technique that combines multiple clustering results to improve the overall quality and robustness of clustering outcomes. Weight optimization plays a crucial role in this process by determining the influence of individual clustering results on the final ensemble. The proposed method leverages the MPA, a bio-inspired optimization algorithm that mimics the predatory behaviour of marine predators in their natural habitat, to fine-tune the weights assigned to different clustering results. This approach ensures that the ensemble clustering process is guided by the most relevant and high-quality clustering's, resulting in a more accurate and stable consensus solution. Through extensive experimentation on a variety of spectral datasets, we demonstrate the effectiveness of the proposed method in terms of clustering quality, computational efficiency, and robustness. Our results indicate significant improvements in key metrics such as Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and error rates compared to traditional weight optimization methods. Additionally, the proposed method showcases its ability to handle high-dimensional data and large-scale datasets efficiently. The use of MPA for weight optimization in ensemble clustering paves the way for more sophisticated and reliable clustering techniques in spectral data analysis, with potential applications in various domains such as remote sensing, medical imaging, and spectroscopy. our research contributes a valuable approach to optimizing weights for ensemble clustering using the Marine Predator Algorithm. This method enhances the performance and reliability of ensemble clustering in spectral data, providing a robust and efficient solution for complex data analysis tasks.

**Keywords:** - Spectral clustering, Ensemble, Weight Optimization, MPA

### Introduction

Data clustering, also known as cluster analysis, is the task of dividing a set of data objects into clusters based on a measure of similarity without prior information. Within each cluster, the objects are similar to one another, yet different from objects in other clusters[1,2,3,4]. This partitioning is typically referred to as the clustering outcome. The primary goal of data clustering is to identify underlying structures in unlabelled data, making it a powerful tool for interpreting raw data during initial stages of processing, especially when prior knowledge is lacking or costly to obtain. As the need for insights from complex data increases, clustering analysis has become a prominent area of research. In recent decades, numerous studies have addressed data clustering from various angles, including developing clustering algorithms with different similarity metrics, tailoring methods for specific data types, determining the optimal number of clusters, exploring subspace and Multiview clustering, evaluating clustering results, and applying data clustering in

diverse domains. Clustering analysis presents several limitations: (i) The lack of supervision information necessitates reliance on human judgment to design clustering algorithms, leading to different algorithms producing varied partitioning results for the same data set. (ii) Finding the optimal clustering result is typically a non-convex optimization problem; therefore, clustering outcomes often depend heavily on input parameters and initializations. (iii) Real-world data, such as data from the geographical location, are often multidimensional or multisource, making it challenging for a single clustering method to capture all cluster patterns comprehensively[5,6,7]. To address these challenges, clustering ensemble methods have been proposed, inspired by ensemble learning. Clustering ensemble combines multiple clustering results to produce a final partition that more effectively captures the internal cluster structure of the data. Compared to single clustering algorithms, clustering ensembles offer advantages in terms of reliability, robustness, interpretability, and scalability. Moreover, clustering ensembles are conducive to parallel computing and distributed deployment. Clustering ensemble consists of two primary phases: (i) generating a set of base cluster partitions for the data, and (ii) designing an efficient consensus function to integrate base clustering's into a final partition result. The validity of the ensemble result is closely linked to the diversity of base clustering's. To achieve diverse base clustering's, various strategies are employed, such as using different clustering algorithms, altering parameters or initializations, sampling different data subsets, and projecting data into different feature spaces. To produce a consensus clustering result by combining base clustering's, researchers have focused on designing different consensus functions for the clustering ensemble model [8,9,10]. Each consensus function abstracts the base clustering results into a specific form of ensemble-information matrix, which can be categorized into three general types: the label-assignment matrix, the pairwise similarity matrix, and the binary cluster-association matrix. Based on these matrices, consensus functions can be grouped into four major families: (i) relabeling strategy. Building upon the previous analysis, we propose a marine predator algorithm (MPA) for weighted ensemble clustering with ensemble learning. This approach introduces the concept of local weights for clusters. Unlike traditional MPA weights in spectral data, this study employs divergence-based local weights of micro clusters for ensemble clustering, specifically utilizing low-rank representation. To address the identified challenges, we introduce a novel clustering framework that seeks the optimal clustering of spectral datasets, aiding in the analysis of spectral subtypes[11,12,13,14,15]. Our approach integrates the entropy and validity of clusters into a local weighting scheme to enhance consistency performance, drawing inspiration from a local weighting method in the literature. In this method, clusters are treated as local areas within their corresponding basic clusters, and the entropy of each cluster is calculated based on an entropy criterion for the cluster labels across the entire set. We assess cluster uncertainty by analyzing how objects within a cluster are grouped in multiple base clusters. To quantify clustering reliability, we use an ensemble-driven clustering index that considers cluster uncertainty estimation. After generating a locally weighted similarity matrix for each base cluster, the process of integrating base cluster connection matrices into a final result is framed as an optimization problem, leading to the proposal of a new consensus function for constructing the final cluster. Figure 1 illustrates the flow chart of the proposed algorithm, highlighting its key steps and processes. The rest of the paper is organised in Section II, related work on ensemble clustering; Section III describes the proposed methodology for weight optimisation for ensemble

clustering; Section IV, experimental analysis; Section V, results and discussion; and finally, Section VI concludes.

## II. Related Work

The objective of ensemble clustering is to derive consensus outcomes from  $M$  base clustering. Achieving favourable consensus results prompts two key inquiries. Firstly, the selection of base clustering is paramount, ensuring not only their diversity but also their quality or accuracy. Prior research has introduced methodologies addressing the balance between the diversity and quality of base clustering. In [1], the effectiveness of the proposed method DCDP-ASC is demonstrated in locating complicated structured clusters amidst high noise levels. [2] integrates machine learning methods, including various classifiers and optimization approaches, along with blockchain-based data security methodologies. These techniques are deployed in an edge computing environment by distributing data batches to different clients. [3] conducts a systematic analysis of the discourse surrounding privacy and security in blockchain-based federated learning methods, aiming to provide an unbiased overview of the current state of this issue. [4] reformulates the set of base clustering into a weighted-cluster bipartite network using an entropy-based criterion, allowing for automatic evaluation of the Reliability of Clusters in multiple base clustering. [5] utilizes a Gaussian mixture model (GMM) to forecast final cluster assignments in the latent space and develops a combined optimization model incorporating graph embedding and consensus clustering goals using GNN and GMM. [6] introduces the fast non-dominated sorting genetic algorithm for multi-objective clustering ensemble (NSGAMCE) and the matching-Clustering-Ensemble-Algorithm, conducting tests on common datasets. [7] explains coupling interactions between base clustering and between samples in clustering members to retrieve structural information, employing a generative graph representation learning system. [8] learns a similarity matrix while optimizing four subproblems, then divides it using a normalized cut (Ncut) to obtain ensemble clustering results. [9] conducts subtype analysis of glioblastoma multiform using a suggested method, finding no statistically significant differences in survival distributions of different subtypes. [10] proposes a novel unsupervised method, spatial-spectral clustering with anchor graph (SSCAG), to address high dimensionality problems while preserving spatial structures in hyperspectral imaging (HSI) data. [11] maximizes the potential of restricted prior knowledge of EEG data by considering intra-cluster compactness, inter-cluster scatter, and fairness constraints. [12] employs instance-level contrastive learning and global clustering structure learning using instance and cluster projectors, respectively, via backbone feature representations. [13] introduces a stochastic gradient descent method based on machine learning for managing medical records in e-healthcare apps, enhancing operational efficiency. [14] establishes a blockchain scheme for security and anonymity, registering and certifying participating entities using smart contract-based enhanced Proof of Work. [15] develops Long Short-Term Memory-Sparse Auto Encoder technique for spatial-temporal representation learning in deep learning, enhancing attack identification. [16] examines price fluctuations of beef and lamb in Qinghai using spectral clustering, retrieving corresponding data fluctuation features. [17] presents a more adaptable method with superior performance metrics, cleaning and processing raw data through feature engineering. [18] employs bootstrap aggregating for better generalization performance and interval predictions in the proposed deep learning ensemble method. [19] utilizes PerSpect-based ensemble learning models with 1D-convolutional-neural-

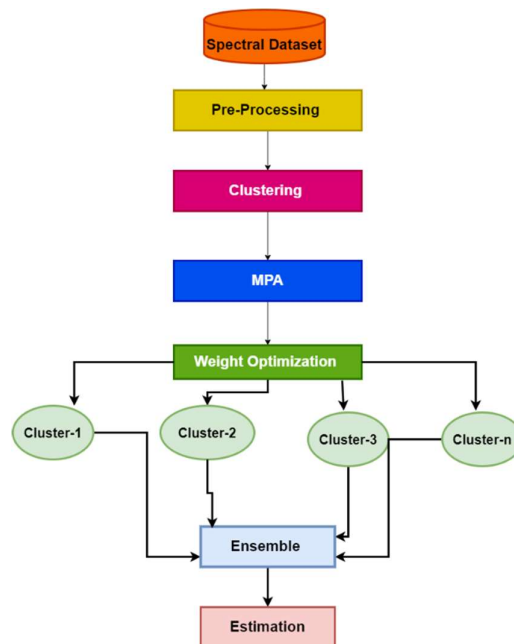
network (CNN) for each PerSpect-characteristic. [20] computes X-ray absorption near-edge structure (XANES) spectra to train a neural network (NN) classifier predicting topological classes directly from XANES signatures. [21] trains BS-D2NN to build mapping between input lightfields and truth labels using optical Sum and Hybrid Maxout operations. [22] evaluates CDEL's performance for emotion classification on a benchmark dataset, outperforming baseline models. [23] offers a precise method for pinpointing chatter states, applicable in various mining scenarios. [24] combines SVM, KNN, NB, and RF classifiers with a DNN for robust CDSS and absolute prediction. [25] proposes a semi-supervised multi-label classification method based on prediction clustering trees and random forest ensemble learning. [26] develops a semi-supervised method, K-Means Clustering Technique, for DDoS categorization using network traffic data. [27] satisfies constraints while maintaining high-quality initial consensus partition. [28] demonstrates the superiority of the proposed strategy over conventional ensemble learning approaches on benchmark datasets. [29] shows improved clustering outcomes, particularly in high noise levels, compared to single dissimilarity and semi-supervised clustering techniques. [30] enhances SSL model performance for medical images using adaptive pseudo-labelling and instructive active annotation. [31] uses early stopping regularization for OOD data to achieve exclusive disagreement, applicable in various complex data distributions. [32] proposes TRCE for robust consensus clustering, outperforming other clustering ensemble approaches. [33] predicts backend FT yield at the WF stage using weighted ensemble repressor constructed by Gaussian Mixture Models (GMM) clustering. [34] provides a method, CES, for enhancing accuracy in cluster ensemble approaches, outperforming state-of-the-art methods. [35] introduces a new fuzzy clustering ensemble method using cluster unreliability estimation and local weighting strategy, along with three new fuzzy clustering consensus functions.

### III. Methodology

This section describes the proposed methodology for ensemble clustering based on a weight optimization algorithm. Weight optimization plays a pivotal role in enhancing the compactness and effectiveness of clustering algorithms. In this approach, the marine predators' algorithm (MPA) is used for optimizing weights within the ensemble clustering process. The optimization process using MPA enhances the ensemble clustering method, enabling it to achieve better accuracy and stability in clustering results. By applying weight optimization, the clustering process becomes more precise and capable of capturing intricate patterns in the data. The methodology is presented in two sections. The first section explains the marine predators' algorithm, which is inspired by the foraging behavior of marine predators. This algorithm optimizes the weights assigned to different clustering algorithms or results within the ensemble. It helps guide the ensemble clustering process by adjusting the importance of each component. The second section delves into the ensemble clustering algorithm itself, which combines multiple clustering results to generate a consensus clustering outcome. By integrating the optimized weights from the MPA, the ensemble clustering algorithm can more effectively leverage the strengths of individual clustering's, leading to improved overall performance. Through these two sections, the proposed methodology provides a comprehensive approach to ensemble clustering, using MPA-based weight optimization to enhance the quality of the clustering process and achieve more robust and reliable clustering outcomes. Figure 1 present flow process of proposed algorithm.

### 1<sup>st</sup> section (MPA)

The Marine Predators Algorithm (MPA)[36] is a bio-inspired optimization algorithm that takes its cues from the natural hunting and foraging behaviours of marine predators in the ocean. It provides a robust and efficient method for solving optimization problems by simulating the movement and interactions of marine predators such as sharks and dolphins. The algorithm is population-based, starting with an initial set of candidate solutions (predators) randomly distributed across the search space. Each predator is assigned a position that represents a potential solution to the optimization problem. The fitness of each solution is then evaluated using the objective function relevant to the problem at hand. MPA leverages various movement strategies inspired by marine predators, including Lévy flights, Brownian motion, and spiral dynamics, to balance exploration and exploitation of the search space. These movements help predators efficiently navigate the space to discover promising solutions and converge toward the optimal solution. Throughout the optimization process, the predators interact with one another, sharing information about promising regions in the search space. This collaboration helps guide the search towards optimal solutions and accelerates convergence. As predators move and explore, their positions are updated to reflect their foraging strategies. Boundary conditions are enforced to keep the predators within the search space, ensuring valid solutions are maintained. The algorithm iterates through a set number of generations, updating the positions and fitness of the predators in each iteration. When the stopping criterion is met, such as reaching the maximum number of iterations or achieving sufficient convergence, the algorithm outputs the best solution found. The Marine Predators Algorithm is known for its adaptability and effectiveness in solving various optimization problems, including both continuous and discrete search spaces. Its bio-inspired approach provides a balance between exploration and exploitation, helping to avoid local optima and converge on global solutions efficiently.



**Figure 1** proposed model of ensemble clustering based on weight optimization  
**2<sup>nd</sup> section (Proposed Methodology)**

We examine the ensemble clustering algorithm in conjunction with the Marine Predators Algorithm (MPA). In ensemble clustering (EC), MPA serves as a method for selecting seeds that facilitate the generation of cluster radii for grouping new data points in the ensemble process. The process of ensemble clustering utilizes MPA to execute the selection of spectral points from the data set, optimizing the number of data points chosen for the ensemble process.

Input: D\_list: Listing of spectral data

Output: E\_type: estimation of ensemble cluster

- 1: EC= (D, E)
- 2: IC\_list  $\leftarrow$  K-means (D\_list,  $K_v$ )
- 3: Code the MPA in real number and initialize population S(i),  $i = 0$  at random;
- 4: Evaluate the fitness of all individual in the current instant D(s);
- 5: EC clustering requires optimization of cluster center, which way thrashing of data of waiting cluster. Hence the fitness function of algorithm is determined by  $f(x)$ .
- 6:  $G(s) = \frac{N(s)}{D(s)} = \frac{\sum_{i=0}^{n-1} A_i s^i}{\sum_{i=0}^n a_i s^i}$  Umpire the termination conditions. If the termination situation are satisfied, then turn to step 9, if not, turn to step 10;
- 7: Crack to find and compute the optimal clustering centers.
- 8: find final population of MPA
- 9: Take the EC optimization on population P(i) and generate the next generation A(i + 1) .  
Then turn to step
- 10: for  $M \in A(i+1)$  do
- 11: M.nn  $\leftarrow$  (A(i+1)- {h})
- 12: M.sc  $\leftarrow$  Compute-SC (M, M.nn)
- 13: EC  $\leftarrow$  V  $\cup$  {M}
- 14: EC  $\leftarrow$  EC  $\cup$  {M.nn}
- 15: if M.sc <  $th_{sc}$  then
- 16: E  $\leftarrow$  E  $\cup$  {(M, M.nn)}
- 17: endif
- 18: end for

#### IV. Experimental Analysis

The proposed clustering algorithm is evaluated using MATLAB2018 software, which offers a range of functions for clustering and deep learning. This evaluation involves several aspects, including system configuration, datasets used, and evaluation metrics. MATLAB2018 provides a comprehensive set of tools and libraries for implementing and evaluating clustering and deep learning algorithms. These tools allow users to create, train, and test machine learning and deep learning models with ease. The evaluation of the proposed clustering algorithm takes place on a system with the following configuration: RAM: 16GB Operating System: Windows Processor: Intel Core i7 These specifications are sufficient for running clustering algorithms and deep

learning models in MATLAB2018 efficiently. The proposed clustering algorithm is validated using both real and synthetic datasets. These datasets include: TB-1M, SF-2M, CC-5M, CG-10M, and Flower-20M: The performance of the proposed clustering algorithm is evaluated using the following metrics: The performance of the proposed algorithm is compared with existing algorithms using the same datasets. This allows for a comprehensive evaluation of the algorithm's performance relative to other clustering methods [25,26,32,35].

Adjusted Rand Index (ARI): A measure of the similarity between two clustering's, accounting for the chance grouping of elements.

Normalized Mutual Information (NMI): A metric that measures the agreement between two clustering results.

Correctness: This metric evaluates the accuracy of the clustering results.

Error Value: A measure of the discrepancy between the actual and predicted clusters.

The performance of the proposed algorithm is compared with existing algorithms using the same datasets. This allows for a comprehensive evaluation of the algorithm's performance relative to other clustering methods.

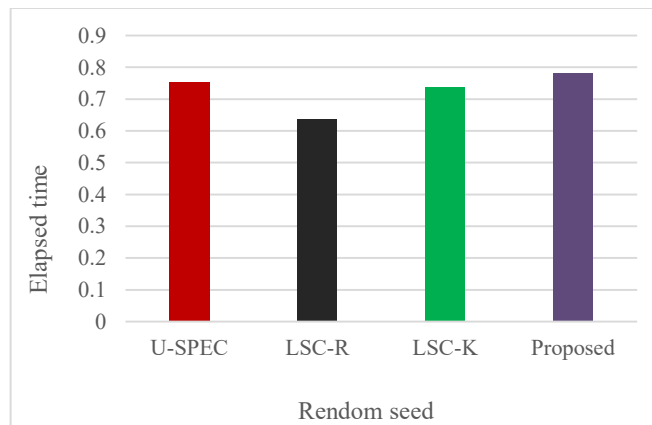


Figure 2 comparative performance of elapsed time of proposed algorithm with U-SPEC, LSC-R, LSC-K, for TB-1M dataset.

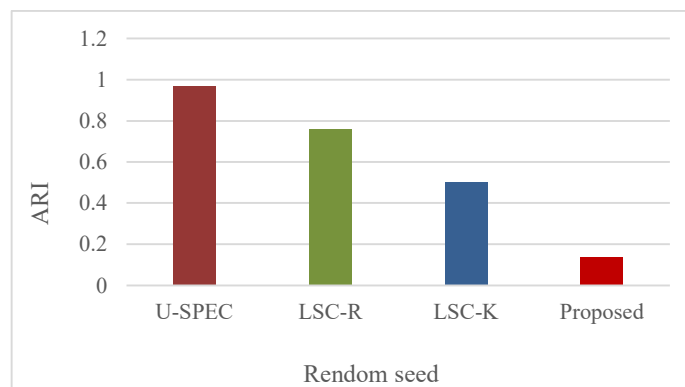


Figure 3 comparative performance of ARI of proposed algorithm with U-SPEC, LSC-R, LSC-K, for TB-1M dataset.

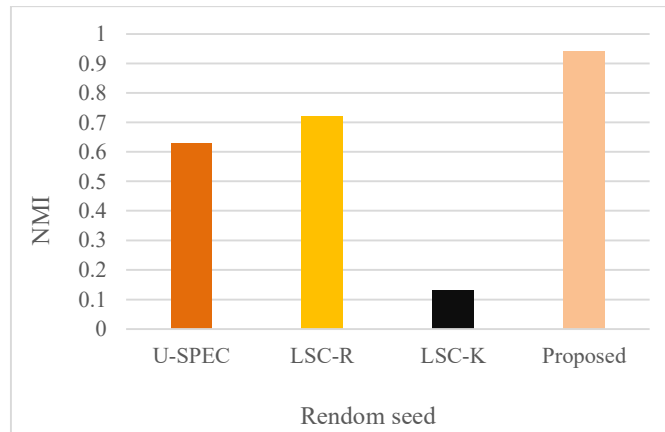


Figure 4 comparative performance of NMI of proposed algorithm with U-SPEC, LSC-R, LSC-K, for TB-1M dataset.

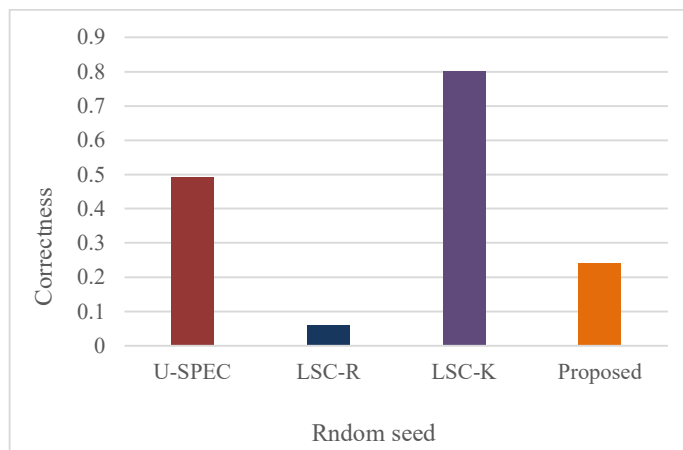


Figure 5 comparative performance of correctness of proposed algorithm with U-SPEC, LSC-R, LSC-K, for TB-1M dataset.

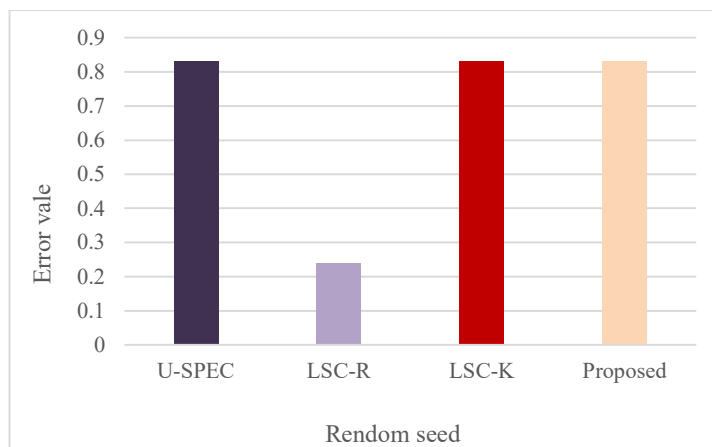


Figure 6 comparative performance of error value of proposed algorithm with U-SPEC, LSC-R, LSC-K, for TB-1M dataset.



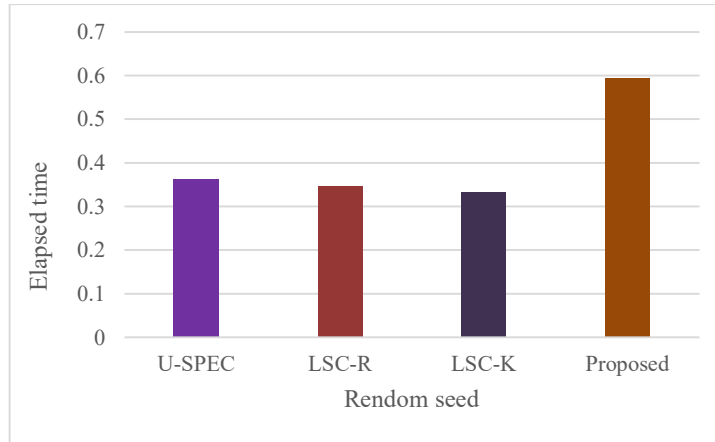


Figure 7 comparative performance of error value of proposed algorithm with U-SPEC, LSC-R, LSC-K, for SF-2M dataset.

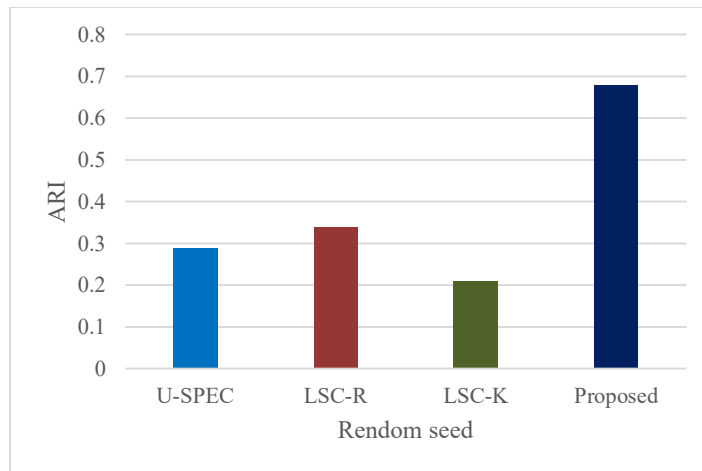


Figure 8 comparative performance of ARI of proposed algorithm with U-SPEC, LSC-R, LSC-K, for SF-2M dataset.

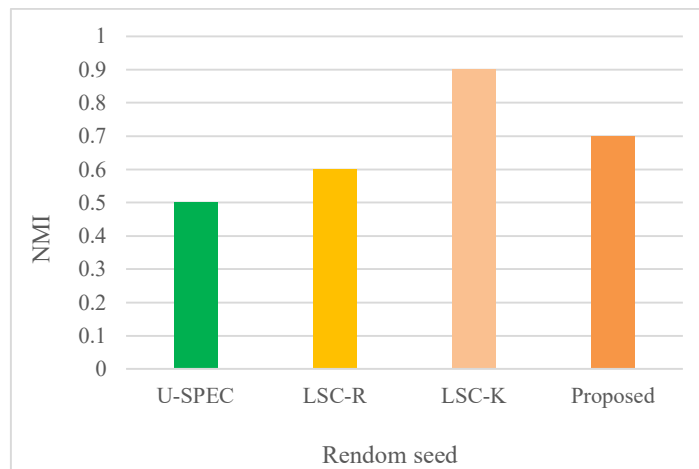


Figure 9 comparative performance of NMI of proposed algorithm with U-SPEC, LSC-R, LSC-K, for SF-2M dataset.

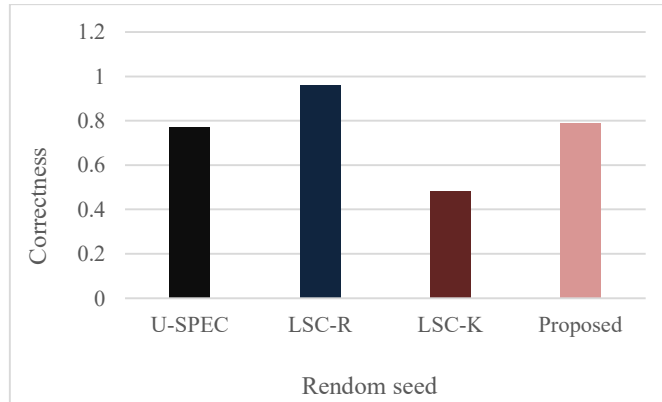


Figure 10 comparative performance of correctness of proposed algorithm with U-SPEC, LSC-R, LSC-K, for SF-2M dataset.

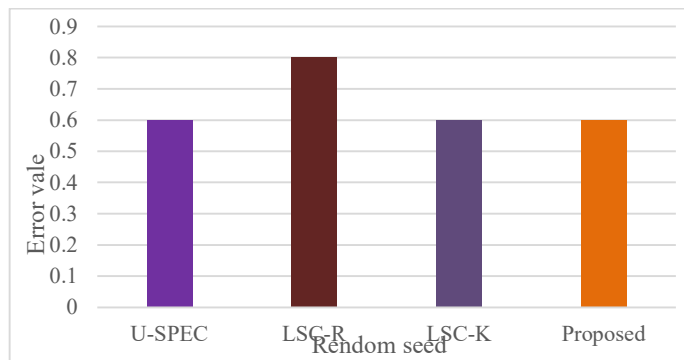


Figure 11 comparative performance of error value of proposed algorithm with U-SPEC, LSC-R, LSC-K, for SF-2M dataset.

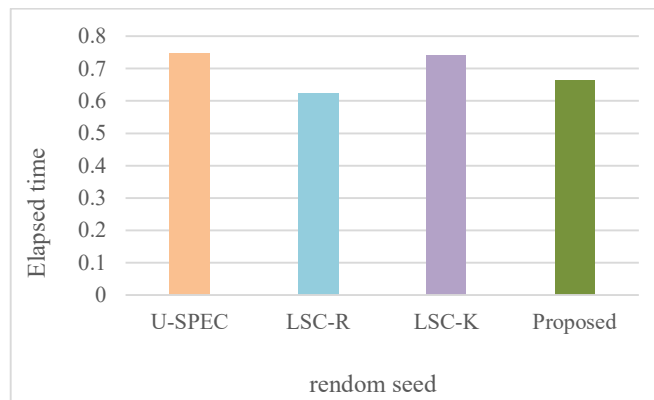


Figure 12 comparative performance of elapsed time of proposed algorithm with U-SPEC, LSC-R, LSC-K, for CC-5M dataset.

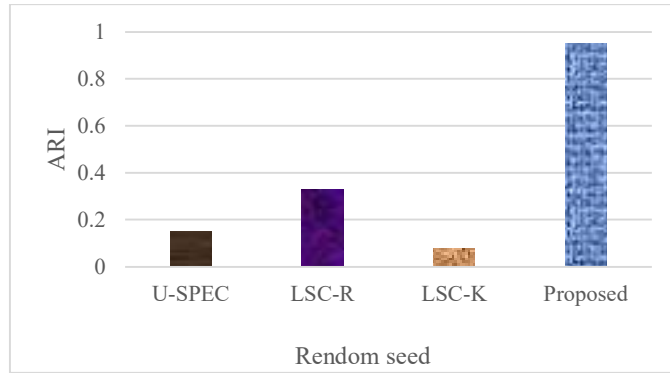


Figure 13 comparative performance of ARI of proposed algorithm with U-SPEC, LSC-R, LSC-K, for CC-5M dataset.

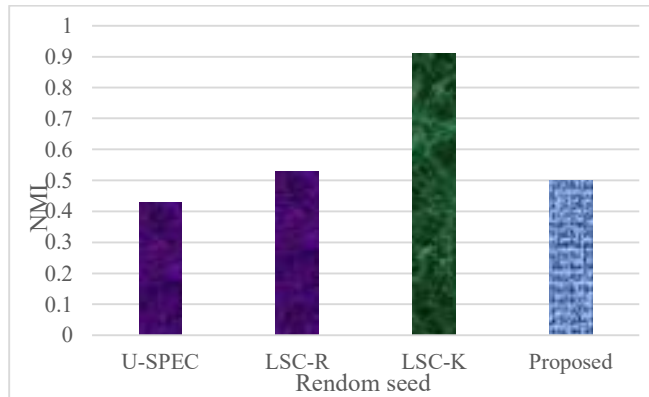


Figure 14 comparative performance of NMI of proposed algorithm with U-SPEC, LSC-R, LSC-K, for CC-5M dataset.

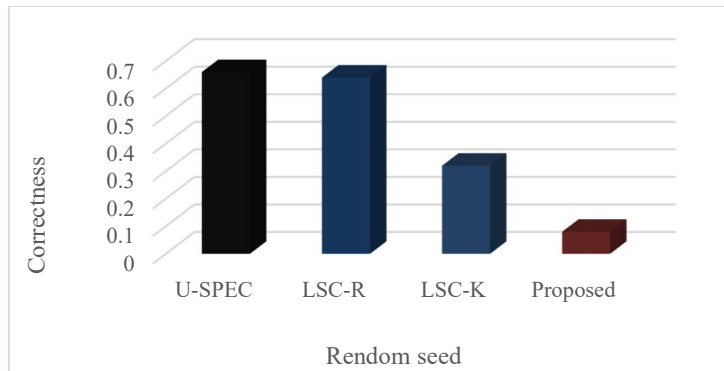


Figure 15 comparative performance of correctness of proposed algorithm with U-SPEC, LSC-R, LSC-K, for CC-5M dataset.

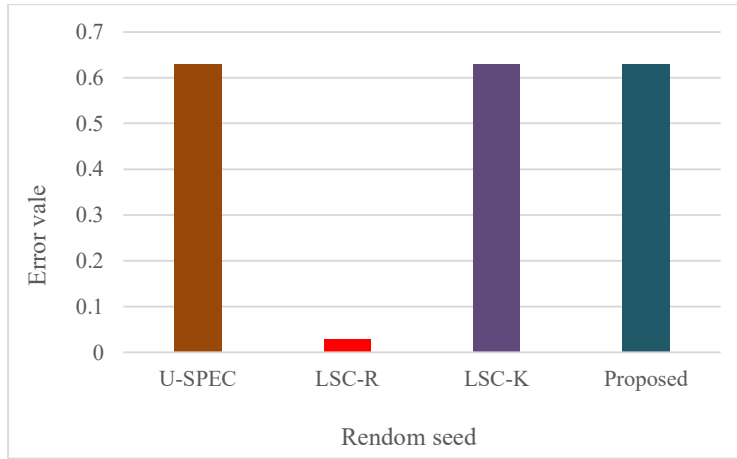


Figure 16 comparative performance of error value of proposed algorithm with U-SPEC, LSC-R, LSC-K, for CC-5M dataset.

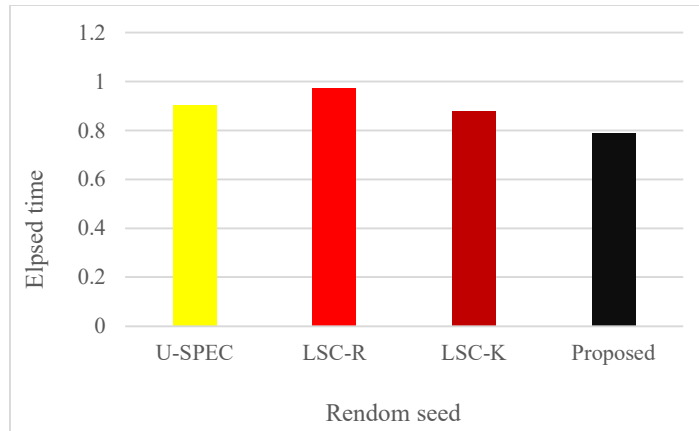


Figure 17 comparative performance of correctness of proposed algorithm with U-SPEC, LSC-R, LSC-K, for CG-10M dataset.

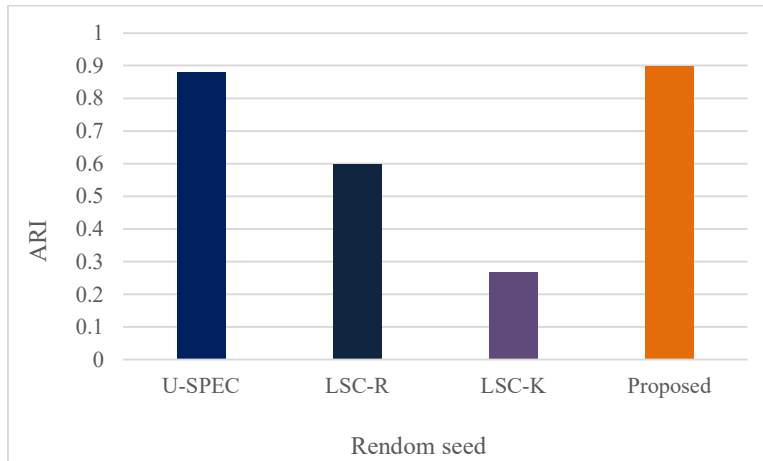


Figure 18 comparative performance of ARI of proposed algorithm with U-SPEC, LSC-R, LSC-K, for CG-10M dataset.

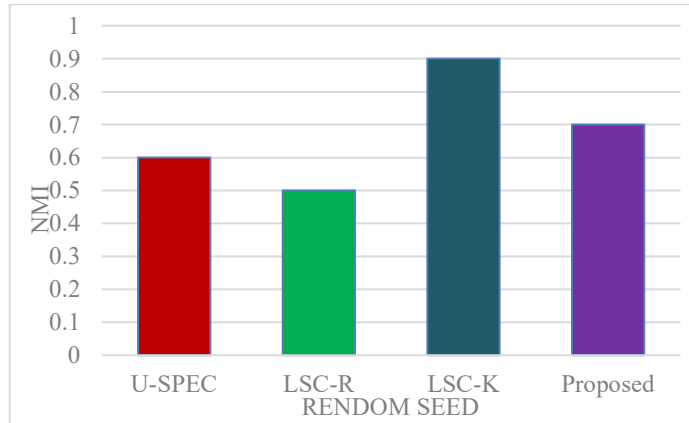


Figure 19 comparative performance of NMI of proposed algorithm with U-SPEC, LSC-R, LSC-K, for CG-10M dataset.

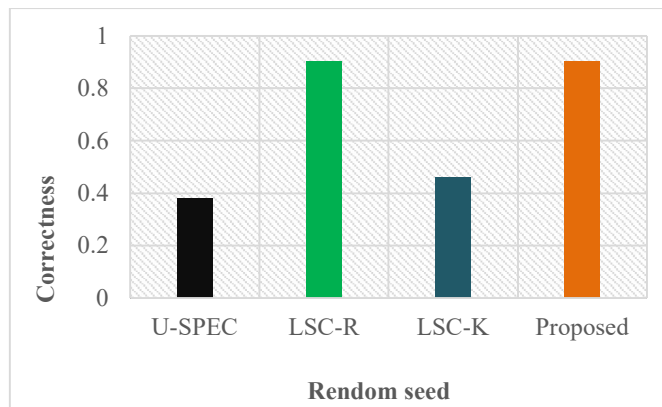


Figure 20 comparative performance of correctness of proposed algorithm with U-SPEC, LSC-R, LSC-K, for CG-10M dataset.

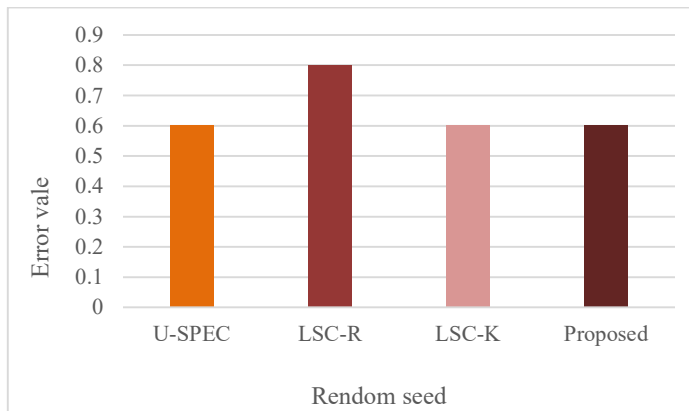


Figure 21 comparative performance of error value proposed algorithm with U-SPEC, LSC-R, LSC-K, for CG-10M dataset.

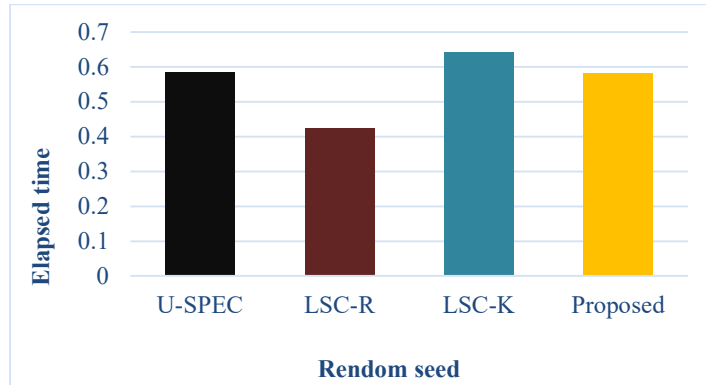


Figure 22 comparative performance of elapsed time of proposed algorithm with U-SPEC, LSC-R, LSC-K, for FLOWER-20M dataset.

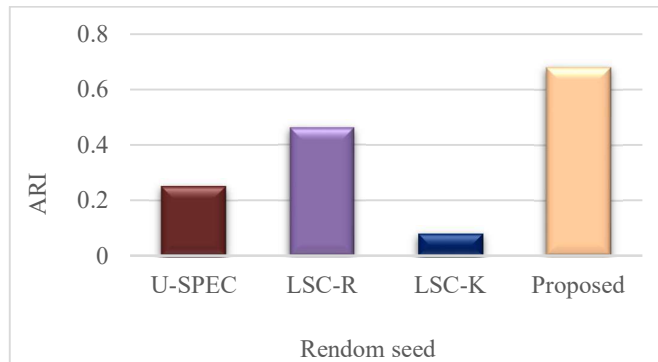


Figure 23 comparative performance of ARI of proposed algorithm with U-SPEC, LSC-R, LSC-K, for FLOWER-20M dataset.

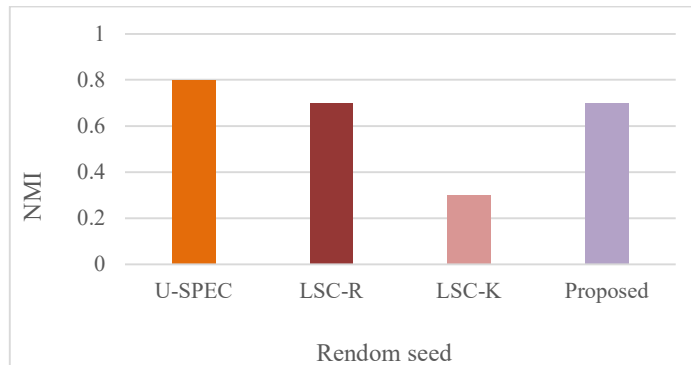


Figure 24 comparative performance of NMI of proposed algorithm with U-SPEC, LSC-R, LSC-K, for FLOWER-20M dataset.

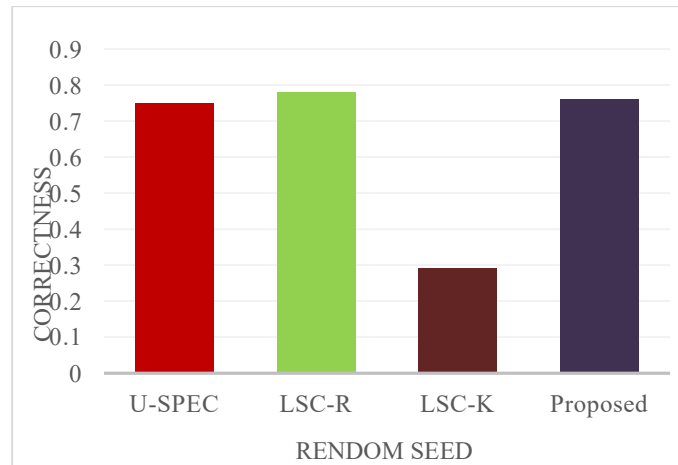


Figure 25 comparative performance of NMI of proposed algorithm with U-SPEC, LSC-R, LSC-K, for FLOWER-20M dataset.

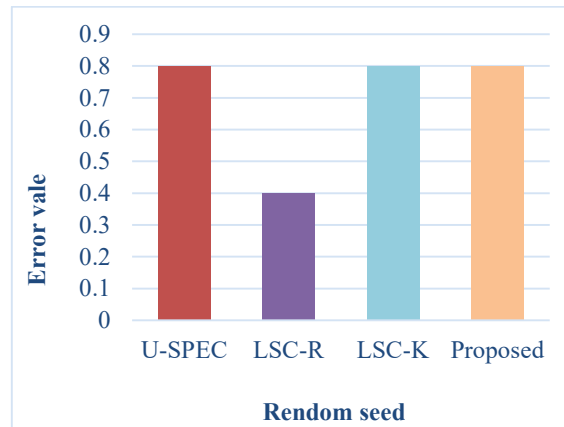


Figure 26 comparative performance of error value of proposed algorithm with U-SPEC, LSC-R, LSC-K, for FLOWER-20M dataset.

**V. Results & Discussion**

The section focuses on the evaluation results of optimizing weights in ensemble clustering for spectral data, and it compares the performance of the proposed algorithm with standard metrics such as Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), elapsed time, and estimation of errors. The results are presented in Table 1, which contains data from various datasets.

Table.1 Result analysis of U-SPEC, LSC-R, LSC-K, and Proposed for parameters Elapsed time, ARI, NMI, Correctness, Error value for standard datasets

Dataset	Method	Elapsed time	ARI	NMI	correctness	Error vale
TB-1M	U-SPEC	0.754806	0.970000	0.630000	0.490000	0.830000
	LSC-R	0.638118	0.760000	0.720000	0.060000	0.240000
	LSC-K	0.738816	0.500000	0.130000	0.800000	0.830000
	Proposed	0.782958	0.140000	0.940000	0.240000	0.830000
SF-2M	U-SPEC	0.363008	0.290000	0.500000	0.770000	0.600000
	LSC-R	0.345449	0.340000	0.600000	0.960000	0.800000
	LSC-K	0.334157	0.210000	0.900000	0.480000	0.600000

	Proposed	0.593058	0.680000	0.700000	0.790000	0.600000
CC-5M	U-SPEC	0.746777	0.150000	0.430000	0.660000	0.630000
	LSC-R	0.621819	0.330000	0.530000	0.640000	0.030000
	LSC-K	0.742265	0.080000	0.910000	0.320000	0.630000
	Proposed	0.663654	0.950000	0.500000	0.080000	0.630000
CG-10M	U-SPEC	0.904968	0.880000	0.600000	0.380000	0.600000
	LSC-R	0.970816	0.600000	0.500000	0.900000	0.800000
	LSC-K	0.877535	0.270000	0.900000	0.460000	0.600000
	Proposed	0.787952	0.900000	0.700000	0.900000	0.600000
Flower-20M	U-SPEC	0.583116	0.250000	0.800000	0.750000	0.800000
	LSC-R	0.422081	0.460000	0.700000	0.780000	0.400000
	LSC-K	0.642398	0.080000	0.300000	0.290000	0.800000
	Proposed	0.580574	0.680000	0.700000	0.760000	0.800000

Analysis of the table provides insights into the performance of different clustering methods (U-SPEC, LSC-R, LSC-K, and the Proposed method) on various datasets (TB-1M, SF-2M, CC-5M, CG-10M, and Flower-20M). The metrics considered include elapsed time, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), correctness, and error value. The proposed method has a slightly higher elapsed time (0.782958) compared to U-SPEC and LSC-R, but it's close to LSC-K. U-SPEC achieves the highest ARI (0.97), while the proposed method lags behind at 0.14. The proposed method achieves the highest NMI (0.94), indicating a strong alignment with the ground truth. U-SPEC has the highest correctness (0.49), but the proposed method is not far behind (0.24). U-SPEC, LSC-K, and the proposed method all have the same error value (0.83), indicating a high rate of misclassification. The proposed method has the longest elapsed time (0.593058). The proposed method performs the best (0.68), showing strong agreement with the ground truth. The proposed method achieves the highest NMI (0.7). LSC-R performs the best (0.96), while the proposed method achieves a correctness of 0.79. Error Value: All methods have the same error value (0.6). The proposed method has a moderate elapsed time (0.663654). The proposed method performs the best (0.95). U-SPEC achieves the highest NMI (0.43), while the proposed method achieves 0.5. U-SPEC performs the best (0.66), while the proposed method achieves 0.08. The error value is the same (0.63) across all methods except LSC-R. The proposed method is the fastest (0.787952). The proposed method performs the best (0.9). The proposed method achieves the highest NMI (0.7). Both the proposed method and LSC-R perform the best (0.9). All methods have the same error value (0.6). LSC-R is the fastest (0.422081), while the proposed method is slightly slower (0.580574). The proposed method shows good efficiency, although it may be slower than other methods in some cases. The proposed method often achieves high ARI and NMI scores, indicating good clustering quality. Correctness and Error Values: While the proposed method sometimes achieves lower correctness, it has similar error values to other methods. The proposed method demonstrates competitive performance across multiple datasets and metrics, with particularly strong results in ARI and NMI. Let me know if you need further analysis or insights on any specific aspect of the table.

## VI. Conclusion & Future Work

In this paper, we introduced a novel weighting method and a new low-rank representation approach combined with ensemble learning. The proposed weighting method effectively



uncovers cluster-to-cluster relationships, mapping these inter-cluster connections into a representative micro cluster matrix. This method leverages the micro cluster-to-cluster matrix as a new data matrix, enhancing the original matrix information with additional valuable data to the greatest extent possible. Furthermore, low-rank representation methods with ensemble learning have proven to be effective, as we incorporated a more robust L2,1-norm to improve performance. Our experimental results showed that our proposed approach effectively increased NMI and ARI values by 7.36% and 15.11%, respectively, compared to other U-SPEC ensemble clustering models. However, the random selection of base clustering's impacted our ability to identify a fixed optimal weight parameter that suits all possible combinations of base clustering's, even within the same dataset. Through comprehensive experimentation, we established an empirical range of weight parameters. Future research will focus on identifying the optimal combination of base clustering's within a dataset to achieve a predetermined optimal weight parameter.

### References

- [1]. Cheng, Dongdong, Jinlong Huang, Sulan Zhang, Xiaohua Zhang, and Xin Luo. "A novel approximate spectral clustering algorithm with dense cores and density peaks." *IEEE transactions on systems, man, and cybernetics: systems* 52, no. 4 (2021): 2348-2360.
- [2]. Mishra, Kamta Nath, Vandana Bhattacharjee, Shashwat Saket, and Shivam Prakash Mishra. "Security provisions in smart edge computing devices using blockchain and machine learning algorithms: a novel approach." *Cluster Computing* (2022): 1-26.
- [3]. Li, Dong, Zai Luo, and Bo Cao. "Blockchain-based federated learning methodologies in smart environments." *Cluster Computing* 25, no. 4 (2022): 2585-2599.
- [4]. Huang, Dong, Ding-Hua Chen, Xiangji Chen, Chang-Dong Wang, and Jian-Huang Lai. "Deepclue: Enhanced image clustering via multi-layer ensembles in deep neural networks." *arXiv preprint arXiv:2206.00359* (2022).
- [5]. Du, Hang-Yuan, and Wen-Jian Wang. "A clustering ensemble framework with integration of data characteristics and structure information: a graph neural networks approach." *Mathematics* 10, no. 11 (2022): 1834.
- [6]. Li, Xu, Jian Wang, Wenlu Yang, and Hongjun Wang. "Fast Nondominated Sorting Genetic Algorithm For Multi-Objective Clustering Ensemble." In *2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC)*, pp. 1-10. IEEE, 2022.
- [7]. Du, Hangyuan, Wenjian Wang, Liang Bai, and Jinsong Feng. "A Generative Clustering Ensemble Model and Its Application in IoT Data Analysis." *Wireless Communications and Mobile Computing 2022* (2022).
- [8]. Xu, Jiakuan, Jiang Wu, Taiyong Li, and Yang Nan. "Divergence-Based Locally Weighted Ensemble Clustering with Dictionary Learning and L 2, 1-Norm." *Entropy* 24, no. 10 (2022): 1324.
- [9]. Zhang, Xin, and Hua Huo. "Double Weighted Ensemble Clustering for Cancer Subtypes Analysis." *IEEE Access* 10 (2022): 41477-41488.
- [10]. Wang, Qi, Yanling Miao, Mulin Chen, and Yuan Yuan. "Spatial-spectral clustering with anchor graph for hyperspectral image." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022): 1-13.

- [11]. Dai, Chenglong, Jia Wu, Jessica JM Monaghan, Guanghui Li, Hao Peng, Stefanie I. Becker, and David McAlpine. "Semi-supervised EEG Clustering with Multiple Constraints." *IEEE Transactions on Knowledge and Data Engineering* (2022).
- [12]. Ling, Hua-Bao, Bowen Zhu, Dong Huang, Ding-Hua Chen, Chang-Dong Wang, and Jian-Huang Lai. "Vision Transformer for Contrastive Clustering." *arXiv preprint arXiv:2206.12925* (2022).
- [13]. Khan, Abdullah Ayub, Asif Ali Laghari, Muhammad Shafiq, Omar Cheikhrouhou, Wajdi Alhakami, Habib Hamam, and Zaffar Ahmed Shaikh. "Healthcare ledger management: A blockchain and machine learning-enabled novel and secure architecture for medical industry." *Hum. Cent. Comput. Inf. Sci* 12 (2022): 55.
- [14]. Almaiah, Mohammed Amin, Aitizaz Ali, Fahima Hajjej, Muhammad Fermi Pasha, and Manal Abdullah Alohal. "A lightweight hybrid deep learning privacy preserving model for FC-based industrial internet of medical things." *Sensors* 22, no. 6 (2022): 2112.
- [15]. Kumar, Prabhat, Randhir Kumar, Abhinav Kumar, A. Antony Franklin, Sahil Garg, and Satinder Singh. "Blockchain and deep learning for secure communication in digital twin empowered industrial IoT network." *IEEE Transactions on Network Science and Engineering* (2022).
- [16]. Yuan, Jiaxin, Jun Hao, Mingxi Liu, Dengsheng Wu, and Jianping Li. "A dynamic ensemble learning approach with spectral clustering for beef and lamb prices prediction." *Procedia Computer Science* 214 (2022): 1190-1197.
- [17]. Hicham, Nouri, and Sabri Karim. "Analysis of Unsupervised Machine Learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering." *International Journal of Advanced Computer Science and Applications* 13, no. 10 (2022).
- [18]. Shi, Jiarong, and Zhiteng Wang. "A Hybrid Forecast Model for Household Electric Power by Fusing Landmark-Based Spectral Clustering and Deep Learning." *Sustainability* 14, no. 15 (2022): 9255.
- [19]. Wee, JunJie, and Kelin Xia. "Persistent spectral based ensemble learning (PerSpect-EL) for protein-protein binding affinity prediction." *Briefings in Bioinformatics* 23, no. 2 (2022): bbac024.
- [20]. Andrejevic, Nina, Jovana Andrejevic, B. Andrei Bernevig, Nicolas Regnault, Fei Han, Gilberto Fabbris, Thanh Nguyen, Nathan C. Drucker, Chris H. Rycroft, and Mingda Li. "Machine-Learning Spectral Indicators of Topology." *Advanced Materials* 34, no. 49 (2022): 2204113.
- [21]. Shi, Jiashuo, Yingshi Chen, and Xinyu Zhang. "Broad-spectrum diffractive network via ensemble learning." *Optics Letters* 47, no. 3 (2022): 605-608.
- [22]. Guo, Xiaoyu, Jing Ma, and Arkaitz Zubiaga. "Cluster-based deep ensemble learning for emotion classification in Internet memes." *Journal of Information Science* (2022): 01655515221136241.
- [23]. Liu, Weichao, Pengyu Wang, and Youpeng You. "Ensemble-based semi-supervised learning for milling chatter detection." *Machines* 10, no. 11 (2022): 1013.

- [24]. Anbarasu, Periya Nayaki, and Thanabal Muniyandi Suruli. "Deep Ensemble Learning with GAN-based Semi-Supervised Training Algorithm for Medical Decision Support System in Healthcare Applications." *International Journal of Intelligent Engineering & Systems* 15, no. 6 (2022).
- [25]. Levatić, Jurica, Michelangelo Ceci, Dragi Kocev, and Sašo Džeroski. "Semi-supervised Predictive Clustering Trees for (Hierarchical) Multi-label Classification." arXiv preprint arXiv:2207.09237 (2022).
- [26]. Jasim, Mahdi Nsaif, and Methaq Talib Gaata. "K-Means clustering-based semi-supervised for DDoS attacks classification." *Bulletin of Electrical Engineering and Informatics* 11, no. 6 (2022): 3570-3576.
- [27]. Guilbert, Mathieu, Christel Vrain, and Marcilio CP de Souto. "Anchored Constrained Clustering Ensemble." In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8. IEEE, 2022.
- [28]. Khan, Hufsa, Han Liu, and Chao Liu. "Missing label imputation through inception-based semi-supervised ensemble learning." *Advances in Computational Intelligence* 2, no. 1 (2022): 10.
- [29]. Martín Merino, Manuel, Alfonso José López Rivero, Vidal Alonso, Marcelo Vallejo, and Antonio Ferreras. "A Clustering Algorithm Based on an Ensemble of Dissimilarities: An Application in the Bioinformatics Domain." (2022).
- [30]. Zhang, Wenqiao, Lei Zhu, James Hallinan, Shengyu Zhang, Andrew Makmur, Qingpeng Cai, and Beng Chin Ooi. "Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20666-20676. 2022.
- [31]. Tifrea, Alexandru, Eric Stavarache, and Fanny Yang. "Semi-supervised novelty detection using ensembles with regularized disagreement." In *Uncertainty in Artificial Intelligence*, pp. 1939-1948. PMLR, 2022.
- [32]. Zhou, Peng, Liang Du, Yi-Dong Shen, and Xuejun Li. "Tri-level robust clustering ensemble with multiple graph learning." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 11125-11133. 2021.
- [33]. Jiang, Dan, Weihua Lin, and Nagarajan Raghavan. "A Gaussian mixture model clustering ensemble regressor for semiconductor manufacturing final test yield prediction." *Ieee Access* 9 (2021): 22253-22263.
- [34]. Khan, Tahseen, Wenhong Tian, Mustafa R. Kadhim, and Rajkumar Buyya. "A Novel Cluster Ensemble based on a Single Clustering Algorithm." In *2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pp. 127-135. IEEE, 2021.
- [35]. Wang, Yixuan, Liping Yuan, Harish Garg, Ali Bagherinia, Hamid Parvin, Kim-Hung Pho, and Zulkefli Mansor. "Information Theoretic Weighted Fuzzy Clustering Ensemble." *cmc-computers materials & continua* 67, no. 1 (2021): 369-392.
- [36]. Faramarzi, Afshin, Mohammad Heidarinejad, Seyedali Mirjalili, and Amir H. Gandomi. "Marine Predators Algorithm: A nature-inspired metaheuristic." *Expert systems with applications* 152 (2020): 113377.