CITE PUBLICATIONS

Journal of Analytical and Applied Computer Sciences
Volume: 2, Issue: 2, 2026 (July - December), Pages: 38-49
ISSN (Online): 3139-3349 | DOI: https://doi.org/10.67914/jaacs/rsa/2.2/38-49
Research Article

Classification of metabolic patterns associated with insulin resistance using Fuzzy C-Means and K-Means algorithms: A comparative applied study

Ahmed Saif Al-Din Mahmoud, Noor Nawzat A. Seyala*

Department of Statistics and Informatics, Faculty of Computer Science and Mathematics, University of Mosul, Mosul, Iraq.

***Correspondence**
Noor Nawzat A. Seyala
nooalior@uomosul.edu.iq

Creative Commons Attribution–NonCommercial–NoDerivatives 4.0 International (CC BY-NC-ND 4.0).

Received: 21 August 2026 / Accepted: 20 October 2026 / Published: 11 November 2026

Insulin resistance is a progressive metabolic disorder, and distinguishing between individuals can be challenging due to overlap in their metabolic characteristics, making categorical classification unsuitable for some cases. This research aims to employ the Fuzzy C-Means (FCM) clustering algorithm to group individuals according to metabolic indicators associated with insulin resistance, and to compare this with the K-Means algorithm. The study relied on health data obtained from specialized endocrinology and diabetes centers in Iraq, including indicators related to obesity and insulin resistance, with the triglyceride-to-high-density lipoprotein (TG/HDL) ratio used as a surrogate indicator associated with insulin resistance. Both algorithms were applied to form three clusters, and the quality of the clustering was evaluated using the partitioning coefficient (PC), partitioning entropy (PE), the Silhouette coefficient, the Davies-Bouldin index (DBI), and the Calinski-Harabasz index (CHI). The FCM results showed values of 0.3772 for the PC coefficient, 1.0307 for the PE coefficient, 0.2497 for the Silhouette index, 1.8875 for the DBI index, and 196.9645 for the CHI index. In contrast, K-Means achieved better values in the cluster quality indices, with a Silhouette coefficient of 0.4176, a DBI index of 1.3160, and a CHI index of 222.6734. Nevertheless, FCM allows for the expression of an individual's degree of belonging to different clusters, which helps in representing overlapping and transitional states. The results indicate that FCM is a suitable tool for studying metabolic patterns associated with insulin resistance, especially when the goal is not only to separate clusters but also to reveal the degree of individual belonging to them and represent overlap between them. This makes it more suitable for classifying health data of an overlapping nature compared to traditional methods based on acute affiliation.

Keywords: *Cluster analysis, Fuzzy C-Means, K-Means, Insulin resistance*

Introduction

Recent years have witnessed a significant increase in the volume of available data across various fields, particularly in healthcare. This has led to a need for statistical and computational methods capable of analyzing this data and uncovering underlying patterns and relationships. Data mining has thus become a crucial field for extracting knowledge from large, multivariable datasets, and cluster analysis is one of its fundamental techniques. Cluster analysis aims to divide observations into groups where observations within a group are more similar, while different groups are more diverse, without requiring prior classification of the observations. The clustering employed by rigid algorithms may not be suitable when clusters overlap or their boundaries are unclear. Some observations may be similar to more than one group

simultaneously. This problem is clearly evident in many real-world applications, including health and vital data, which are often characterized by overlap and a gradation in the characteristics of the observations. Not every observation needs to be clearly associated with a single cluster, and forcing it into one cluster may lead to the loss of information regarding its proximity to other clusters. This is where the concept of fuzzy logic emerged, its theoretical foundations laid by Zadeh (1965). This logic provided a mathematical framework that allows for dealing with partial membership degrees between 0 and 1, rather than the absolute binary (yes or no) membership. Building on this concept, Dunn in (1973) laid the groundwork for fuzzy clustering, and Bezdek later developed the Fuzzy C-Means (FCM) algorithm, which has become one of the most widely used fuzzy clustering algorithms.

Research problem

The research problem lies in the difficulty of collecting health data related to insulin resistance due to the potential overlap and similarity of metabolic characteristics among individuals. This may limit the ability of rigid clustering methods, such as the K-Means algorithm, to represent this overlap.

Research objective:

This research aims to employ the Fuzzy Means (FCM) algorithm to cluster individuals according to a set of metabolic indicators associated with insulin resistance. (FCM) provides organic scores that help clarify the extent to which observations belong to different clusters. The results are then compared with K-Means results using a set of clustering validity indicators to determine the suitability of each algorithm in identifying metabolic patterns in the studied data.

Lecturer review

The initial foundations of fuzzy clustering were presented by Dunn in (1973) in his study titled " A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters". His contribution lay in developing a concept that allowed for the representation of the gradual belonging of elements to clusters, rather than imposing a definitive assignment for each element. Then, in (1981), Bezdek developed the mathematical basis for fuzzy objective function algorithms, including the Fuzzy C-Means (FCM) algorithm, which later became one of the most widely used fuzzy clustering algorithms. Numerous theoretical and applied studies have been conducted in this field, in addition to comparisons between fuzzy clustering and solid clustering algorithms. In (2013), Ghosh and Dubey conducted a study titled "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms," comparing the two algorithms by varying the number of observations and clusters. They found that (FCM) yielded results close to those of K-Means but required a longer computation time. In (2020), Wiharto and Suryani conducted a study titled " The comparison of clustering algorithms K-means and fuzzy C-means for segmentation retinal blood vessels " comparing the two algorithms on DRIVE and STEARE data. They found that (FCM) significantly outperformed K-Means and had a higher Area Under Curve (AUC), but it required a longer computation time. In (2023), Nagase and colleagues conducted a study titled " Serum fatty acid composition balance by fuzzy C-means method in individuals with or without metabolic dysfunction-associated fatty liver disease", using the (FCM) algorithm to analyze serum fatty acid composition balance. Blood, and (FCM) was a suitable choice due to its ability to handle overlapping and fuzzy data. In (2025), Bora and Mishra conducted a study entitled " Comparative evaluation of hard and soft clustering for precise brain tumor segmentation in MR imaging " in which they compared solid and fuzzy clustering in segmenting brain tumors from MRI images. The result was clearly in favor of (FCM), as it identified the boundaries of irregular and overlapping tumors with higher accuracy, and its Dice similarity coefficient was 0.67 compared to 0.43 for the K-Means algorithm.

Theoretical framework:

Cluster analysis

Cluster analysis is a key technique in multivariate statistics and data analysis, and one of the most prominent unsupervised learning techniques. This analysis aims to discover patterns and clusters in data by dividing observations or individual data into relatively homogeneous groups called clusters. This division is based on the principle that observations belonging to the same cluster are more similar to each other, while observations belonging to different clusters differ to varying degrees depending on the nature of the data, the algorithm, and the similarity or distance measure used (Jaeger & Banks, 2023). Cluster analysis is not limited to a single algorithm; rather, it encompasses a variety of methods that differ according to the nature of the data, the similarity or dissimilarity measure used, the shape of the clusters, and the mechanism for their formation. Among the most prominent of these methods are partitioning, hierarchical clustering, density-based clustering, and model-based clustering. The choice of the appropriate method depends on the nature of the research problem, the characteristics of the data, and the objective of the clustering process (Singh & Singh, 2024).

k-means algorithm:

The k-means algorithm is one of the fundamental partitioning clustering tools in unsupervised learning. It aims to distribute numerical observations across (k) distinct clusters. The algorithm is based on the idea of representing each cluster with a geometric center that represents the arithmetic mean of its points. Its mechanism seeks to minimize the sum of the squares of the distances between an observation and the center of its cluster, thus ensuring the proximity of elements within the cluster and their distance from other clusters (MacQueen, 1967; Jain, 2010). The algorithm assumes a dataset containing n observations and p variables. Each observation, x_i, can be represented as a mathematical vector: $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$. The K-means algorithm essentially aims to partition these observations into a specific number of clusters, k i.e., $(c1, c2, \dots, c_k)$. The goal is to make the observations within each cluster as close together as possible. To achieve this, the algorithm minimizes the sum of squared distances within the clusters (WCSS) (MacQueen, 1967; Jain, 2010).

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 \quad (1)$$

Where μ_j represents the center point of cluster j. This center is nothing but the arithmetic average of all observations belonging to the same cluster, and is calculated as follows:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \quad (2)$$

As for measuring how close or far the observer x_i is from the center μ_j, the most common method is to rely on the Euclidean distance:

$$\boldsymbol{d(x_i, \mu_j) = \sqrt{\sum_{r=1}^{p} (x_{ir} - \mu_{jr})^2}} \quad (3)$$

Based on this distance, the algorithm assigns each observation to the cluster with the nearest center, according to the rule:

$$\boldsymbol{C(x_i) = argmin_j\, d(x_i, \mu_j)} \quad (4)$$

This process is not a single step but rather iterative, with the algorithm repeatedly redistributing observations and updating the positions of the centers. The algorithm stops and stabilizes when it reaches a state of convergence, which occurs when the change in the positions of the centers or in the assignment process becomes very small (less than a certain threshold value (ε)), or when the maximum number of iterations allowed is reached (Hartigan & Wong, 1979; Jain, 2010).

Fuzzy clustering

Fuzzy clustering is an unsupervised learning technique that aims to divide observations into clusters while allowing partial membership in more than one cluster. The Fuzzy C-Means (FCM) algorithm is one of the most prominent algorithms used in this field. It iteratively estimates cluster centers and updates observation membership scores to minimize the objective function, which depends on the distances between observations and cluster centers. Membership scores are assigned weights that control their impact on the clustering process. This algorithm is used in multivariable data analysis when representing the interaction between clusters is important for understanding data structure and interpreting results (Krasnov et al., 2023; Hamini et al., 2025)

Fuzzy C-Means clustering (FCM)

The Fuzzy C-Means (FCM) algorithm is one of the most prominent unsupervised clustering techniques. It is used to divide a dataset into a predetermined number of clusters based on the degree of similarity between individual elements and cluster centers. FCM differs from the traditional K-Means algorithm in that it does not force an element to belong to a single cluster but allows it to belong to multiple clusters simultaneously, with membership scores ranging from 0 to 1. Therefore, FCM is considered a soft fuzzy clustering technique, where the membership score indicates the degree to which an observation belongs to each cluster (Bezdek, 1981). The Fuzzy C-Means Clustering (FCM) algorithm relies on the principle of minimizing the quadratic objective function, where the terms are weighted by membership scores and is defined by the following formula:

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} \|x_i - v_j\|^2 \quad (5)$$

Where c represents the number of clusters, (n) the number of observations, u_{ij} the degree of observer membership x_i in cluster i, v_j represents the center of cluster i, and m is the fuzzy coefficient, while $\| x_i - v_j \|^2$ squared distance between the observation and the center of the cluster. The algorithm aims to find the membership values and cluster centers that yield the lowest possible value for the objective function (Krasnov et al., 2023; Hashemi et al., 2023). The membership matrix $U = [u_{ij}]$ is subject to a set of constraints that ensure the integrity of the clustering process, most notably $0 \leq u_{ij} \leq 1$ and $\sum_{i=1}^{c} u_{ij} = 1$ where j = 1, 2, ..., n. Additionally, each cluster must have a non-zero membership in the dataset, $0 < \sum_{j=1}^{c} u_{ij} < n$, where i = 1, 2, ..., c. Under these constraints, the membership values must be within the interval [0, 1], and the sum of the membership scores for each observation across all clusters must equal one. The formation of empty clusters is prohibited. The (FCM) algorithm operates iteratively to reach the cluster centers and observation membership scores, where the number of clusters (c), the fuzzy coefficient (m), and the convergence threshold are specified. $U^{(0)}$ then initialize the membership matrix ϵ randomly such that $\sum_{i=1}^{c} u_{ij} = 1$, then determine the center of each cluster using the following formula:

$$v_i^{(k)} = \frac{\sum_{j=1}^{n} \left(u_{ij}^{(k)}\right)^m x_j}{\sum_{j=1}^{n} \left(u_{ij}^{(k)}\right)^m} \qquad i = 1, \ldots, c \quad (6)$$

Then calculate the Euclidean distance between each x_j observation and the cluster center v_i, and then update the membership scores using the following relationship:

$$u_{ij}^{(k+1)} = \left[\sum_{l=1}^{c} \left(\frac{d_{ij}}{d_{lj}}\right)^{\frac{2}{m-1}}\right]^{-1} \quad (7)$$

The membership score falls within the interval [0,1] such that the sum of the membership scores for each observation across all clusters equals one. The change in the membership matrix is calculated as follows:

$$\| \mathrm{U}^{(k+1)} - \mathrm{U}^{(k)} \| < \varepsilon \quad (8)$$

Therefore, if the change in the membership matrix meets the predefined convergence condition, the algorithm is stopped. If it does not meet it, $U^{(k)}$ is replaced with $U^{(k+1)}$ and the steps are repeated starting from the calculation of the cluster centers. This iterative procedure continues until the convergence condition is met or the maximum number of iterations is reached.

Cluster validity indices

Cluster validity indices are crucial for quantitatively measuring the quality of results and enabling comparisons between the performance of different algorithms when applied to the same data. In this study, these indices are divided into two tracks: general indices applied to hard clustering outputs and fuzzy outputs after hard clustering conversion, and exclusive fuzzy indices calculated directly from the Membership Matrix to evaluate the (FCM) algorithm without requiring hard clustering.

Within-cluster sum of squares (WCSS)

This indicator reflects the sum of the squares of the Euclidean distances between each data point and the cluster center to which it is attributed. Its structure constitutes the objective function, which the K-Means algorithm iteratively minimizes. The lower the value of the index, the higher the degree of internal coherence, but it automatically decreases as the number of clusters (c) increases; therefore, its use is limited to comparing solutions based on the same number of clusters:

$$WCSS = \sum_{i=1}^{c} \sum_{x_j \in C_i} \|x_j - c_i\|^2 \quad (9)$$

Silhouette score

This measure the degree of consistency of an element within its cluster compared to its distance from the nearest neighboring cluster, and is calculated for each element *i* using the formula:

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)} \tag{10}$$

Where $a(i)$ represents the average distance between element *i* and the other elements of the same cluster (a measure of cohesion), while $b(i)$ represents the lowest average distance between it and elements of neighboring clusters (a measure of separation). Index values are set between -1 and 1, with higher values indicating greater homogeneity and clearer separation between clusters (Rousseeuw, 1987).

Davis-Bouldin Index (DBI)

This index calculates the average of the highest similarity between each cluster and its nearest corresponding cluster, where similarity is defined as a function relating the internal dispersion of clusters to the spacing between their centers.

$$DBI = \frac{1}{c}\sum_{i=1}^{c} \max_{j \neq i}\left(\frac{\sigma_i + \sigma_j}{d\big(c_i, c_j\big)}\right) \tag{11}$$

Where σ_i and σ_j represent the average internal dispersion of clusters *i* and *j*, while $d\big(c_i, c_j\big)$ represents the distance between their centers. Lower values of the index reflect better cluster quality, representing more cohesive and widely spaced clusters (Davies & Bouldin, 1979).

The Kalinski -Harabasz Index (CHI) Known as the Variance Ratio Criterion, it measures the ratio of intercluster dispersion to internal dispersion divided by the corresponding degrees of freedom.

$$CHI = \frac{SS_B}{c-1} \Big/ \frac{SS_W}{n-c} \tag{12}$$

Where SS_B represents the sum of squared variances between clusters, SS_W the sum of squared variances within clusters, (n) the sample size, and (c) the number of clusters. A higher value indicates more efficient clustering and clearer separation (Caliński & Harabasz, 1974).

Special fuzzy indices (PC, PE)

Unlike general indices, this group relies directly on the Fuzzy Membership Matrix generated by the FCM algorithm without requiring hard classification, to assess the clarity and overlap within the fuzzy structure itself.

Partition coefficient (PC)

Developed by Bezdek (1974) to measure the "sharpness" of a fuzzy distribution based on the squared membership scores:

$$PC = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \tag{13}$$

Where u_{ij} represents the degree to which observation j belongs to cluster *i*, n is the number of observations, and *c* is the number of clusters. The index ranges between *1/c* and 1; a value close to 1 indicates a sharp classification approaching solid clustering, while a value close to *1/c* reflects wide fuzzy overlaps, a behavior expected when processing complex medical data (Bezdek, 1984).

Partition entropy (PE) measures the degree of uncertainty (entropy) in the distribution of affiliations across clusters, and was based by Bezdek (1981) on Shannon's concept of entropy.

$$PE = -\left(\frac{1}{n}\right)\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\log(u_{ij}) \quad (14)$$

The index ranges between 0 and $log(c)$; lower values indicate clearer taxonomic boundaries, while higher values indicate greater overlap between clusters. It is used in conjunction with the PC coefficient to provide a comprehensive reading of the level of ambiguity in the classification (Bezdek, 1981).

Practical aspect:

Data description

To study the collection of health data related to insulin resistance, a local Iraqi database from Baghdad's Medical City Hospital and Al-Kindi Specialized Center for Endocrinology and Diabetes, specifically the Diabetes Dataset for (2023), was used via the Mendeley Data platform [1]. The data were verified to be free of missing values for all variables. Statistical analysis was then performed using MATLAB (version R2026a), employing the Fuzzy Logic Toolbox to implement the Fuzzy C-Means algorithm and the Statistics and Machine Learning Toolbox to implement the K-Means algorithm. The dataset comprises (1000) observations (patients) and nine variables relevant to the research topic. Variables irrelevant to the study were excluded, as detailed in Table (1).

Table 1. Data description

Variable	Code Variable Description
x_1	Sex (1 = Male, 2 = Female)
x_2	Age (years)
x_3	Glycated Hemoglobin (HbA1c) (%)
x_4	Total Cholesterol
x_5	Triglycerides
x_6	HDL (Good) Cholesterol
x_7	LDL (Bad) Cholesterol
x_8	VLDL (Very Low-Density Lipoprotein) Cholesterol

Data preprocessing

To detect strong correlations between variables that might indicate information duplication or multi-collinearity, a correlation matrix was calculated for all quantitative variables, which could affect the clustering results. The results showed that the correlations between the variables were weak to moderate (less than 0.42), as shown in Table 2. This is a positive finding because it supports the independence of each variable from the others, meaning that each variable provides different and useful information for the analysis.

Table 2. Correlation matrix between variables

	AGE	**HbA1c**	**Chol**	**TG**	**HDL**	**LDL**	**VLDL**	**BMI**
AGE	**1.00**	0.38	0.04	0.15	-0.02	0.02	-0.09	0.38
HbA1c	0.38	**1.00**	0.18	0.22	0.03	0.01	0.07	0.41
Chol	0.04	0.18	**1.00**	0.32	0.10	0.42	0.08	0.01
TG	0.15	0.22	0.32	**1.00**	-0.08	0.02	0.14	0.11
HDL	-0.02	0.03	0.10	-0.08	**1.00**	-0.14	-0.06	0.07
LDL	0.02	0.01	0.42	0.02	-0.14	**1.00**	0.06	-0.07
VLDL	-0.09	0.07	0.08	0.14	-0.06	0.06	**1.00**	0.20
BMI	0.38	0.41	0.01	0.11	0.07	-0.07	0.20	**1.00**

Alternative insulin resistance index

According to academic practices and recent medical studies, the triglyceride/HDL cholesterol ratio (TG/HDL ratio) has been used as a globally accepted alternative for assessing insulin resistance and metabolic sensitivity. This is due to the unavailability of direct fasting insulin testing, without which the HOMA-IR index cannot be calculated using the traditional method. This alternative index is well-known in the medical literature as it reliably reflects insulin resistance status, particularly in epidemiological and clinical studies dealing with large patient groups. The TG/HDL ratio

demonstrated excellent predictive power for insulin resistance in an Iraqi study conducted on a sample of healthy adults, which supports the adoption of this index as a practical alternative within the context of the Iraqi health data under study.

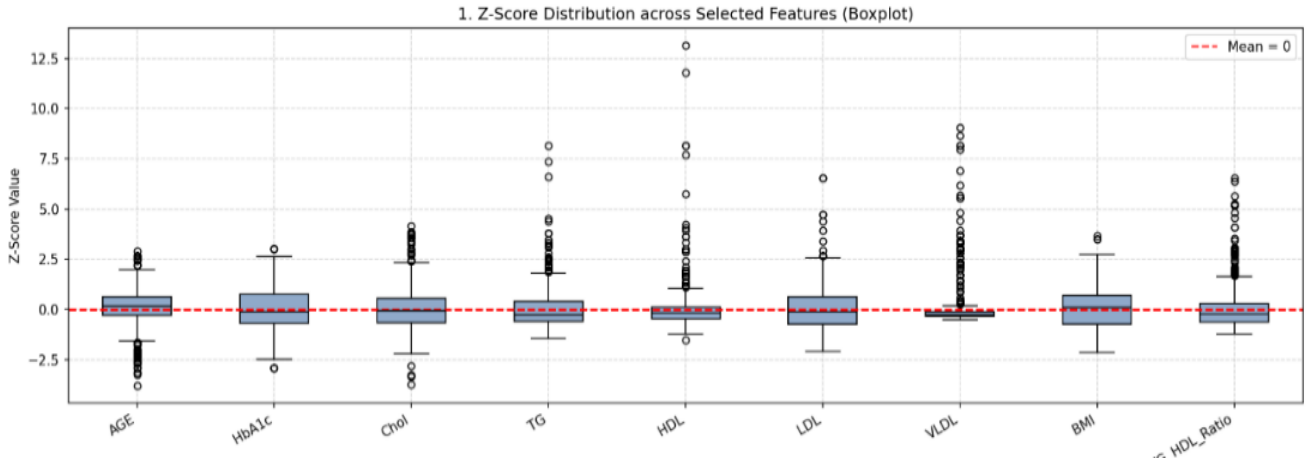

Figure 1. Box plot of variables in their standard form

Figure (1) shows box plots of the values for the variables (for the selected metabolic traits) after standardization to place all variables on a uniform scale, thus reducing the impact of variations in measurement units and numerical ranges during the data collection process. This reveals variations in the dispersion and distribution patterns of the selected traits. While central values generally cluster around zero, some variables exhibit wider interquartile ranges, indicating greater inter-individual variability. Furthermore, several variables show notable outliers, particularly in the positive direction, suggesting the presence of individuals with relatively high metabolic measurements. These observations reflect a degree of heterogeneity in the metabolic profiles of the studied individuals. This variability, coupled with the presence of overlapping metabolic characteristics, highlights the complexity of representing individuals within clearly defined, separate groups. Consequently, the use of fuzzy clustering (FCM) is a suitable analytical framework, as it allows each individual to belong to multiple groups to varying degrees, rather than being exclusively classified within a single group.

Data collection (data clustering)

Using the Fuzzy C-Means (FCM) algorithm

The FCM algorithm was applied with a fuzzy coefficient ($m = 2$) and a number of clusters ($c = 3$) in accordance with the applied objective of the study and the common clinical classification of risk levels (low, medium, high). The results are shown in Table (3).

Table 3. Cluster centers using the FCM algorithm

Cluster	AGE	HbA1c	Chol	LDL	VLDL	BMI	TG\HDL Ratio
Cluster 1	0.253	0.271	0.031	-0.026	-0.010	0.263	0.054
Cluster 2	-0.686	-0.711	-0.084	0.068	-0.057	-0.743	-0.185
Cluster 3	0.252	0.267	0.016	-0.043	-0.012	0.264	0.047

Table (3) shows the final positions of the three clusters at the convergence of the algorithm. These values represent the standard means of the variables in each cluster, allowing for an effective comparison of the relative trend of the variables. The results showed a significant difference in the central characteristics of the clusters. Cluster 2 clearly diverged from clusters 1 and 3, particularly in the variables of age (AGE), glycated hemoglobin (HbA1c), and body mass index (BMI). The center values of cluster 2 were lower than the overall sample mean, registering negative standard values of -0.68580, -0.71053, and -0.74284, respectively. In contrast, clusters 1 and 3 showed a high degree of similarity in their central behavior, with positive and closely related standard values ranging between 0.25 and 0.27.

Table 4. Cluster sizes and average highest degree of belonging

Cluster	Number of cases	Percentage %	Average highest degree of affiliation	Cases with a affiliation of 0.5 or more
First	342	34.2	0.3941	0
Second	274	27.4	0.5377	164
Third	384	38.4	0.3997	0
Total	**1000**	**100**	**0.4356**	**164**

After assigning each observation to the cluster with the highest membership score, the cases were distributed across the three clusters with a clear disparity, as shown in Table (4). The third cluster contained the largest number of cases, with 384 (38.4%), followed by the first cluster with 342 (34.2%), while the second cluster recorded only 274 (27.4%). This decisive outcome is due to the adoption of the highest value, and because the FCM algorithm essentially relies on assigning multiple fuzzy membership scores to each element before assignment. Figure (2) illustrates the distribution of observations across the three clusters, as shown.

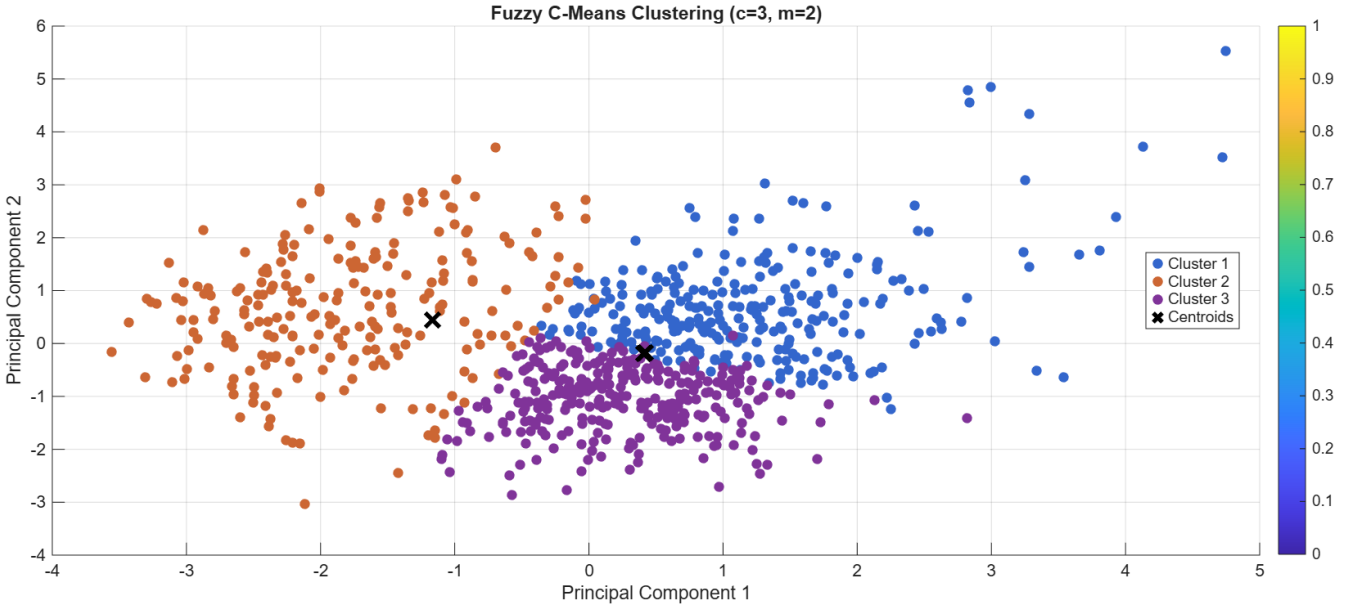

Figure 2. The distribution of observations across the three clusters is shown using the FCM algorithm

We observe overlapping transition zones between clusters, a phenomenon consistent with the fundamental concept of fuzzy clustering (FCM), which allows for the distribution of membership scores for a single observation in varying proportions across multiple clusters.

The cluster validity indices, as shown in Table (5), exhibited a relatively high degree of fuzzy affiliation, along with limited geometric differentiation between clusters.

Table 5. Assembly quality indicators

Indicator	**Value**	**Interpretation**
Partition Coefficient (PC)	0.3772	Relatively high degree of overlap
Partition Entropy (PE)	1.0306	Relatively high fuzziness
Silhouette Coefficient	0.2497	Weak to moderate cohesion
Davies–Bouldin Index (DBI)	1.8875	Limited separation between cluster centers
Calinski–Harabasz Index (CHI)	196.9645	Acceptable ratio of between-cluster to within-cluster variance

The validity indicators of fuzzy clustering support this observation; the partitioning coefficient (PC) reached 0.3772, while the partitioning entropy (PE) reached 1.0306, indicating the spread of membership scores among clusters and the high degree of fuzzy affiliation. Principal component representation also revealed overlapping regions between observations, particularly between clusters 1 and 3. Therefore, the FCM results reveal the importance of representing data with flexible membership scores, especially when the boundaries between different patterns are not sufficiently clear. This provides a more detailed description of the nature of observation affiliation compared to categorical assignment.

Using the K-Means algorithm

Table 6. Cluster sizes resulting from the K-Means algorithm

Cluster	Number of Cases	Percentage (%)
First	**50**	**5.0**
Second	**234**	**23.4**
Third	**716**	**71.6**
Total	**1000**	**100.0**

The data were analyzed using the K-Means algorithm based on three clusters (k=3). This algorithm assigns observations based on the principle of categorical assignment to the nearest center according to Euclidean distance, aiming to form clusters with the highest degree of internal homogeneity while achieving significant differentiation between different clusters. The results of applying the K-Means algorithm showed an uneven distribution of observations among the three clusters. Cluster 3 had the largest share with 716 observations (71.6% of the total sample), followed by cluster 2 with 234 observations (23.4%), while cluster 1 had only 50 observations (5.0%). Table 6 details this distribution, including the number and percentage of observations among the resulting clusters. The clustering results were visually represented to show the spatial distribution and visual interpretation of the resulting observations and clusters, as in Figure (3).

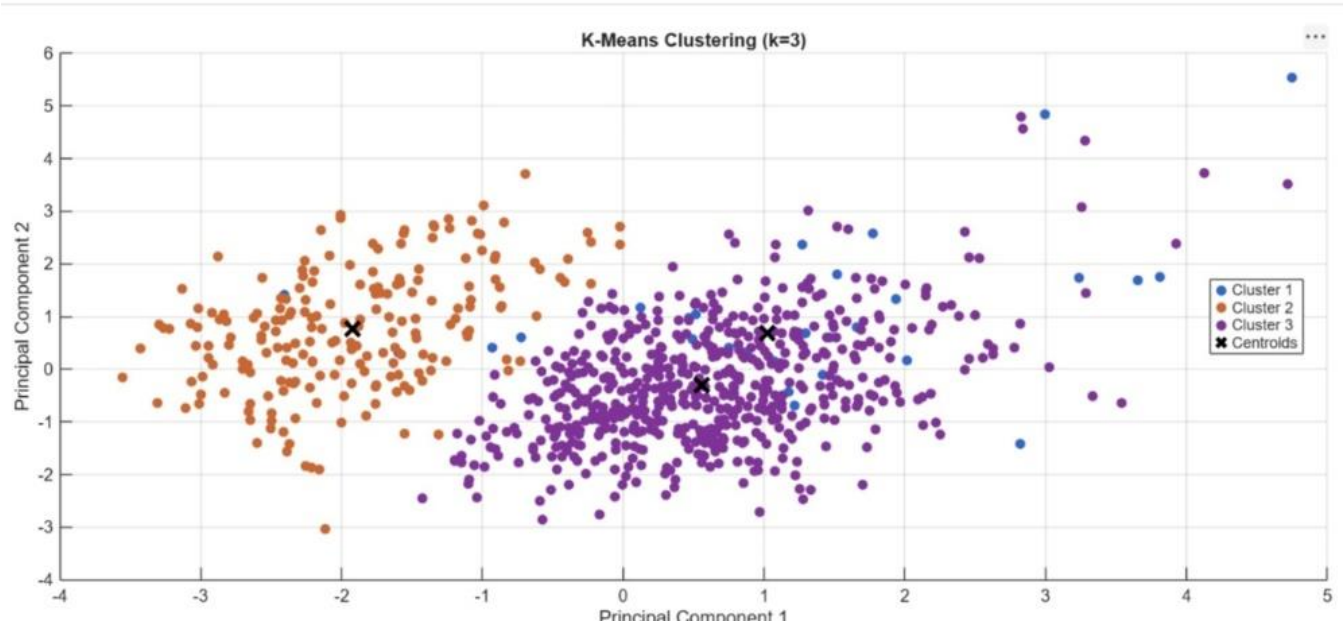

Figure 3. The distribution of observations across the three clusters is shown using the K-Means algorithm

As can be seen from the figure, the cluster centers are marked with an (X), reflecting their central positions relative to the observations within each cluster. There is some overlap between observations, particularly in the central region, indicating that the boundaries between some clusters are not perfectly sharp. This is consistent with the nature of insulin resistance data, as metabolic variables do not necessarily change independently between specific groups; rather, there may be cases with shared or similar characteristics across multiple metabolic profiles.

Table 7. K-Means algorithm cluster centers with standard values

clusters	AGE	HbA1c	Chol	LDL	VLDL	BMI	TG\HDL_Ratio
Cluster 1	-0.612	0.150	0.066	0.343	3.841	0.896	-0.015
Cluster 2	-1.210	-1.106	-0.071	0.103	-0.226	-1.227	-0.258
Cluster 3	0.438	0.351	0.019	-0.058	-0.195	0.338	0.085

The table above represents the cluster positions after applying the algorithm to the standard variables, as shown in Table (7). The first cluster exhibited a significantly high VLDL value, with a center score of 3.841, indicating a mean that was approximately 3.841 standard deviations higher than the overall sample mean. This same cluster also showed relatively high BMI and LDL values, with centers reaching 0.896 and 0.343, respectively, while the age variable (AGE) had a negative standard value of -0.612. The second cluster's positions were clearly below the overall mean for several key variables, most notably age (-1.210), HbA1c (-1.106), and BMI (-1.227). These values deviated by more than one standard deviation, while the remaining variables cholesterol, LDL, VLDL, and TG-HDL ratio—were closer to the overall mean. The third cluster showed positive and moderate standard values in age (0.438), glycated hemoglobin (0.351), and body mass index (0.338), with the centers of the other variables remaining stable near zero, reflecting their direct proximity to the overall mean of the sample.

Table (8) above represents the cluster validity indices resulting from the application of the K-Means algorithm. The sum of squared deviations within clusters (WCSS) reached a value of (4833.8033). This value was distributed across the three

clusters as follows: (472.06) for the first cluster, (957.91) for the second cluster, and (3403.83) for the third cluster. Considering the different cluster sizes, the mean squared distance to the cluster center was (9.441) for the first cluster, (4.094) for the second cluster, and (4.754) for the third cluster. These values indicate that the first cluster has a relatively higher internal dispersion when considering the number of observations in each cluster, while the second cluster shows a relatively lower dispersion according to this measure.

Table 8. Validity indicators for the K-Means algorithm

indicator	Value
Within-Cluster Sum of Squares (WCSS)	4833.8033
Silhouette Coefficient	0.4176
Davies-Bouldin Index (DBI)	1.3160
Calinski-Harabasz Index (CHI)	222.6734

The ratio of inter-cluster variance to total variance was approximately 30.9%, reflecting the amount of variance that cluster segmentation could explain between clusters compared to the total variance in the data. The average Silhouette coefficient was 0.4176, with values of 0.2769 for cluster 1, 0.4918 for cluster 2, and 0.4032 for cluster 3. These values indicate that cluster 2 exhibits the highest degree of relative differentiation, while cluster 1 shows a greater degree of overlap with other clusters. Additionally, 14 observations out of 1000 (1.4%) showed negative values, suggesting a limited number of observations that may be closer to other clusters. Overall, the Silhouette value of 0.4176 indicates moderate cluster differentiation with some overlap between clusters. In general, the K-Means results reveal three relatively distinct patterns of metabolic characteristics within the study data. The differences are most pronounced in the VLDL, BMI, AGE, HbA1c, and TG/HDL ratio variables, which represent different aspects of metabolic status related to insulin resistance. The first cluster is characterized by a clear pattern of significantly elevated VLDL with high BMI. The second cluster is characterized by relatively low levels of several metabolic markers. The third cluster appears to be the most prevalent pattern in the sample, with relatively high levels of AGE, HbA1c, and BMI. Therefore, the clustering results do not provide a direct diagnosis of insulin resistance, but rather reveal sub-patterns within the data that can be interpreted in light of the metabolic markers associated with insulin resistance.

Discussion

The comparison between the results of the Fuzzy C-Means and K-Means algorithms showed a difference in the nature of the resulting clustering, despite their application to the same dataset and the same variables and number of clusters. The K-Means results showed better geometric differentiation according to traditional validity indices, with the Silhouette coefficient (0.4176), the Davies-Bouldin index (1.3160), and the Calinski-Harabasz index (222.6734), compared to the corresponding FCM results of (0.2497), (1.8875), and (196.9645), respectively (Rousseeuw, 1987; Davies & Bouldin, 1979; Caliński & Harabasz, 1974). In contrast, the (FCM) algorithm provided additional information not offered by hard clustering, namely the membership scores of each observation within different clusters (Bezdek et al., 1984; Graziani et al., 2024). This allowed for the detection of varying levels of clarity in membership, rather than simply assigning each observation to a single cluster. The Partition Coefficient and Partition Entropy indices showed a relatively high degree of fuzziness in the distribution of membership scores, reflecting the indefinite nature of membership for a large number of observations (Pal & Bezdek, 1995). Therefore, the results do not indicate the absolute superiority of either algorithm, but rather illustrate the different analytical roles of each. K-Means provides a more geometrically clear segmentation based on traditional indices, while (FCM) allows for a more flexible representation of the degree of belonging and overlap between different patterns (Graziani et al., 2024). The importance of combining these two algorithms lies in providing a comprehensive analytical view of the data, combining clear segmentation with the ability to represent uncertainty and gradations in belonging.

Conclusion

Based on the results obtained from applying K-Means and Fuzzy C-Means to the study data, the main conclusions are as follows:

1. The application of K-Means and Fuzzy C-Means (FCM) revealed distinct patterns of metabolic characteristics among the study observations, highlighting the effectiveness of clustering techniques in exploring multivariate data.
2. FCM identified overlapping characteristics, particularly between clusters one and three, while cluster two showed distinct patterns in age, HbA1c, and BMI. The PC (0.3772) and PE (1.0306) values further indicated substantial uncertainty in cluster membership.

3. K-Means produced a more distinct cluster structure according to the traditional validity indices, with a Silhouette Coefficient of 0.4176 and a DBI of 1.3160. However, 71.6% of observations were concentrated in the third cluster, indicating an unequal cluster distribution.
4. The findings highlight the differences between hard and fuzzy clustering approaches. While K-Means assigns each observation to a single cluster, FCM captures the degree of membership and overlapping characteristics.
5. Neither algorithm can be considered universally superior, as their suitability depends on the data structure and analytical objectives. Combining both approaches provides complementary insights into the underlying metabolic patterns.

Acknowledgement

The authors are very grateful to the College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq, which helped improve this work's quality.

Funding information

This study was conducted with the support of a scholarship awarded by the Tertiary Education Trust Fund (TETFund), Nigeria, under its scholarship programme for my program (Master of Science (MSc) in Computer Science programme). The scholarship provided financial support throughout the study, including the research and related academic activities.

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Author contributions

Ahmed Saif Al-Din conducted the research, including data preparation and analysis, and the application of software techniques to analyze the data using two algorithms: fuzzy C-means clustering and the k-means algorithm. He then interpreted the results and compared them. The research was conducted under the supervision and guidance of Dr. Noor Nawzat A. Seyala, the author's academic supervisor. Her valuable intellectual guidance, direction, constructive contributions, encouragement, and comprehensive supervision played a crucial role in the successful completion of this research.

Ethics

Ethical approval was not required for this study as it did not involve human participants, personal data.

AI tool usage declaration

The authors should declare that no AI and associated tools are used for writing scientific content in the article.

References

Bezdek, J. C. (1973). Cluster validity with fuzzy sets. *Journal of Cybernetics, 3*(3), 58–73. https://doi.org/10.1080/01969727308546047

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Springer. https://doi.org/10.1007/978-1-4757-0450-1

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences, 10*(2–3), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Bora, D. J., & Mishra, M. K. (2025). Comparative evaluation of hard and soft clustering for precise brain tumor segmentation in MR imaging. *Journal of Advances in Mathematics and Computer Science, 40*(9), 127–141. https://doi.org/10.9734/jamcs/2025/v40i92050

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Deza, M. M., & Deza, E. (2016). *Encyclopedia of distances* (4th ed.). Springer. https://doi.org/10.1007/978-3-662-52844-0

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics, 3*(3), 32–57. https://doi.org/10.1080/01969727308546046

Ghosh, S., & Dubey, S. K. (2013). Comparative analysis of K-means and fuzzy C-means algorithms. International *Journal of Advanced Computer Science and Applications, 4*(4). https://doi.org/10.14569/IJACSA.2013.040406

Golzari Oskouei, A., Samadi, N., Khezri, S., Najafi Moghaddam, A., Babaei, H., Hamini, K., Fath Nojavan, S., Bouyer, A., & Arasteh, B. (2025). Feature-weighted fuzzy clustering methods: An experimental review. *Neurocomputing, 619*, 129176. https://doi.org/10.1016/j.neucom.2024.129176

Graziani, S., et al. (2024). Fuzzy k-Means: History and applications. Econometrics and Statistics, 30, 110–123. https://doi.org/10.1016/j.ecosta.2021.11.008

Hartigan, J. A., & Wong, M. A. (1979). A K-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 28*(1), 100–108. https://doi.org/10.2307/2346830

Hashemi, S. E., Gholian-Jouybari, F., & Hajiaghaei-Keshteli, M. (2023). A fuzzy C-means algorithm for optimizing data clustering. *Expert Systems with Applications, 227*, 120377. https://doi.org/10.1016/j.eswa.2023.120377

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Krasnov, D., Davis, D., Malott, K., Chen, Y., Shi, X., & Wong, A. (2023). Fuzzy C-means clustering: A review of applications in breast cancer detection. *Entropy, 25*(7), 1021. https://doi.org/10.3390/e25071021

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.

Nagase, Y., Satoh, T., Shigetome, K., Tokumaru, N., Matsumoto, E., Yamada, K. D., Imafuku, T., Watanabe, H., Maruyama, T., Ogata, Y., Yoshida, M., Saruwatari, J., & Oniki, K. (2023). Serum fatty acid composition balance by fuzzy C-means method in individuals with or without metabolic dysfunction-associated fatty liver disease. *Nutrients, 15*(4), 809. https://doi.org/10.3390/nu15040809

Pal, N. R., & Bezdek, J. C. (1995). On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy Systems, 3(3), 370–379. https://doi.org/10.1109/91.413225

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Wiharto, W., & Suryani, E. (2020). The comparison of clustering algorithms K-means and fuzzy C-means for segmentation retinal blood vessels. *Acta Informatica Medica, 28*(1), 42–47. https://doi.org/10.5455/aim.2020.28.42-47

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zhou, K., & Yang, S. (2019). Fuzzifier selection in fuzzy C-means from cluster size distribution perspective. *Informatica, 30*(3), 613–628. https://doi.org/10.15388/Informatica.2019.221