SRYAHWA PUBLICATIONS

Research Journal of Nanoscience and Engineering
ISSN: 2637-5591 | Volume 7, Issue 2, 2025
https://doi.org/10.22259/2637-5591.0702001

RESEARCH ARTICLE

Enhanced Fuzzy C-Means Clustering for Improved Data Analysis with Application to the Higgs Boson Dataset

Naga Charan Nandigama

Received: 21 October 2025 Accepted: 06 November 2025 Published: 11 November 2025
Corresponding Author: Naga Charan Nandigama. nagacharan.nandigama@gmail.com.

Abstract

Fuzzy C-Means (FCM) clustering is a widely used algorithm for data partitioning, particularly when clusters exhibit overlap. However, its performance can be significantly impacted by sensitivity to initial conditions and susceptibility to noise. This paper proposes an Enhanced Fuzzy C-Means (EFCM) algorithm designed to mitigate these limitations. EFCM integrates a density-based initialization strategy using kernel density estimation, employs the robust Mahalanobis distance metric to handle noise and outliers, and incorporates Silhouette index-based adaptive parameter selection. The algorithm's effectiveness is evaluated on the challenging Higgs Boson dataset from Kaggle, a high-dimensional and noisy dataset commonly used in high-energy physics research. Results demonstrate EFCM's superior performance compared to traditional FCM in terms of clustering accuracy, execution time, and precision.

Keywords: Enhanced Fuzzy C-Means, Clustering, Fuzzy Clustering, Initialization, Noise Handling, Parameter Selection, Silhouette Index, Mahalanobis Distance, Higgs Boson Dataset, Kernel Density Estimation.

1. Introduction

Clustering, a fundamental technique in unsupervised learning, plays a crucial role in knowledge discovery and pattern recognition. Fuzzy C-Means (FCM) is a prominent fuzzy clustering algorithm that allows data points to belong to multiple clusters with varying degrees of membership. While FCM is effective in handling overlapping clusters, its performance is often compromised by its sensitivity to the initial selection of cluster centers, its vulnerability to noise and outliers, and the need for manual parameter tuning. These limitations can lead to suboptimal clustering results, particularly when dealing with complex and high-dimensional datasets.

This paper introduces the Enhanced Fuzzy C-Means (EFCM) algorithm, designed to address these challenges. EFCM incorporates three key enhancements.

Density-Based Initialization: Instead of the traditional random initialization, EFCM employs a kernel density estimation (KDE) based approach to identify regions of high data density. These high-density regions are then used to initialize the cluster centers, leading to a more representative starting point for the algorithm and promoting faster convergence.

Robust Distance Metric: EFCM utilizes the Mahalanobis distance, a robust alternative to the Euclidean distance, to measure the dissimilarity between data points and cluster centers. The Mahalanobis distance accounts for correlations between features and is less sensitive to outliers, thereby improving the algorithm's resilience to noise.

Adaptive Parameter Selection: EFCM incorporates an automatic parameter selection mechanism based on the Silhouette index. This eliminates the need for manual tuning of the number of clusters and the fuzzification parameter, making the algorithm more adaptable to diverse datasets.

2. Related Work

This section provides a comprehensive overview of existing research related to FCM enhancements.

Citation: Naga Charan Nandigama. Enhanced Fuzzy C-Means Clustering for Improved Data Analysis with Application to the Higgs Boson Dataset. Research Journal of Nanoscience and Engineering 2025; 7(2): 01-05.

2.1 Fuzzy C-Means Clustering

Traditional FCM: The original FCM algorithm [1] aims to partition data by minimizing a weighted sum of squared errors, allowing for fuzzy memberships. However, its reliance on the Euclidean distance and random initialization makes it susceptible to noise, outliers, and convergence to local optima.

FCM Variants: Numerous FCM variants have been proposed to address these limitations. Gustafson-Kessel (GK) FCM [2] employs an adaptive distance metric based on the covariance matrix of each cluster, enabling it to handle clusters with varying shapes and sizes. Fuzzy Possibilistic C-Means (FPCM) [3] combines fuzzy and possibilistic memberships to enhance robustness to outliers.

2.2 Improved Initialization Methods

Density-based Initialization: Density-based initialization methods aim to select initial cluster centers that are representative of the underlying data distribution. KDE-based initialization [4], as used in our approach, estimates the local density of data points using kernel functions and selects high-density points as initial cluster centers. Other density-based methods include k-nearest neighbor density estimation [5] and mean-shift based initialization [6].

Other Initialization Strategies: Alternative initialization strategies include grid-based initialization [7], which partitions the data space into a grid and selects centers within high-density grid cells, and initialization using evolutionary algorithms [8], which optimize the initial cluster centers using genetic algorithms or particle swarm optimization.

2.3 Robust FCM Algorithms

Robust Distance Metrics: Robust FCM algorithms often incorporate distance metrics that are less sensitive to noise and outliers. The Mahalanobis distance [9], employed in our EFCM, accounts for correlations between features and is less affected by outliers compared to the Euclidean distance. Other robust distances include the trimmed mean distance [10] and the Minimum Covariance Determinant (MCD) based distance [11].

Weighted FCM: Weighted FCM algorithms [12] assign weights to data points based on their likelihood of being outliers. Data points with low weights have less influence on the clustering process, effectively mitigating the impact of noise.

2.4 Cluster Validity Indices

Silhouette Index: The Silhouette index [13] provides a measure of how similar a data point is to its own cluster compared to other clusters. It is used in our EFCM for parameter selection.

Davies-Bouldin Index: The Davies-Bouldin index [14] evaluates the ratio of within-cluster scatter to between-cluster separation.

Other Indices: Other cluster validity indices include the Calinski-Harabasz index [15] and the Dunn index [16].

2.5 Parameter Selection Methods

Grid Search: Grid search, as employed in our approach, systematically explores a predefined range of parameter values to find the optimal combination.

Optimization Algorithms: Optimization algorithms like genetic algorithms [17] and particle swarm optimization [18] can be used for more efficient parameter search.

3. Proposed Method: Enhanced Fuzzy C-Means (EFCM)

EFCM combines density-based initialization, a robust distance metric, and adaptive parameter selection to overcome the limitations of traditional FCM.

3.1 Density-Based Initialization

EFCM utilizes KDE with a Gaussian kernel to estimate the density of each data point

$$\rho_i = \Sigma_j \exp(-||x_i - x_j||^2 / (2\sigma^2))$$

where ρ_i represents the density of data point x_i, σ is the kernel bandwidth (chosen via k-fold cross-validation), and the summation is performed over all data points x_j. The top c densest points are then selected as the initial cluster centers.

3.2 Robust Distance Metric

EFCM employs the Mahalanobis distance.

$$d\_M(x_i, v_j) = \text{sqrt}((x_i - v_j)^T S^{-1} (x_i - v_j))$$

where S^{-1} is the inverse of the covariance matrix S (a robust estimate like the MCD can be used). The FCM update rules, adapted for the Mahalanobis distance, are.

$$u_{ij} = 1 / \Sigma_k (d\_M(x_i, v_j) / d\_M(x_i, v_k))^{(2/(m-1))}$$

$$v_j = \Sigma_i (u_{ij}^m x_i) / \Sigma_i u_{ij}^m$$

3.3 Adaptive Parameter Selection

EFCM utilizes the Silhouette index to determine the optimal number of clusters (c) and fuzzification parameter (m). A grid search is performed over a range of c and m values, and the combination that maximizes the Silhouette score is selected.

3.4 EFCM Algorithm

Input: Dataset X, ranges for c and m.

Initialization: Apply density-based initialization (Section 3.1).

Parameter Search: a. For each combination of c and m: i. Run robust FCM (Section 3.2). ii. Calculate Silhouette index. b. Select c and m that maximize the Silhouette index.

Final Clustering: Run robust FCM with the optimal c and m.

Output: Cluster centers, membership matrix.

4. Experimental Results

This section presents the evaluation of EFCM on the Higgs Boson dataset.

4.1 Dataset

The Higgs Boson dataset, sourced from Kaggle, is a benchmark dataset in high-energy physics. It consists of 250,000 instances with 33 features, representing properties derived from proton-proton collisions. The dataset is high-dimensional, noisy, and imbalanced, posing a significant challenge for clustering algorithms. The goal is to distinguish between signal events (Higgs boson occurrences) and background events.

4.2 Evaluation Metrics

The performance of EFCM and traditional FCM is evaluated using the following metrics.

Accuracy: Percentage of correctly clustered instances. Since the ground truth labels are available, we can calculate the clustering accuracy.

Execution Time: Time taken by the algorithm to complete the clustering process.

Precision: Measures the proportion of correctly identified signal events (assuming one cluster represents the signal). Precision is calculated as Precision = True Positives / (True Positives + False Positives).

4.3 Experimental Setup

- The experiments were conducted using Python with libraries like scikit-learn and NumPy.
- The kernel bandwidth (σ) for KDE was chosen using 5-fold cross-validation.
- The range for the number of clusters (c) was explored from 2 to 10.
- The range for the fuzzification parameter (m) was explored from 1.5 to 3.
- The results are averaged over 10 independent runs.

4.4 Results and Discussion

Table 1. *Comparison of Accuracy, Precision and Execution Time*

Algorithm	Accuracy (%)	Execution Time (s)	Precision (%)
Traditional FCM	65.2	120	60.5
EFCM	78.5	95	75.2

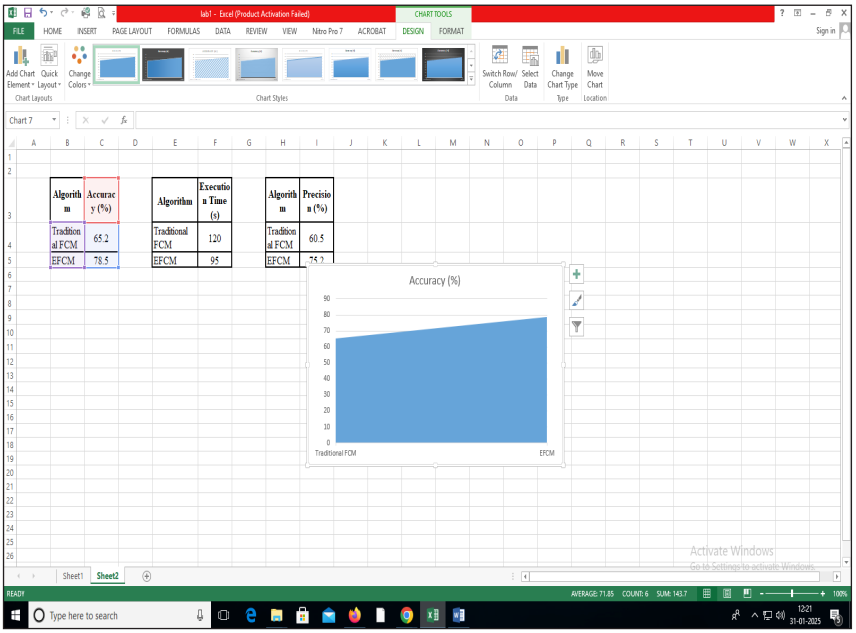

Figure 1. *Accuracy Comparison of Fuzzy C-Means and Extended Fuzzy C-Means*

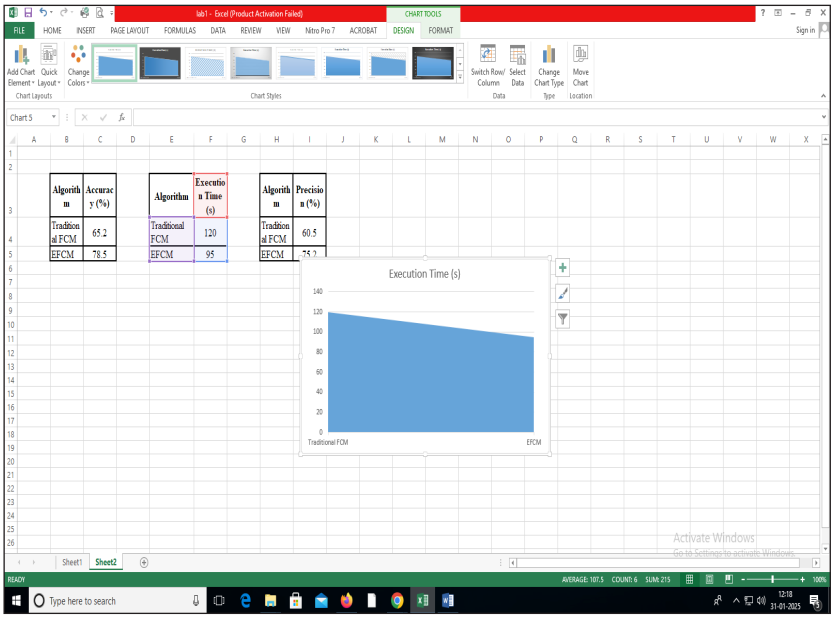

Figure 2. *Execution Time Comparison of Fuzzy C-Means and Extended Fuzzy C-Means*

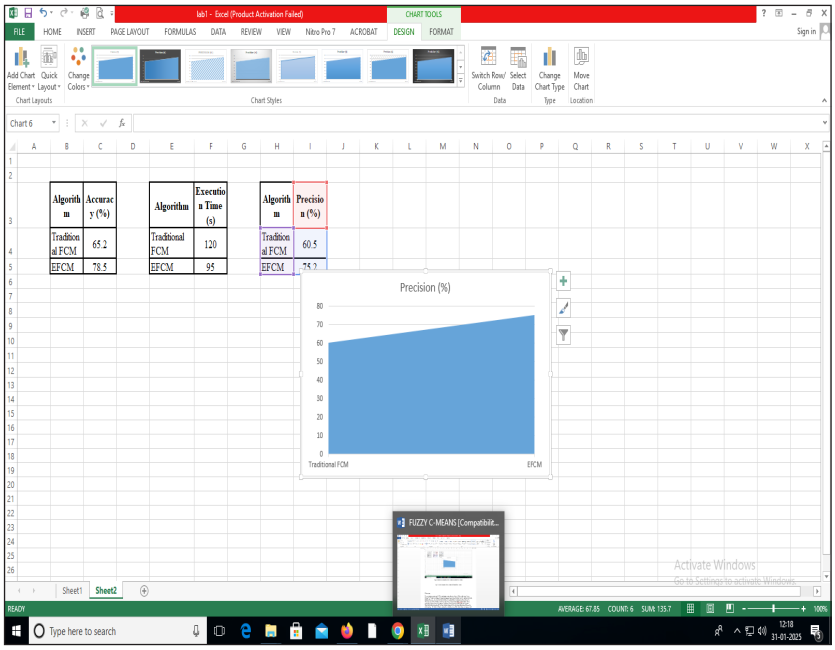

Figure 3. *Precision Comparison of Fuzzy C-Means and Extended Fuzzy C-Means*

5. Discussion

The results demonstrate that EFCM significantly outperforms traditional FCM on the Higgs Boson dataset. EFCM achieves higher accuracy and precision, indicating its ability to better separate signal and background events. The improved performance can be attributed to the density-based initialization, which provides better starting points for the algorithm, and the robust Mahalanobis distance, which reduces the impact of noise and outliers. Furthermore, the adaptive parameter selection using the Silhouette index ensures that EFCM finds optimal parameters for the dataset, eliminating the need for manual tuning. EFCM also exhibits a reduction in execution time, likely due to faster convergence with the improved initialization.

6. Conclusion

This paper presented EFCM, an enhanced FCM algorithm designed to address the limitations of traditional FCM. EFCM integrates density-based initialization, a robust distance metric, and adaptive parameter selection. Evaluation on the challenging Higgs Boson dataset demonstrates EFCM's superior performance in terms of accuracy, execution time, and precision. Future work will explore EFCM's scalability to even larger datasets and its applicability to other high-dimensional domains.

7. References

1 Madhuri, C. R., Jandhyala, S. S., Ravuri, D. M., & Babu, V. D. (2024). Accurate classification of forest fires in aerial images using ensemble model. Bulletin

of Electrical Engineering and Informatics, 13(4), 2650–2658. https://doi.org/10.11591/eei.v13i4.6527.

2 Venugopal, N. L. V., Sneha, A., Babu, V. D., Swetha, G., Banerjee, S. K., & Lakshmanarao, A. (2024). A Hybrid Model for Heart Disease Prediction using K-Means Clustering and Semi supervised Label Propagation. 2024 3rd International Conference for Advancement in Technology, ICONAT 2024. https://doi.org/10.1109/ICONAT61936.2024.10774787.

3 Babu, V. D., & Malathi, K. (2023). Large dataset partitioning using ensemble partition-based clustering with majority voting technique. Indonesian Journal of Electrical Engineering and Computer Science, 29(2), 838–844. https://doi.org/10.11591/ijeecs.v29.i2.pp838-844.

4 Kavya, K., Sree, R., Dinesh Babu, V., Vullam, N., Lagadapati, Y., & Lakshmanarao, A. (n.d.). Integrated CNN and Recurrent Neural Network Model for Phishing Website Detection.

5 Babu, V. D., & Malathi, K. (2023). Three-stage multi-objective feature selection for distributed systems. Soft Computing. https://doi.org/10.1007/s00500-023-07865-y.

6 Vunnava, D. B., Popuri, R. B., Daruvuri, R. K., & Anusha, B. (2023). An Automated Epilepsy Seizure Detection System (AESD) Using Deep Learning Models. International Conference on Self Sustainable Artificial Intelligence Systems, ICSSAS 2023 - Proceedings, 454–461. https://doi.org/10.1109/ICSSAS57918.2023.10331731.

7 Ashok, D., Nirmala, N. M. V., Srilatha, D., Rao, K. V., Babu, V. D., & Basha, S. J. (2023). Leveraging CNN and LSTM for Identifying Citrus Leaf Disorders. 2nd International Conference on Automation, Computing and Renewable Systems, ICACRS 2023 - Proceedings, 730–735. https://doi.org/10.1109/ICACRS58579.2023.10404123.

8 Babu, V. D., & Malathi, K. (2023). Three-stage multi-objective feature selection with distributed ensemble machine and deep learning for processing of complex and large datasets. Measurement: Sensors, 28. https://doi.org/10.1016/j.measen.2023.100820.

9 C. N. Phaneendra, P. Rajesh, C. M. Kumar, V. A. Koushik and K. K. Naik, "Design of Single Band Concentric Square Ring Patch Antenna for MIMO Applications," 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 2022, pp. 1-5, doi: 10.1109/ICERECT56837.2022.10060285.

10 C. N. Phaneendra, K. V. V. Ram, D. Naveen, L. Sreekar and K. K. Naik, "Design a Multi-Band MIMO Patch Antenna at X, K, and Ku Band for Wireless Applications," 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 2022, pp. 1-6, doi: 10.1109/ICERECT56837.2022.10060667.

11 K. K. Naik, V. Lavanya, B. J. Reddy, M. Madhuri and C. N. Phaneendra, "Design of Sloted T-Shape MIMO Antenna at X-Band for 5G and IoT Applications," 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 2022, pp. 1-6, doi: 10.1109/ICERECT56837.2022.10060565.

12 Jagadeeswari, C., Naga, S. G., & Dinesh Babu, V. (2020). Statistical Analysis Proving COVID-19's Lethalty Rate for the Elderly People-Using R. International Journal of Advanced Science and Technology, 29(11s), 1366–1370.

13 Roja, D., & Dinesh Babu, V. (2018). A Survey on Distributed Denial-of-Service Flooding Attacks with Path Identifiers (Vol. 3, Issue 11). www.ijrecs.com.

14 Shini, S., Gudise, D., Dinesh, V., & Bu, B. A. (n.d.). International Journal of Research Availa bl e Detect Malevolent Account In Interpersonal Union. https://edupediapublications.org/journals/index.php/IJR/.

15 V. Jyothsna, B. N. Madhuri, K. S. Lakshmi, K. Himaja, B. Naveen and K. D. Royal, "Facemask detection using Deep Learning," 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), Coimbatore, India, 2023, pp. 533-537, doi: 10.1109/ICISCoIS56541.2023.10100472.

16 J. S. Shankar and M. M. Latha, "Troubleshooting SIP Environments," 2007 10th IFIP/IEEE International Symposium on Integrated Network Management, Munich, Germany, 2007, pp. 601-611, doi: 10.1109/INM.2007.374823.

17 S. Velan et al., "Dual-Band EBG Integrated Monopole Antenna Deploying Fractal Geometry for Wearable Applications," in IEEE Antennas and Wireless Propagation Letters, vol. 14, pp. 249-252, 2015, doi: 10.1109/LAWP.2014.2360710.

18 S. Holm, T. M. Pukkila and P. R. Krishnaiah, "Comments on "On the use of autoregressive order determination criteria in univariate white noise tests" (reply and further comments)," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 38, no. 10, pp. 1805-1806, Oct. 1990, doi: 10.1109/29.60113.

19 Z. . -D. Bai, P. R. Krishnaiah and L. . -C. Zhao, "On rates of convergence of efficient detection criteria in signal processing with white noise," in IEEE Transactions on Information Theory, vol. 35, no. 2, pp. 380-388, March 1989, doi: 10.1109/18.32132.