IRFCA: A Hybrid Fuzzy-Possibilistic Clustering Algorithm for High-Dimensional Data

Sumrana Siddiqui
Computer Science Engineering Department, GITAM School of Technology, India
ssiddiqu@gitam.in

Nandita Bhanja Chaudhuri
Computer Science Engineering Department, GITAM School of Technology, India
nchaudhu@gitam.edu (corresponding author)

Received: 11 December 2025 | Revised: 29 December 2025, 13 January 2026, and 25 January 2026 | Accepted: 27 January 2026

 DOI: https://doi.org/10.48084/etasr.16837

ABSTRACT

Clustering high-dimensional datasets poses significant challenges due to the "curse of dimensionality," noise, and outliers, which render traditional algorithms like FCM and K-Means ineffective. The primary objective of this study was to develop a robust framework that balances noise resilience with the scalability required for large-scale data. The proposed Intelligent Robust Fuzzy Clustering Algorithm (IRFCA) is a hybrid pipeline that integrates Truncated SVD for dimensionality reduction, K-Means++ for stable initialization, and a fused FCM-PCM mechanism with dynamic outlier trimming. Tested on large-scale datasets including Network Attack Data (10.9 GB) and CORD-19, IRFCA demonstrates superior clustering quality, achieving Silhouette scores up to 0.63 and Davies-Bouldin Indices as low as 0.34, significantly outperforming baselines (FCM, FCMedian, PCM). These results quantify IRFCA's ability to produce compact, well-separated clusters while maintaining competitive scalability. IRFCA offers a resilient solution for high-dimensional analysis, with transformative potential for cybersecurity, business analytics, and scientific research.

Keywords-hybrid clustering; high-dimensional data; Principal Component Analysis (PCA); fuzzy clustering; Possibilistic C-Means (PCM); outlier trimming

I. INTRODUCTION

High-dimensional data introduces the "curse of dimensionality," where distance metrics lose discriminatory power, rendering traditional algorithms such as K-Means and FCM ineffective [1]. Furthermore, real-world datasets often contain noise and outliers that skew cluster centroids in standard soft clustering models. Existing solutions typically fail to balance the trade-off between robustness to noise and the computational efficiency required for large-scale datasets. Processing and analyzing large volumes of data face significant challenges, necessitating advanced optimization techniques to enhance convergence speed and accuracy in high-dimensional spaces [2]. This study addresses this gap by introducing IRFCA, a hybrid algorithm integrating PCA, K-Means++ initialization, and a fused FCM-PCM approach with dynamic outlier trimming.

- Design of IRFCA: The study introduces Intelligent Robust Fuzzy Clustering, a novel three-phase hybrid pipeline that integrates Principal Component Analysis (PCA), K-Means++ initialization, and a fused FCM-PCM approach with dynamic outlier trimming.
- Enhanced Clustering Quality: Empirical results across multiple domains demonstrate that IRFCA consistently outperforms baseline algorithms—including FCM, FCMedian, and PCM—by achieving superior internal validity indices, specifically higher Silhouette scores and lower Davies-Bouldin Indices.
- Robustness to Noise and Outliers: By combining the noise tolerance of Possibilistic C-Means with a dynamic trimming mechanism that removes statistically extreme points, the algorithm provides a resilient solution for datasets where overlapping boundaries and class imbalances are prevalent.
- Scalability for Big Data: The framework addresses the "curse of dimensionality" and computational overhead through Truncated SVD and Dynamic Subset Selection, enabling the processing of large-scale datasets (such as the 10.9 GB Network Attack dataset) while maintaining competitive runtimes.
- Cross-Domain Validation: This study provides a comprehensive evaluation of the algorithm using diverse

real-world datasets from cybersecurity, data science, and healthcare, proving its broad applicability for identifying meaningful sub-patterns in complex data environments.

- Practical Deployment Insights: Through extensive stability and parameter sensitivity analysis, this study offers actionable guidance on balancing clustering precision and computational costs in practical high-stakes applications such as intrusion detection or medical research.

Through this synergistic design, this study effectively bridges the gap between noise robustness and computational efficiency required for modern large-scale data analysis.

A. Related Work and Research Gap

Recent advances in fuzzy clustering have focused on hybridizing algorithms to improve robustness. For instance, in [3], an accelerated FCM was based on affinity filtering to handle noise, while in [4], sensitivity clustering was explored for parameter estimation. Recent studies have integrated evolutionary algorithms with K-Means to optimize performance in high-dimensional spaces [1]. However, these methods often process data in batches without explicit mechanisms for dynamic outlier removal during the iterative process, leading to centroid distortion. IRFCA differentiates itself by embedding a dynamic trimming step directly into the hybrid FCM-PCM update loop, specifically targeting the trade-off between noise resilience and convergence speed.

Unlike FCM and PCM, which rely solely on distance or typicality, IRFCA integrates K-Means++ initialization and dynamic outlier trimming to enhance stability. Although fuzzy clustering captures uncertainty through probabilistic memberships, integrating it with non-fuzzy methods (such as K-Means) has been shown to improve robustness and predictive performance in complex datasets [2].

Research in fuzzy clustering initially focused on improving how clusters are separated and evaluated, particularly in noisy datasets and uncertain environments. The early frameworks laid the mathematical foundation for distance-driven membership assignment and cluster validation metrics, helping improve decision accuracy in clustering tasks. Later studies introduced stronger comparative indices and algorithmic structures that provided more reliable interpretations of clustering outcomes in practical applications.

Recent developments have moved toward hybrid and adaptive clustering techniques that refine membership values and produce more stable clusters. Improved optimization strategies and parameter-sensitive models enhance the reliability of clustering under complex conditions. In addition, scalable and noise-tolerant methods have been introduced to support high-dimensional and streaming data, making fuzzy clustering more suitable for real-time analytical and machine learning systems [5].

II. PROPOSED METHODOLOGY

The proposed IRFCA addresses the fundamental limitations of classical soft clustering through a three-phase hybrid pipeline integrating dimensionality reduction, dual-membership learning, and adaptive outlier control.

First, dimensionality reduction using Truncated SVD projects high-dimensional vectors into a compact latent space while retaining structural variance and suppressing noise, thus improving distance reliability and reducing computational cost. Second, the algorithm simultaneously executes Fuzzy C-Means and Possibilistic C-Means updates, allowing clusters to form even when the data contains overlapping boundaries or class imbalance. The fusion coefficient α balances uncertainty handling (FCM) and strong prototype anchoring (PCM), enabling IRFCA to stabilize cluster centers against non-Gaussian outliers. Finally, dynamic trimming removes statistically extreme points based on percentile distance thresholds, preventing them from skewing center updates while preserving cluster interpretability.

This hybrid design enables IRFCA to adaptively reorganize clusters as the structure of the dataset evolves, making it suitable for large-scale real-world data in cybersecurity, scientific literature, and behavioral analytics. The framework prioritizes robustness, compactness, and scalable computation, resulting in improved cluster validity indices without sacrificing runtime feasibility for big data workloads.

A. Algorithm Design and Components

Although traditional PCA is standard, this study employs Truncated SVD to project sparse high-dimensional data into a lower-dimensional latent space. This approach retains structural variance while suppressing noise and reducing computational complexity [6]. Modified Fuzzy C-Means combines FCM and PCM memberships with a weighting factor $\alpha = 0.7$. The following equation shows how the hybrid membership is computed to balance probabilistic and possibilistic information:

$$U_{ij}^{hybrid} = \alpha \cdot U_{ij}^{FCM} + (1-\alpha) \cdot U_{ij}^{PCM} \quad (1)$$

where U_{ij}^{hybrid} is the final integrated membership of data point i in cluster j, α is the fusion weighting factor (set to 0.7), which determines the relative influence of the fuzzy versus possibilistic components, U_{ij}^{FCM} is the probabilistic membership value derived from the standard Fuzzy C-Means logic, and U_{ij}^{PCM} is the typicality value derived from Possibilistic C-Means, representing how well a point belongs to a cluster regardless of other clusters.

The IRFCA framework minimizes a dual-term objective function iteratively to ensure both structural accuracy and noise resilience. The objective function is given by:

$$J_{hybrid} = \sum_{i=1}^{n}\sum_{j=1}^{c}\left(U_{ij}^{hybrid}\right)^{m} || X_i - C_j ||^2 + \sum_{j=1}^{c}\eta_j \sum_{i=1}^{n}\left(1 - U_{ij}^{PCM}\right)^{q} \quad (2)$$

minimized iteratively with K-means++ initialization, fuzzifier $m \in \{1.2, 1.5, 2.0\}$, and outlier trimming at proportions $\{0.001, 0.01\}$.

A random 5% subset of the data for initial clustering and iteratively refining the clusters for up to three iterations, stopping early if the change between successive iterations falls below a convergence threshold of 0.5. Table I presents the notations used in hybrid membership computations.

TABLE I. NOTATIONS USED IN HYBRID MEMBERSHIP

Notations	Meaning
J_{hybrid}	The total cost (objective) function to be minimized.
n	Total number of data points in the processed dataset.
c	Number of clusters specified for the algorithm.
m	Fuzzification exponent (fuzzifier) that controls the degree of overlap between clusters.
X_i	The i^{th} high-dimensional data vector.
C_j	The centroid (mean) of the j^{th} cluster.
η_j	Scale parameter for cluster j, determining the distance at which the possibilistic membership value becomes 0.5.
q	PCM penalty exponent, typically set equal to m to maintain balance during the optimization process.

Algorithm 1: Proposed IRFCA

Step 1: A small subset is selected to reduce computation and improve initialization. Centers are initialized using K-Means++.

$$X_s \subset X, \ |X_s| = S \quad (3)$$

$$V^{(0)} = KMeans^{++}(X_s)$$

Step 2: Probabilistic memberships are assigned based on relative distance to cluster centers.

$$U_{ij}^{fcm} = 1/\left(\sum_{k=1}^{c}\left(\frac{||x_i - v_j||^2}{||x_i - v_c||^2}\right)^{1/(m-1)}\right) \quad (4)$$

Step 3: A scale parameter is computed and PCM evaluates how typical a point is to a cluster.

$$\eta_j = \frac{\left(\sum_{\{i=1\}}^{N} U_{\{ij\}}^{\{fcm\}} ||x_i - v_j||^2\right)}{\left(\sum_{\{i=1\}}^{N} U_{\{ij\}}^{\{fcm\}}\right)} \quad (5)$$

$$u_{ij} = \frac{1}{1+\left(\frac{||x_i - v_j||}{\eta_j}\right)^{\frac{1}{m-1}}} \quad (6)$$

Step 4: FCM and PCM memberships are combined with a weight α and normalized.

$$U_{ij}^{hybrid} = \alpha \cdot U_{ij}^{FCM} + (1-\alpha) \cdot U_{ij}^{PCM}$$

$$U_{ij} = \frac{U_{ij}}{\sum_{c=1}^{k} U_{ic}} \quad (7)$$

Step 5: Centers are updated using weighted membership values.

$$v_j = \frac{\sum_{i=1}^{N}(U_{ij}^m \cdot x_i)}{\sum_{i=1}^{N} U_{ij}^m} \quad (8)$$

Step 6: Outliers are removed based on percentile distance threshold. x_i is an outlier if

$$\| x_i - v_j \|^2 > P_{(1-\theta)}\left(\| X - v_j \|^2\right) \quad (9)$$

Step 7: The algorithm stops when centroid changes fall below ε or maximum iterations are reached.

$$\| V^{(t+1)} - V^{(t)} \| < \varepsilon \quad (10)$$

III. EXPERIMENTAL SETUP

A. Datasets, Preprocessing, and Parameters

Preprocessing involved encoding categorical features, converting CORD-19 text to TF-IDF vectors, and standardizing all features to zero mean and unit variance. Missing values were imputed using the column mean. For the IRFCA parameters, the fuzzifier m was set to 2.0 to provide standard soft boundaries, and the fusion weighting factor α was set to 0.7 based on empirical testing to prioritize the probabilistic structural information of FCM while leveraging PCM for noise suppression. The trim proportion was set to 0.001 to remove extreme outliers without discarding valid data clusters.

To ensure practical utility, the selected datasets reflect real-world challenges of high dimensionality, noise, and scale, testing IRFCA's adaptability across cybersecurity, data science, and healthcare domains [7]. Table III details the datasets.

TABLE II. DATASET SUMMARY

Dataset	Source	Instances (rows)	Features	Domain
Network Attack Data	[8]	40,77,265	48	Cybersecurity / Network Security
Kaggle Public Metadata	[9]	2,34,23,543	12	Data Science / Competition Platforms
CORD – 19	[10]	10,58,609	~30,000 (TF-IDF)	Healthcare / Medical Research

B. Experimental Configuration

Parameters: Fuzzifier $m \in \{1.2 - 2.0\}$, trim proportion $\in \{0.001, 0.01\}$. Experiments used Python 3.9 on a 16-core, 64GB RAM workstation.

C. Visualization of Clustering Results

This section provides a comprehensive visual exploration of IRFCA's performance across three diverse datasets: Network Attack Data (10.9 GB) [8], Kaggle Public Metadata (4.45 GB) [9], and CORD-19 (1.53 GB) [10].

Figure 1 displays Silhouette score distributions, confirming IRFCA's superior cluster stability and quality against baselines on the Network Attack dataset [8]. Figure 2 compares performance metrics, illustrating IRFCA's favorable balance of low Davies-Bouldin indices and high stability despite increased computational demands.

Figure 3 shows Silhouette scores for the Kaggle dataset [9], highlighting IRFCA's consistent performance and robust cluster formation across diverse parameter settings. Figure 4 illustrates scalability trade-offs, demonstrating that IRFCA's higher memory usage is justified by significantly improved Inertia and Davies-Bouldin scores.

Figure 5 highlights clustering consistency on CORD-19 [10], where IRFCA achieves high Silhouette scores with minimal variance compared to unstable baselines. Figure 6 details efficiency metrics, confirming IRFCA's competitive runtime and superior cluster separation (Silhouette > 0.6) on unstructured textual data.

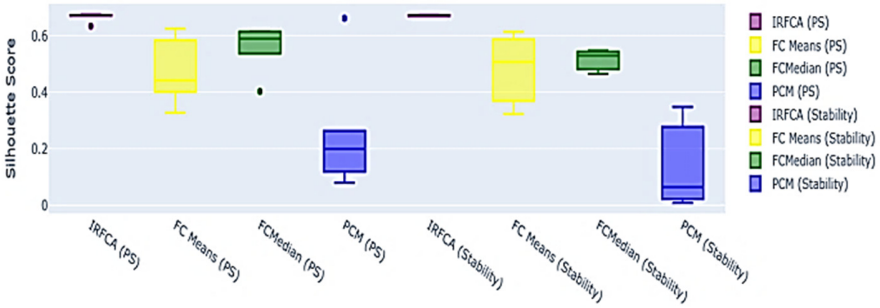

Fig. 1. Silhouette score distribution across all runs on the Network Attack dataset [8].

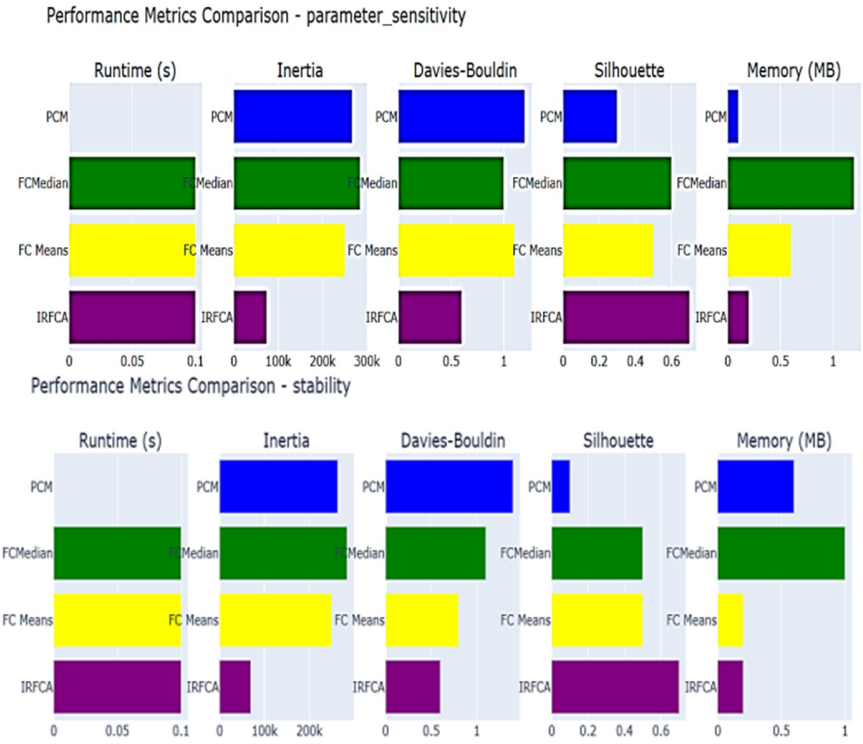

Fig. 2. Comparison of performance metrics on the Network Attack dataset [8].

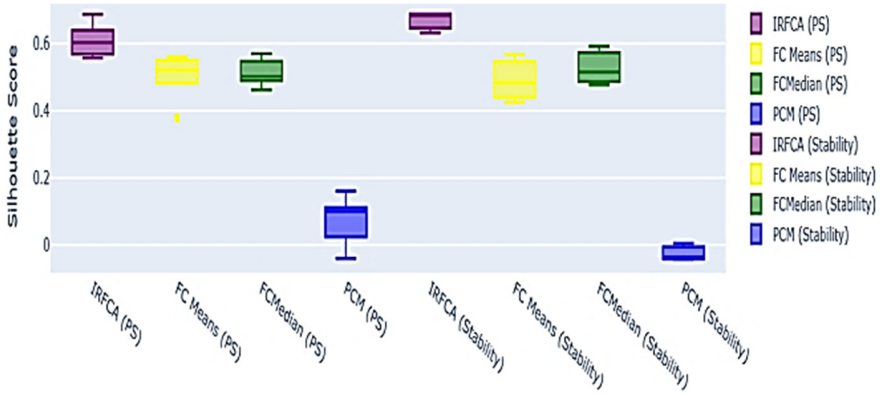

Fig. 3. Silhouette score distribution across all runs on Kaggle public metadata [9].

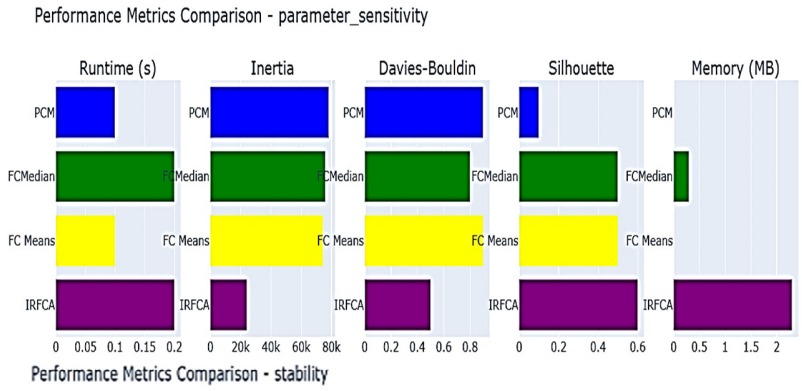

Fig. 4. Comparison of performance metrics on Kaggle public metadata [9].

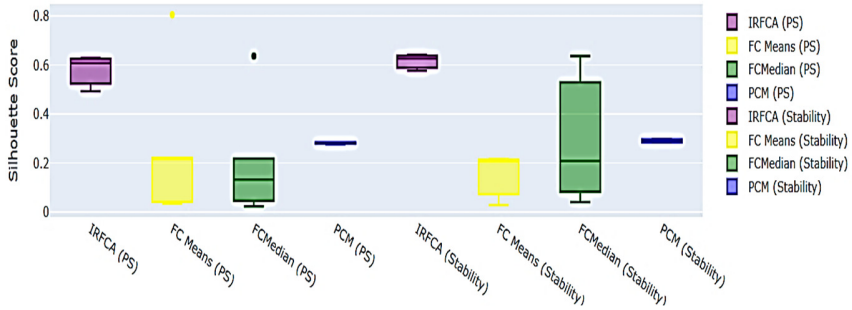

Fig. 5. Silhouette score distribution across all runs on the CORD-19 dataset [10].

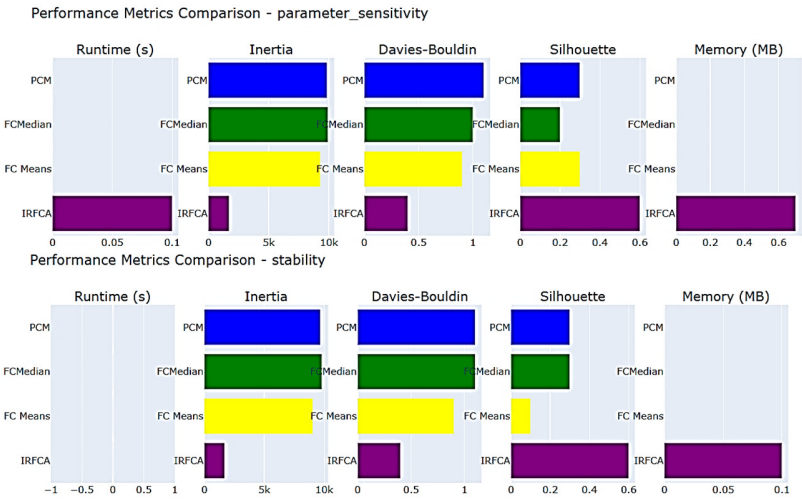

Fig. 6. Comparison of performance metrics on the CORD-19 dataset [10].

IV. RESULTS AND DISCUSSION

This section presents a detailed evaluation of IRFCA against baseline clustering methods—Fuzzy C-Means (FCM), Fuzzy C-Median (FCMedian), and Possibilistic C-Means (PCM)—across the three datasets. Performance was assessed using clustering quality metrics (Silhouette score, Davies-Bouldin Index, and Adjusted Rand Index-ARI) and efficiency metrics (runtime, memory usage), with insights derived from visualizations [3, 11-14]. Table III summarizes IRFCA's clustering performance compared to baseline algorithms across the three datasets, highlighting its superior Silhouette Scores and Davies-Bouldin Indices.

TABLE III. METRICS COMPARISON

Dataset	Algorithm	Silhouette score	Davies-Bouldin index	Key observations
Network Attack Data	IRFCA	0.63	0.66	Tighter, well-separated clusters.
	FCM	0.33	1.07	Struggles with noise.
	FCMedian	0.61	1.17	Good cohesion but poor separation (High DBI).
	PCM	0.21	0.91	Lacks scalability.
Kaggle Public Metadata	IRFCA	0.62	0.42	Robust pattern capture.
	FCM	0.55	0.66	Sensitive to high-dim noise.
	FCMedian	0.55	0.65	Limited by hard assignment constraints.
	PCM	0.02	1.02	Fails to identify structure.
CORD-19	IRFCA	0.62	0.34	Groups noisy text effectively.
	FCM	0.04	1.23	Sensitive to textual noise.
	FCMedian	0.02	1.22	Ineffective on sparse text data.
	PCM	0.28	1.14	Poor thematic clustering.

IRFCA's enhanced clustering quality comes with a trade-off in computational demands, but it remains competitive with baseline methods. Efficiency analysis shows that IRFCA incurs higher runtime and memory usage than simpler baselines, such as FCM, due to its hybrid architecture, which includes PCA, FCM-PCM integration, and dynamic outlier trimming [15]. However, this resource usage is justified by superior clustering quality, particularly in precision-critical applications such as network intrusion detection or thematic grouping of medical literature [16]. PCA's dimensionality reduction enhances scalability by reducing computational overhead, making IRFCA practical for large-scale datasets. Future optimizations, such as parallel processing or adaptive subset selection, could further reduce its computational footprint [17].

A. Parameter Sensitivity and Robustness

Parameter sensitivity tests provide insight into IRFCA's robustness across varying configurations. The higher fuzziness enhances the algorithm's ability to handle overlapping clusters, which is particularly beneficial for noisy datasets such as CORD-19, where textual metadata may include incomplete or ambiguous entries. Additionally, a trim proportion of 0.001 was found to maximize cluster precision across all datasets, with increasing it to 0.01 leading to diminishing returns due to elevated computational costs. These findings underscore the importance of parameter tuning to balance clustering quality and efficiency, particularly in datasets with varying noise levels and dimensionality. The results align with the research focus on outlier management, demonstrating IRFCA's adaptability to diverse data characteristics through its dynamic outlier trimming mechanism [18].

B. Stability and Convergence

Stability and convergence analyses highlight both strengths and limitations of IRFCA. Across all datasets, the algorithm exhibits rapid convergence, with centroid changes stabilizing within two to three iterations. Similarly, on the Network Attack and Kaggle Public Metadata datasets, convergence occurs within three iterations, reflecting the efficiency of K-Means++ initialization and hybrid FCM-PCM updates. However, stability analysis reveals a challenge; Although IRFCA achieved superior Silhouette scores, the ARI values were lower than baselines [13]. The discrepancy between superior internal validity (Silhouette) and lower external validity (ARI) indicates that IRFCA identifies compact latent structures based on intrinsic feature geometry that do not strictly align with pre-assigned semantic labels. For instance, in the Network Attack dataset, while ground truth labels broadly categorize attacks (e.g., DoS), IRFCA's dynamic outlier trimming likely segregates them into more granular subclusters based on packet volume intensity or protocol anomalies. This suggests that the algorithm is discovering novel, physically meaningful subpatterns in the high-dimensional space that traditional broad labels fail to capture, rather than producing erroneous groupings [19].

V. CONCLUSION AND FUTURE WORK

IRFCA is a hybrid clustering algorithm that addresses the limitations of high-dimensional data analysis. By integrating PCA, K-Means++ initialization, and dynamic outlier trimming, IRFCA outperformed FCM, FCMedian, and PCM across cybersecurity, data science, and healthcare datasets. The results confirm that IRFCA provides superior cluster validity indices (Silhouette scores consistently >0.60) and robustness to noise, albeit with a marginal increase in computational overhead. Future work will explore parallelization strategies to further enhance scalability.

AVAILABILITY OF DATA

The datasets used in this study are publicly available open-access datasets obtained from Kaggle and the Allen Institute for AI. Specifically, the Network Attack (Kitsune) dataset, the Meta Kaggle dataset, and the CORD-19 dataset are freely accessible for research and educational use under their respective licenses and terms of service. No proprietary or restricted data were used. All datasets were utilized in compliance with the original providers' usage and attribution requirements.

REFERENCES

[1] N. L. G. P. Suwirmayanti, I. K. G. D. Putra, M. Sudarma, I. M. Sukarsa, E. Setyaningsih, and R. A. N. Diaz, "Invasive Weed Optimization K-Means Performance Robust Operations (IWOKM PRO) in High-

Dimensional Datasets," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 24390–24395, Aug. 2025, https://doi.org/10.48084/etasr.11112.

[2] E. A. Sarwoko, E. Vianita, and A. Wibowo, "Evaluating the Integration of Fuzzy and Non-Fuzzy Clustering Approaches into LSTM for the Power Consumption Forecasting Utilizing the Case Study Dataset of Tetuan City," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 26689–26696, Oct. 2025, https://doi.org/10.48084/etasr.11938.

[3] D. Li, S. Zhou, and W. Pedrycz, "Accelerated Fuzzy C-Means Clustering Based on New Affinity Filtering and Membership Scaling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12337–12349, Dec. 2023, https://doi.org/10.1109/TKDE.2023.3273274.

[4] H. Chhajer and R. Roy, "Rationalised experiment design for parameter estimation with sensitivity clustering," *Scientific Reports*, vol. 14, no. 1, Oct. 2024, Art. no. 25864, https://doi.org/10.1038/s41598-024-75539-2.

[5] P. S. Thakur and S. Mohapatra, "A Survey of Fuzzy Clustering Methods and Validation Approaches," in *Data Science and Applications*, vol. 1266, S. J. Nanda, R. P. Yadav, A. H. Gandomi, and M. Saraswat, Eds. Springer Nature Singapore, 2025, pp. 183–194.

[6] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, Jan. 2011, https://doi.org/10.1137/090771806.

[7] X. Zhang *et al.*, "ResTune: Resource Oriented Tuning Boosted by Meta-Learning for Cloud Databases," in *Proceedings of the 2021 International Conference on Management of Data*, June 2021, pp. 2102–2114, https://doi.org/10.1145/3448016.3457291.

[8] "Kitsune Network Attack Dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/ymirsky/network-attack-dataset-kitsune.

[9] "Meta Kaggle." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/kaggle/meta-kaggle.

[10] "COVID-19 Open Research Dataset Challenge (CORD-19)." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge.

[11] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, https://doi.org/10.1109/TPAMI.1979.4766909.

[12] M. J. Warrens and H. Van Der Hoef, "Understanding the Adjusted Rand Index and Other Partition Comparison Indices Based on Counting Object Pairs," *Journal of Classification*, vol. 39, no. 3, pp. 487–509, Nov. 2022, https://doi.org/10.1007/s00357-022-09413-z.

[13] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[14] H. Li and J. Wang, "From Soft Clustering to Hard Clustering: A Collaborative Annealing Fuzzy c-Means Algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 32, no. 3, pp. 1181–1194, Mar. 2024, https://doi.org/10.1109/TFUZZ.2023.3319663.

[15] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, https://doi.org/10.1109/TIT.1982.1056489.

[16] M. M. U. Rony *et al.*, "Augmenting Visualizations with Predictive and Investigative Insights to Facilitate Decision Making," in *Companion Proceedings of the ACM Web Conference 2023*, Dec. 2023, pp. 77–81, https://doi.org/10.1145/3543873.3587317.

[17] F. Ros, R. Riad, and S. Guillaume, "PDBI: A partitioning Davies-Bouldin index for clustering evaluation," *Neurocomputing*, vol. 528, pp. 178–199, Apr. 2023, https://doi.org/10.1016/j.neucom.2023.01.043.

[18] N. Cicekli and I. Cicekli, "Formalizing the specification and execution of workflows using the event calculus," *Information Sciences*, vol. 176, no. 15, pp. 2227–2267, Aug. 2006, https://doi.org/10.1016/j.ins.2005.10.007.

[19] D. Feldman, M. Schmidt, and C. Sohler, "Turning Big Data Into Tiny Data: Constant-Size Coresets for k-Means, PCA, and Projective Clustering," *SIAM Journal on Computing*, vol. 49, no. 3, pp. 601–657, Jan. 2020, https://doi.org/10.1137/18M1209854.