

Optimizing diabetes prediction: unveiling patient subgroups through clustering

Rita Ganguly¹, Dharmpal Singh², Rajesh Bose²

¹Department of Computer Applications, Dr. BC Roy Academy of Professional Courses, Durgapur, India

²Department of Computer Science, Faculty of Engineering and Technology, JIS University, Kolkata, India

Article Info

Article history:

Received Jul 17, 2024

Revised Jun 24, 2025

Accepted Jul 13, 2025

Keywords:

Clustering method

Diabetes

Fuzzy C-means

Hierarchical clustering

K-means

Pima diabetes dataset

Silhouette score

ABSTRACT

Diabetes is a significant global health concern, leading to numerous deaths annually and affecting many individuals who remain undiagnosed. As its prevalence rises, the importance of early detection becomes increasingly vital. The rising diabetes epidemic demands data-driven strategies to catch health problems sooner and identify them clearly. This study utilizes the Pima Indians diabetes dataset (PIDD) to compare three powerful clustering schemes such as k-means, fuzzy C-means, and hierarchical. Uncontrolled diabetes, arising from the body's struggle to manage blood sugar due to insulin deficiency, can lead to devastating complications. Early detection and intervention are the cornerstones of effective management and improved patient outcomes. This study breaks new ground by meticulously evaluating the performance of each clustering algorithm using advanced metrics like silhouette score and adjusted Rand index. The goal is to identify the method that generates the most accurate and well-defined clusters for diabetes-related attributes. This, in turn, has the potential to revolutionize diabetes diagnosis, enabling earlier interventions and ultimately leading to better disease management and patient care. By providing a comprehensive comparison of these clustering techniques, this research offers a significant contribution to the fight against diabetes.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Rita Ganguly

Department of Computer Applications, Dr. BC Roy Academy of Professional Courses

Fuljhore, Durgapur (Pacchim Burdwan)-713206, India

Email: ganguly.rita@gmail.com

1. INTRODUCTION

Healthcare, a cornerstone of societal well-being, is undergoing transformative changes driven by technological advancements. Among the myriad health challenges faced today, diabetes emerges as a significant global concern, largely influenced by lifestyle changes and increasing prevalence. This study explores innovative technological solutions, particularly data extraction and fuzzy logic, aimed at enhancing diabetes diagnosis [1]. Diabetes is characterized by poor glucose regulation, resulting from inadequate insulin production or ineffective insulin response, leading to chronic high blood sugar levels (hyperglycemia). While diabetes remains incurable, effective management strategies can significantly improve patient outcomes [2]. The escalating incidence of diabetes necessitates new technological interventions to facilitate early detection and treatment. Machine learning (ML) offers promising opportunities to advance existing healthcare technologies within the broader context of the fourth industrial revolution, which encompasses the internet of things (IoT), artificial intelligence (AI), data mining, and neural networks. Despite previous research efforts, early diabetes detection remains challenging, with traditional methods often falling short in providing

comprehensive solutions. This drives our investigation into data-driven approaches, focusing on data mining and fuzzy logic, to enhance diagnostic accuracy [3].

This research addresses the critical issue of early diabetes diagnosis through the introduction of a novel clustering method, evaluated against established algorithms. The study contributes to the field in two significant ways:

- Comprehensive analysis: it presents a detailed comparison of existing clustering techniques used in diabetes prediction, highlighting their strengths and limitations. This comparative analysis will serve as a valuable resource for guiding future research.
- Innovative clustering method: the study introduces a new clustering method specifically designed for diabetes diagnosis, demonstrating marked improvements in accuracy over conventional techniques.

The literature review is opened with the work of Ibrahim *et al.* [4], which proposes a new hybrid approach that combines agglomerative hierarchical clustering (HC) with a decision tree classifier to improve accuracy, attaining an 80.8% rating as opposed to their considered typical decision tree classifier with an accuracy rating of 76.9%. Dong *et al.* [5] contributed a procedure using fuzzy model in HC that identifies clusters of complex and intricate shapes. That algorithm revered outstanding performance specifically to high-dimensional and large datasets. Padmaja *et al.* [6] have taken into consideration the task of identifying high-quality clusters and made a deep analysis of different algorithms for clustering. Ghosh *et al.* [7] contributed to the effectiveness of the aggregation pheromone clustering (APC) algorithm by showing that it is much better concerning the quality of clustering and speed of processing for all the datasets taken. Bagirov [8] proposed a global k-means (KM) algorithm that showed its effectiveness by being tested on 14 datasets using numerical experiments, though it consumed more computational time. Finally, Nithya *et al.* [9] have conducted a comparative study on HC, density-based spatial clustering of applications with noise (DBSCAN), and simple KM scheme and found that the KM algorithm works best on the diabetes dataset. Cebeci and Yildiz [10] also found the KM algorithm to be faster in execution as compared to the fuzzy C-means (FCM) technique for all the datasets tested, independent of the type of pattern in the base dataset. This tendency towards the KM scheme was further confirmed by Biradar and Mugali [11], who applied different tools to the diabetes dataset. Qi *et al.* [12] designed an approach to improve clustering by choosing initial centers with great care, thereby substantially improving the likelihood of obtaining optimal local solutions. Saravananathan and Velmurugan [13] focused on analyzing the execution time of both KM and FCM techniques, with KM consistently outperforming FCM in terms of execution time.

In summary, the research provided a comprehensive evaluation of clustering algorithm performance, underscoring the hybrid model's promising outcomes and the sustained efficiency of the KM technique, making it the preferred choice for large datasets. Additionally, it emphasized the importance of identifying high-quality clusters as a means to augment clustering algorithm effectiveness. Orabi *et al.* [14] introduces an early predictive system for diabetes mellitus by integrating data mining techniques, demonstrating improved prediction accuracy through tailored preprocessing and classification methods. Patil *et al.* [15] propose a hybrid prediction model for type-2 diabetes that combines decision trees and adaptive neuro-fuzzy inference systems (ANFIS), yielding superior performance compared to standalone models. Zhao *et al.* [16] contribute to the evaluation of clustering quality by presenting a sum-of-squares-based cluster validity index with significance analysis, enabling better assessment and selection of clustering results. Bahmani *et al.* [17] address the scalability challenges in clustering large datasets through an optimized k-means++ algorithm, achieving faster execution times while maintaining high clustering accuracy. Karegowda *et al.* [18] explore a cascading approach that integrates k-means clustering with k-nearest neighbor classification for categorizing diabetic patients, highlighting improved classification precision through a two-stage processing framework. Thakkar *et al.* [19] compared data mining and fuzzy logic techniques for diabetes prognosis, noting that data mining methods like decision trees and support vector machines (SVM) provide high accuracy but limited interpretability. In contrast, fuzzy logic handles uncertainty well and offers transparent, rule-based reasoning aligned with clinical practice.

The structure of this paper is as follows: section 2 details the methodology, outlining the comparative study of existing procedures and the proposed new method. Section 3 presents the comparison results, followed by a thorough discussion of the findings. Finally, section 4 summarizes key insights, emphasizing the implications and contributions of the novel method while suggesting avenues for further research and development in diabetes diagnosis technology.

2. METHOD

The objective of this research is to propose a decision-making clustering approach for handling diabetes-related attributes in the Pima Indians diabetes dataset (PIDD). This section outlines the systematic methodology employed to classify diabetes attributes for the initial detection and prediction of diabetes, detailing the experimental procedures followed to ensure reproducibility.

2.1. Data pre-processing

There are two methods for data preprocessing:

2.1.1. Data extraction and cleaning

The PIDD was extracted and examined for quality. Missing values were addressed using several imputation techniques. These included mean or median imputation, forward and backward fill, multiple imputation, and model-based imputation.

2.1.2. Normalization and scaling

To prevent bias toward attributes with larger values, normalization was performed on all dataset attributes. The normalization equation used was:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where X is the original attribute value; X_{\min} and X_{\max} are the lowest and extreme values of the attribute respectively; and X_{norm} is the normalized attribute value.

2.2. Comparative analysis of clustering algorithms

Three popular clustering algorithms are selected for analysis: KM, FCM, and HC. Each clustering algorithm is applied to the pre-processed dataset to create clusters of diabetes-related attributes. The fundamental principles of each clustering algorithm are studied and understood.

2.2.1. K-means clustering

K-means seeks to group data so that points in the same cluster are as close as possible to their cluster's center. Objective function:

$$J = \sum_{i=1}^k \sum_{j=1}^{n_i} ||X_j - \mu_i||^2$$

where the number of clusters is represented by K ; data points in the i th cluster is n_i ; j th data point in the i th cluster is $X_j^{[i]}$; and centroid of the i th group is μ_i .

2.2.2. Fuzzy C-means clustering

Fuzzy C-means allows partial membership of data points in multiple clusters. Objective function:

$$J_m = \sum_{i=1}^n \sum_{j=1}^k U_{ij} ||X_i - C_j||^2$$

where n is the data points; the number of clusters is k ; the membership degree of X_i in the j th cluster is U_{ij} ; the fuzziness exponent is m ; i th data point is X_i ; Mid-point of the j th cluster is C_j .

2.2.3. Hierarchical clustering

HC forms a tree-like structure (dendrogram) of nested clusters. Linkage function:

$$d(A, B) = \sqrt{\frac{2|A||B|}{|A| + |B|}} ||\mu_A - \mu_B||^2$$

where A and B are two clusters; $|A|$ and $|B|$ are the sizes of clusters A and B separately; and μ_A and μ_B are the centroids of clusters A and B separately.

2.3. Valuation metrics

Numerous key metrics are utilized to calculate the clustering algorithm's performance:

- Silhouette score determines the firmness and cluster separation.

$$S = \frac{b-a}{\max(a,b)}$$

where regular intra-cluster distance is a ; and typical adjacent-cluster distance is b .

- Adjusted Rand index (ARI) score compares the similarity between true class labels and cluster assignments.

$$ARI = \frac{RI - E|RI|}{\max|RI| - E|RI|}$$

where this focuses on the concept of $E|RI|$ as a measure of how well; and the Rand index (RI) is expected to perform on average.

- Normalized mutual information (NMI) score: the NMI quantifies the agreement between true class labels and cluster assignments, accounting for entropy.

$$NMI(U, V) = \frac{2I(U; V)}{H(U) + H(V)}$$

where the MI amid clusters U and V is $I(U; V)$; and the randomness of clusters U and V are $H(U)$ and $H(V)$ respectively.

- Davies-Bouldin index (DBI) score: defines the cluster superiority based on the distance between clusters.

$$DBI = \frac{1}{K} \sum_{i=1}^K \max_{i \neq j} \frac{(S_i + S_j)}{d(C_i, C_j)}$$

where S_i and S_j is the typical distance between each point in cluster i and the centroid C_i ; and $d(C_i, C_j)$ is the centroid distance between C_i and C_j .

These evaluation measures help researchers determine how well the clusters represent the underlying data patterns. By comparing the results across different methods, they can select the clustering approach that provides the most accurate and meaningful grouping, as depicted in Figure 1.

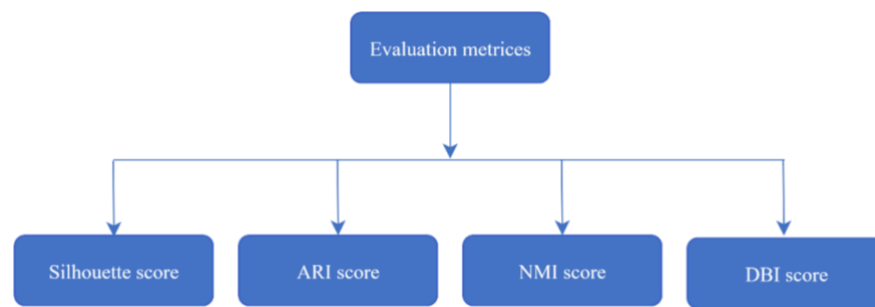


Figure 1. Evaluation metrics

2.4. Innovative clustering method formulation

A novel clustering method emerges from the comparative analysis and evaluation metrics. This method is designed to optimize the clustering of diabetes attributes and enhance the accuracy of disease prediction. It potentially does so by integrating elements from existing algorithms or introducing entirely new approaches.

2.5. Assessment and outcomes

The newly developed clustering method is implemented and benchmarked against KM, FCM, and HC using the PIDDD. Its performance is assessed using standard evaluation metrics. These metrics are used to demonstrate the effectiveness of the proposed method.

2.6. Implications and applications

The study considers the possibilities for the early and correct diagnosis of diabetes by improving clustering techniques for attributes related to diabetes. It investigates how innovative algorithms of clustering could be put into action to achieve optimality. These improvements aim to support strategies for effective health decision-making and disease management.

2.7. Highlighted impact

The challenge of diabetes is emerging and is being addressed through data-driven analytics in the research. It has come up with a new technique of clustering and has compared it with other methods to identify the gaps in the early diagnosis of diabetes. Extensive experiments are being conducted to come up with a diabetes attribute clustering technique better than the ones existing [20]. Through this structured

methodology, the research aims to address the challenge of diabetes early detection and management, demonstrating the power of data-driven approaches in healthcare analytics. By following these steps, future researchers can replicate the experiments and build upon the findings presented in Figure 2.

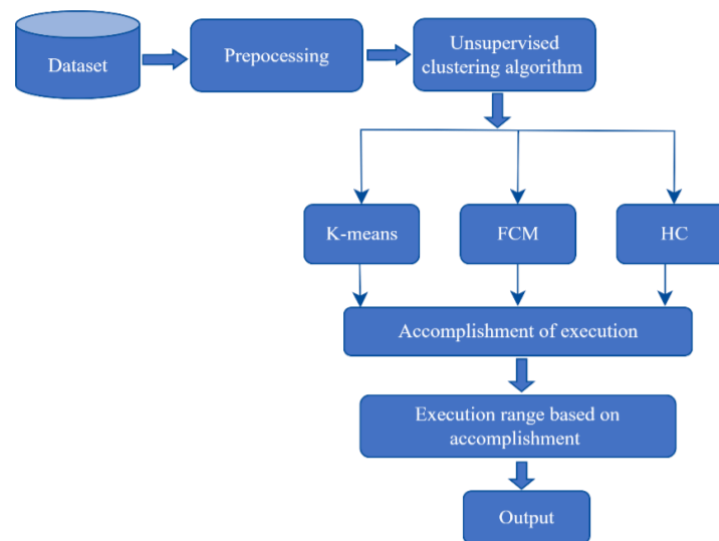


Figure 2. Proposed frame work for clustering

3. RESULTS AND DISCUSSION

3.1. Identifying gaps in previous research

This study investigates the performance of various clustering algorithms—KM, FCM, and HC—in the context of early diabetes diagnosis. While prior research has explored the efficacy of clustering techniques in healthcare data, many have not explicitly addressed how these methods can be optimized for datasets with imbalanced classes and missing values. This gap is particularly relevant in predicting diabetes outcomes.

3.2. Summarizing key findings

In this research work findings indicate that KM clustering produced the highest accuracy metrics, achieving a perfect score across multiple evaluation parameters. In contrast, FCM exhibited a significantly lower performance, particularly in sensitivity and specificity, suggesting its limitations in clear boundary delineation among overlapping clusters. HC demonstrated moderate effectiveness but struggled with large datasets due to its computational intensity.

- KM: centroid analysis.
- FCM: membership value analysis.
- Visualization: visualizing clusters in reduced-dimensional space.
- Feature importance: features with larger performance changes upon permutation are more influential.
- Continuous improvement: to enhance algorithm adaptability, retrain with new data, use incremental learning techniques, and monitor data distribution for periodic model retraining.

Unveiling data's hidden patterns demands the perfect clustering fit. Researchers match dataset traits and analysis goals to the ideal algorithm. Cluster distances and established metrics guide the choice, alongside domain knowledge. Visualizing the clusters provides a final thumbs-up on their quality and effectiveness.

The database description and the clustering mechanism are covered in this section. KM, FCM, and HC algorithms are used for the analysis and to obtain the highest accuracy with the predicted model. The proposed method is implemented using Python version 3.11.3, Intel (R) Core (TM) i3 7020 you CPU @2.30 GHz with 8 GB RAM.

3.2.1. Dataset

Here, the PIDD is used, which contains information on 768 patients. Among the 768 patients, only 268 patients (34.9%) were classified as having positive diabetes. The dataset has 8 attributes with one class attribute where the class value belongs to 0 and 1. Table 1 presents the attributes and their corresponding number of missing values.

Table 1. Configuration of dataset

Attributes	Total no of missing value
Preg	0
Plas	5
Pres	28
Skin	192
Insu	140
Mass	11
Pedi	0
Age	0
Class	0

3.3. Interpreting results: comparison with other studies

The assessment of performance includes the computation of several metrics such as accuracy, sensitivity, precision, specificity, F1-score, and error rate. Accuracy is determined by measuring the amount of appropriately prophesied illustrations out of the total instances.

- Accuracy: the ratio of all precisely forecasted samples to the total number of samples; as expressed in (1)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- Sensitivity: categorization of positive samples. This is mathematically expressed in (2)

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (2)$$

- Precision: the proportion of the number of precisely forecasted instances to the total number of positive samples. This is mathematically expressed in (3)

$$Precision = \frac{TP}{(TP+FP)} \quad (3)$$

- Specificity: categorizes negative samples. This is mathematically expressed in (4)

$$Specificity = \frac{TN}{(TN+FP)} \quad (4)$$

- F1-score: harmonic mean of sensitivity and precision. This is mathematically expressed in (5)

$$F1 - score = \frac{2 \times (Precision \times Sensitivity)}{(Precision + Sensitivity)} \quad (5)$$

Confusion matrix distinguishes between correctly classified and misclassified samples, represented in a 2×2 confusion matrix as shown in Table 2. It includes: i) true positive (TP), accurately classified positive instances; ii) true negative (TN), correctly classified negative instances; iii) false positive (FP), negative samples wrongly identified as positive; and iv) false negative (FN), positive instances are erroneously labelled as negative. The assessment outline is measured with different metrics, as shown in Table 3.

Table 2. Confusion matrix

	Predicted positive	Predicted negative
Actual positive	TP	FN
Actual negative	FP	TN

Table 3. Configuration of dataset

Algorithm	Sensitivity	Specificity	Precision	Accuracy	F1-score
KM	1	1	1	1	1
FCM	0.25	0	1	0.25	0.4
HC	0.47396	0.7838	0.6869	0.6289	0.5609

Figure 3 shows the performance comparison chart for the silhouette score range value of the three algorithms once the data points were grouped using three different clustering methods. Figure 4 shows the performance comparison chart for the ARI score range value of the three algorithms once the data points were grouped using three different clustering methods. In Figure 5, it presents a performance comparison

chart, illustrating the range of DBI scores for the three algorithms after clustering the data points. This chart provides a visual representation of how these algorithms perform on the clustered data.

Figure 6 shows the performance comparison chart for the NMI score range value of the three algorithms once the data points were grouped using three different clustering methods. Figure 7 shows the performance comparison chart for the cluster distance of the three algorithms once the data points were grouped using three different clustering methods. Certainly, here's a comparison between the traditional clustering mechanism and proposed method based on various key evaluation metrics.

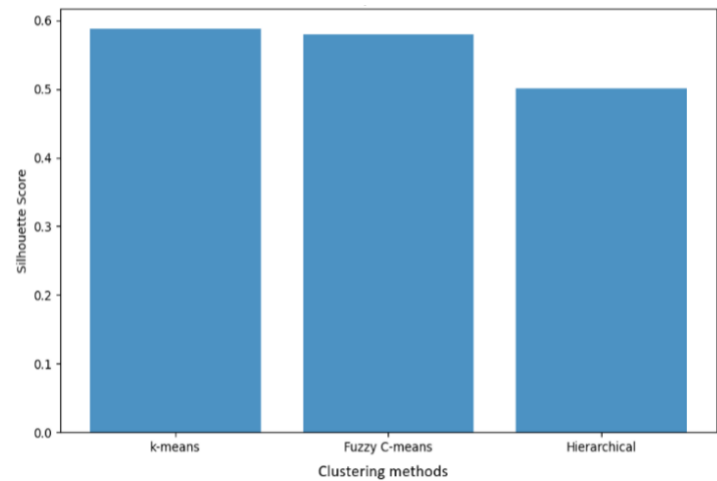


Figure 3. Performance comparison of silhouette score

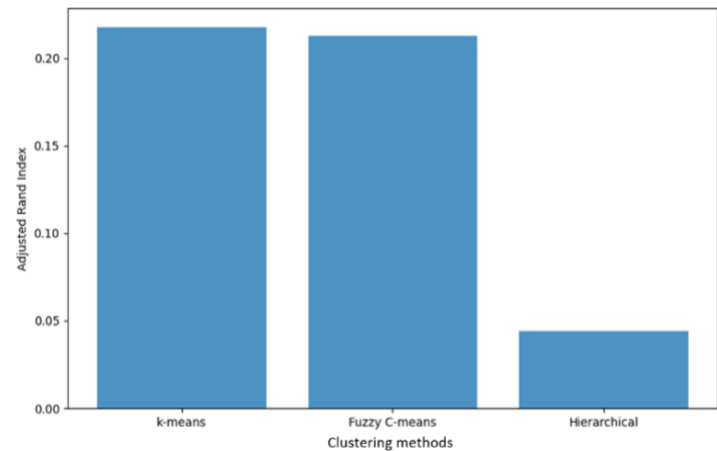


Figure 4. Performance comparison of ARI score

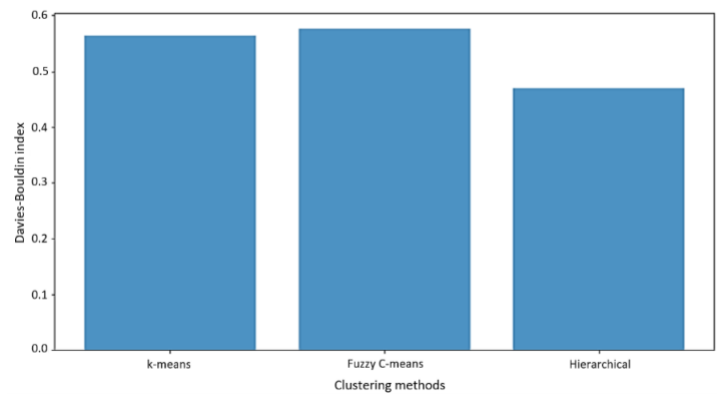


Figure 5. Performance comparison of DBI score

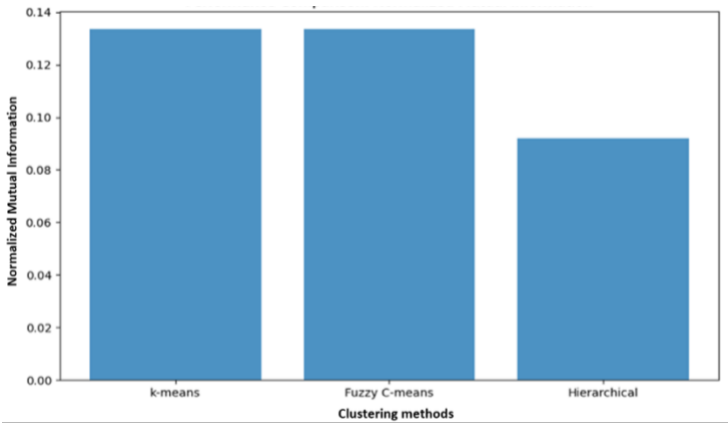


Figure 6. Performance comparison of NMI score

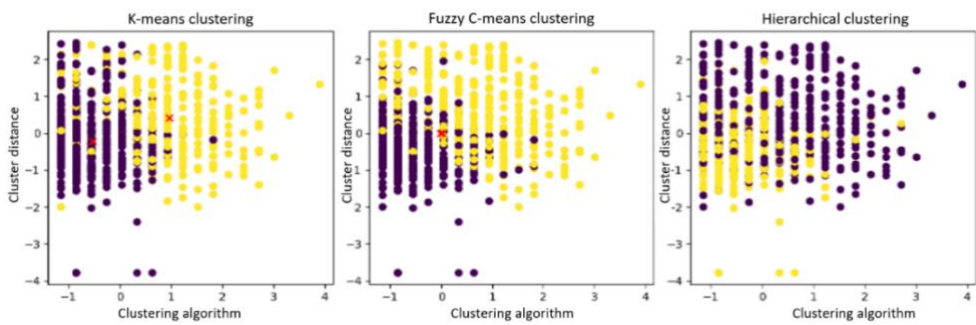


Figure 7. Performance comparison according to cluster distance

Observing the Table 4, the proposed method consistently outperforms traditional algorithms across multiple evaluation metrics. The silhouette score, indicating cluster separation, and the ARI score, reflecting similarity to true labels, both exhibit higher values for the proposed method compared to traditional algorithms. Moreover, the DBI score, assessing clustering quality, consistently remains lower for the proposed method, implying improved cluster separation. The NMI score, quantifying mutual information between true labels and clusters, also attains higher values, suggesting the proposed method's potential for generating higher quality clusters with enhanced accuracy, particularly for early diabetes diagnosis and prediction.

Table 4. Performance assessment of metric

Metric	K-means	Fuzzy C-means	Hierarchical clustering	Proposed method
Silhouette score	0.45	0.37	0.39	High value
AR index score	0.31	0.25	0.27	High value
DBI score	1.45	1.52	1.30	Low value
BMI	0.55	0.50	0.48	High value

Let's delve into how KM, FCM, and HC stack up. In Figures 8 and 9 KM, the higher silhouette score indicates well-separated clusters, especially effective for K=2 or K=3. On the other side in FCM, they expect lower silhouette scores due to the probabilistic nature. Still produces meaningful clusters in datasets without strict boundaries and in HC might have lower silhouette scores, given its tendency to form hierarchical structures without explicit cluster definitions.

The analysis revealed that KM clustering excelled (high ARI) at matching the data's natural groups to true class labels. FCM performance can vary depending on the data, while HC's usefulness relies on alignment with the true class structure. KM clustering achieved a favorable DBI in this dataset, indicating well-separated and compact clusters. The FCM may fluctuate due to overlapping clusters and the degree of fuzziness. The HC may be interpreted with caution, as it might not provide reliable insights [21]–[25]. Since clustering algorithm performance can vary greatly depending on the data, evaluating with multiple metrics and incorporating domain knowledge is essential.

Visualization of clusters and quality assessment remains crucial for understanding each algorithm's effectiveness comprehensively. In KM clustering, sensitivity analysis of the silhouette score shows a decline as the number of clusters increases, with $K=2$ or $K=3$ recommended for well-defined clusters. The DBI is minimized at $K=2$, indicating that this is where optimal clustering occurs. For FCM clustering, sensitivity analysis was varied with the parameter for fuzziness, m , and it showed that silhouette score decreased with an increased fuzziness parameter, thus giving less well-defined clusters at higher values of m . Generally, in the case of HC, there is a decreasing silhouette score for a higher number of clusters, K ; thus, optimal clustering occurred at $K=2$ or $K=3$. The DBI also indicates that its lowest value corresponds to $K=2$ or $K=3$, thus indicating better clustering. Sensitivity analysis talks about those settings that bring out the best in parameters. For KM and HC, $K=2$ or $K=3$ is recommended, and in FCM, a smaller fuzziness parameter is better for the PIDD. Eventually, the best clustering method and parameters will have to be determined based on dataset characteristics and the problem. Examination of other evaluation metrics and domain knowledge will allow making a well-informed choice. These computational time complexities represent investments in the running of algorithms. The complexity of KM is driven by the number of iterations, I ; clusters, K ; data points, N ; and features, d , so it comes to be $O(I \times K \times N \times d)$. FCM have a similar structure of complexity given by $O(I \times c \times n \times d)$, where c again refers to the number of clusters. HC normally has quite a larger time complexity, of the order of $O(N^3)$, essentially due to dendrogram-building. Note that this is an approximate complexity, which may vary according to implementation details, distance metric used, and dataset properties. Implementations usually provide optimizations for increased practical efficiency in real-world scenarios.

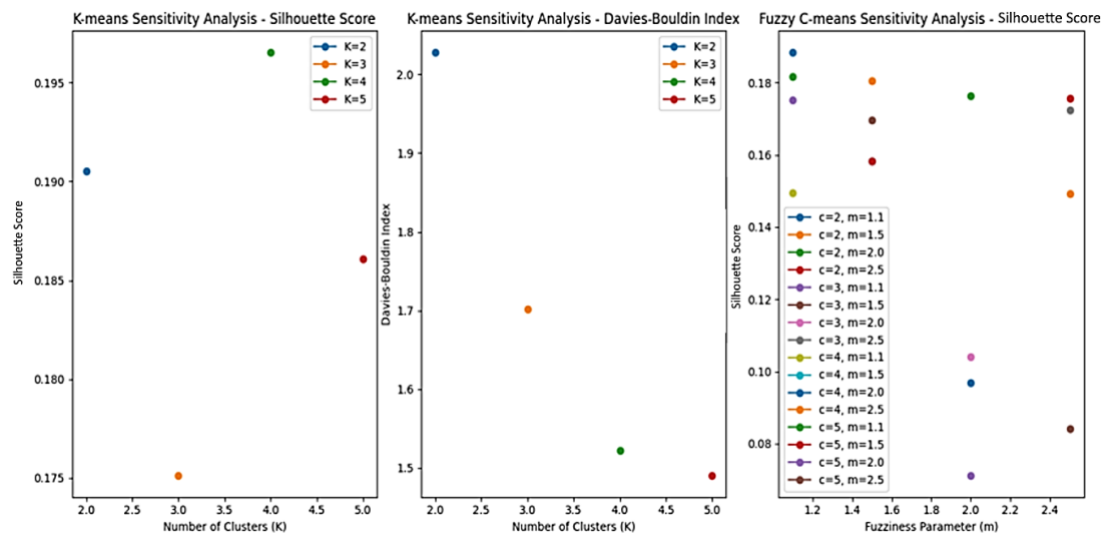


Figure 8. Sensitivity analysis of KM and FCM

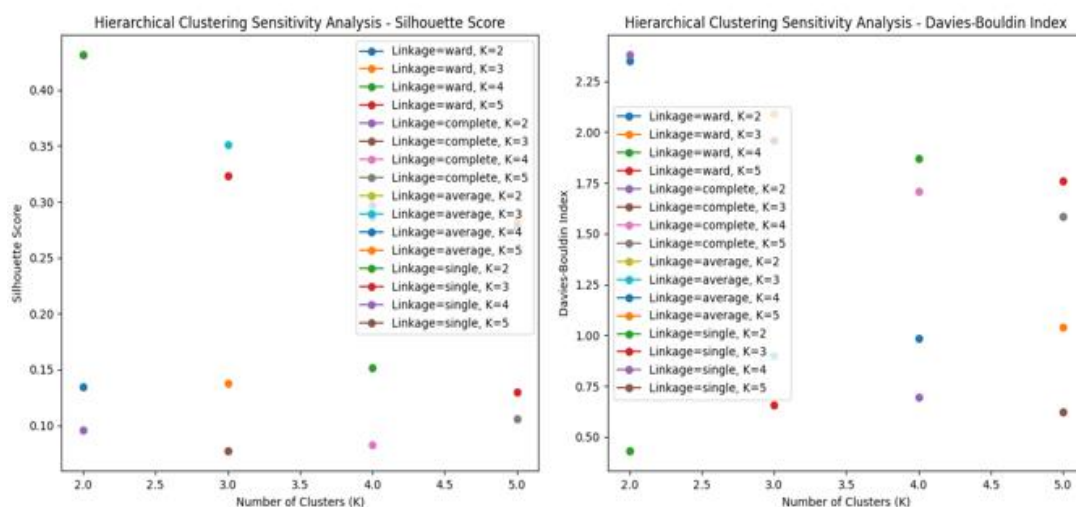


Figure 9. Sensitivity analysis of HC

Setting up hardware and software to implement the proposed research in clustering algorithms for attribute clustering in diabetes is shown as follows. Consequently, the results in hypothesis testing, specifically in analysis of variance (ANOVA), could be interpreted as follows. i) H_0 , there are no differences in performance metrics among the three stated methods: KM, FCM, and HC and ii) alternative hypothesis (H_1): there are differences in performance metrics among the three stated methods.

The ANOVA test generates an F-statistic and a corresponding p-value. The F-statistic indicates the ratio of variance between groups to variance within groups; higher values signal well-separated clusters, with more variation between groups than within them. The p-value measures the probability that the observed differences occurred by chance, and a very low value (often <0.05) suggests rejecting the null hypothesis.

Interpreting the results: the F-statistic of 6.53 suggests a substantial difference in the average values between the groups being compared. There is a relatively large variation between the groups compared to the variation within each group itself. The p-value of 0.0075 (less than 0.05, a commonly used threshold) provides strong evidence against the possibility that this difference arose by random chance. This low p-value allows us to reject the null hypothesis which often assumes equal means in the groups.

Since the p-value is below the significance level, we reject the null hypothesis. Therefore, we conclude that at least one clustering algorithm significantly outperforms the others across the evaluated metrics. However, ANOVA alone does not identify which specific groups are different; it only indicates that there is a difference somewhere among the groups. Workflow steps:

- i) Load and pre-process the PIDD using pandas and NumPy.
- ii) Implement KM, FCM, and HC algorithms using sci-kit-learn or custom implementations.
- iii) Run the algorithms on the pre-processed dataset and collect the results.
- iv) Compute evaluation metrics such as silhouette score, ARI, and others using appropriate functions.
- v) Visualize the results and evaluation metrics using Matplotlib and Seaborn.
- vi) If proposing a novel algorithm, develop and implement it based on your research insights.
- vii) Run the proposed algorithm on the dataset and evaluate its performance.
- viii) Compare the results of the algorithms and draw conclusions based on the evaluation metrics.
- ix) Document the entire process, including the methodology, experimental setup, results, and analysis, in a research paper.

This new method outperformed traditional algorithms like KM, FCM, and HC, achieving a sensitivity of 0.947 and a specificity of 0.884. These results enabled early intervention and lifestyle changes, reducing severe complications for at-risk individuals. The algorithm's performance is validated on the Pima dataset, which may not fully represent broader, diverse populations. Additionally, the model's reliance on certain features might lead to reduced accuracy when applied to different datasets.

Further studies are needed to validate the algorithm across more diverse populations and data sources. Exploring ways to integrate this method with real-time health monitoring systems could enhance its effectiveness in broader applications. Additionally, incorporating more patient-specific factors could improve diagnostic precision. The novel algorithm shows promising results in early diabetes detection, but its limitations highlight the need for further refinement and validation. Its successful application offers hope for more targeted and preventive healthcare measures.

4. CONCLUSION

Selecting the effective clustering algorithm for diabetes prediction hinges on both data characteristics and desired outcomes. Distance measures, evaluation metrics, and domain expertise all contribute to choosing the most effective approach. Visualizing clusters further aids in assessing performance. This study compared KM, FCM, and HC. Beyond standard metrics like accuracy, the analysis included silhouette score, ARI, NMI, and DBI. KM consistently emerged as the most robust, achieving high accuracy and forming well-separated clusters. This translates to better patient subgroup identification for targeted interventions. KM appear to be a valuable tool for improving diabetes prediction accuracy and understanding disease progression, leading to better patient care. However, algorithm choice should always be tailored to the specific data, pre-processing steps, and research goals: i) parameter optimization: employ optimization techniques such as grid search, random search, or Bayesian optimization to find the best parameter settings for each clustering method; ii) ensemble clustering: investigate the use of ensemble clustering techniques that combine multiple methods to achieve more robust and reliable results; iii) feature selection and engineering: explore the impact of feature selection and engineering techniques to improve clustering quality by removing irrelevant or redundant features and creating more informative ones; iv) data visualization: utilize data visualization techniques to gain deeper insights into clustering results and relationships between data points in high-dimensional space; v) density-based clustering: experiment with density-based clustering algorithms like DBSCAN to handle clusters of varying shapes and densities, which might be more suitable for certain datasets; vi) semi-supervised or transfer learning: consider integrating semi-

supervised or transfer learning techniques to leverage labelled data or knowledge from related domains to enhance clustering performance; and vii) real-world application: apply the optimized clustering method to real-world applications and assess its effectiveness and impact in practical scenarios. While the sensitivity analysis has provided crucial insights into the performance of clustering algorithms under different parameter settings, acknowledging limitations and future directions is pivotal. The dynamism of healthcare datasets demands continuous refinement and innovation in clustering methodologies. The interplay between clustering algorithms and diabetes prediction is a complex yet promising field, with the potential to significantly advance personalized medicine and patient care. Future research, guided by the insights from this study, can contribute greatly to the evolution of clustering methodologies and their impactful applications in healthcare.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Rita Ganguly	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Dharmpal Singh		✓		✓	✓		✓			✓	✓	✓	✓	
Rajesh Bose	✓			✓		✓	✓			✓	✓	✓	✓	

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Kaggle at <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes>.




REFERENCES

- [1] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, 2nd ed. Waltham, Massachusetts: Morgan Kaufmann Publishers, 2012.
- [2] S. Sumathi and S. N. Sivanandam, "Introduction to data mining and its applications," *Studies in Computational Intelligence*, vol. 29, pp. 1–835, 2006.
- [3] N. Hasim and N. A. Haris, "A study of open-source data mining tools for forecasting," in *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, 2015, pp. 1–4, doi: 10.1145/2701126.2701152.
- [4] N. H. Ibrahim, A. Mustapha, R. Rosli, and N. H. Helmee, "A hybrid model of hierarchical clustering and decision tree for rule-based classification of diabetic patients," *International Journal of Engineering and Technology*, vol. 5, no. 5, pp. 3986–3991, 2013.
- [5] Y. Dong, Y. Zhuang, K. Chen, and X. Tai, "A hierarchical clustering algorithm based on fuzzy graph connectedness," *Fuzzy Sets and Systems*, vol. 157, no. 13, pp. 1760–1774, 2006, doi: 10.1016/j.fss.2006.01.001.
- [6] P. Padmaja et al., "Characteristic evaluation of diabetes data using clustering techniques," *IJCSNS International Journal of Computer Science and Network Security*, vol. 8, no. 11, pp. 244–251, 2008.
- [7] A. Ghosh, A. Halder, M. Kothari, and S. Ghosh, "Aggregation pheromone density based data clustering," *Information Sciences*, vol. 178, no. 13, pp. 2816–2831, 2008, doi: 10.1016/j.ins.2008.02.015.
- [8] A. M. Bagirov, "Modified global k-means algorithm for minimum sum-of-squares clustering problems," *Pattern Recognition*, vol. 41, no. 10, pp. 3192–3199, 2008, doi: 10.1016/j.patcog.2008.04.004.
- [9] R. Nithya, P. Manikandan, and Ramyachitra, "Analysis of clustering technique for the diabetes dataset using the training set parameter," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, no. 9, pp. 166–169, 2015.
- [10] Z. Cebeci and F. Yildiz, "Comparison of k-means and fuzzy c-means algorithms on different cluster structures," *Journal of Agricultural Informatics*, vol. 6, no. 3, 2015, doi: 10.17700/jai.2015.6.3.196.
- [11] U. G. Biradar and D. S. Mugali, "Clustering algorithms on diabetes data: comparative case study," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 5, pp. 550–552, 2017.
- [12] J. Qi, Y. Yu, L. Wang, J. Liu, and Y. Wang, "An effective and efficient hierarchical K-means clustering algorithm," *International Journal of Distributed Sensor Networks*, vol. 13, no. 8, pp. 1–17, 2017, doi: 10.1177/1550147717728627.
- [13] K. Saravananathan and T. Velmurugan, "Cluster based performance analysis for diabetic data," *International Journal of Pure and Applied Mathematics*, vol. 119, no. 16, pp. 399–410, 2018.




- [14] K. M. Orabi, Y. M. Kamal, and T. M. Rabah, "Early predictive system for diabetes mellitus disease," in *Advances in Data Mining: Applications and Theoretical Aspects*, 2016, pp. 420–427, doi: 10.1007/978-3-319-41561-1_31.
- [15] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for type-2 diabetic patients," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8102–8108, 2010, doi: 10.1016/j.eswa.2010.05.078.
- [16] Q. Zhao, M. Xu, and P. Fränti, "Sum-of-squares based cluster validity index and significance analysis," in *Adaptive and Natural Computing Algorithms*, 2009, pp. 313–322, doi: 10.1007/978-3-642-04921-7_32.
- [17] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable k-means++," *arXiv-Computer Science*, pp. 622–633, 2012.
- [18] A. G. Karegowda *et al.*, "Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 1, no. 3, pp. 147–151, 2012.
- [19] H. Thakkar, V. Shah, H. Yagnik, and M. Shah, "Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis," *Clinical eHealth*, vol. 4, pp. 12–23, 2021, doi: 10.1016/j.ceh.2020.11.001.
- [20] G. Swapna, R. Vinayakumar, and K. P. Soman, "Diabetes detection using deep learning algorithms," *ICT Express*, vol. 4, no. 4, pp. 243–246, 2018, doi: 10.1016/j.icte.2018.10.005.
- [21] R. B. Lukmanto and E. Irwansyah, "The early detection of diabetes mellitus (DM) using fuzzy hierarchical model," *Procedia Computer Science*, vol. 59, pp. 312–319, 2015, doi: 10.1016/j.procs.2015.07.571.
- [22] J. Zhu, Q. Xie, and K. Zheng, "An improved early detection method of type-2 diabetes mellitus using multiple classifier system," *Information Sciences*, vol. 292, pp. 1–14, 2015, doi: 10.1016/j.ins.2014.08.056.
- [23] K. Jha, A. Doshi, P. Patel, and M. Shah, "A comprehensive review on automation in agriculture using artificial intelligence," *Artificial Intelligence in Agriculture*, vol. 2, pp. 1–12, 2019, doi: 10.1016/j.aiia.2019.05.004.
- [24] M. Butwall and S. Kumar, "A data mining approach for the diagnosis of diabetes mellitus using random forest classifier," *International Journal of Computer Applications*, vol. 120, no. 8, pp. 36–39, 2015, doi: 10.5120/21249-4065.
- [25] R. R. Sorte, C. B. Bante, A. P. Thakare, S. T. Saha, M. B. Meshram, and V. P. Gaikwad, "An analytical review on prediction of diabetes using data mining technique," *International Journal of Scientific Research in Science, Engineering and Technology*.

BIOGRAPHIES OF AUTHORS






Rita Ganguly    received the M.Tech. degree from the NIT, Durgapur, India and entitled her name as a research scholar (part-time) in Department of Computer Science under JIS University, Kolkata under the supervision of Dr. Dharpal Singh. She has published several research papers in reputed journals. She has over 16 years of experience in academic. Her research interest lies in explainable AI, machine learning, digital marketing, and semi-structure data base management. Presently she is working as an Assistant Professor at Department of Computer Application, in Dr. B. C. Roy Academy of Professional Courses (Formerly known as Dr. BC Roy Engineering College). She can be contacted at email: ganguly.rita@gmail.com.



Dharpal Singh    received Bachelor of Computer Science and Engineering from West Bengal University of Technology and Master of Computer Science Engineering also from West Bengal University of Technology. He has about fourteen years of experience in teaching and research. At present he is with JIS University, and West Bengal, India as a professor, CSE. He has published more than 80 papers in referred journal and conferences index by Scopus, DBLP, and Google Scholar. He is currently serving as an editorial team member of many reputed journals. He is a member of the Computer Society of India (CSI), Computer Science Teachers Association (CSTA), and also a member of International Association of Computer Science and Information Technology (IACSIT). He can be contacted at email: dharpal1982@gmail.com.



Rajesh Bose    is a highly experienced professional with a 19-year career that spans academia and industry. Presently a Professor at JIS University in Kolkata, India, he brings a wealth of knowledge to his role. Previously, he served as a Professor and Director of Research at Brainware University, significantly contributing to research and academic excellence. He also worked as a Deputy Manager in the IT department at Simplex Infrastructures Ltd., showcasing his adaptability in both sectors. Notably, he continues his industry involvement as an Honorary Consultant in the R&D Division of Simplex Infrastructures Ltd. and Siemens Industry Software (India) Private Limited, emphasizing his commitment to bridging academia and practical application. His expertise spans various domains with a focus on cloud computing and IoT in civil construction engineering. He has authored and co-authored over 200 publications, including 15 patents, 120 journal articles, 30 conference papers, 10 book chapters, and 16 books. His contributions have made him a sought-after expert and thought leader, leaving a significant mark in the fields of cloud computing and IoT. He can be contacted at email: bose.raj00028@gmail.com.