Research Paper
Computational Methods for Differential Equations
http://cmde.tabrizu.ac.ir
Vol. *, No. *, *, pp. 1-16
DOI:10.22034/cmde.2026.73378.3742

Unsupervised discovery of patient health patterns using K-Means clustering in healthcare data

Ali Subhi Alhumaima[1,*], Taqwa Salim Mohammed[2], Bashar Talib Al-Nuaimi[3], Hussein Alkattan[4], and Mostafa Abotaleb[4]

[1]Electronic Computer Centre, University of Diyala, Diyala, Iraq.

[2]College of Medicine, University of Diyala, Diyala, Iraq.

[3]Computer Science Department, University of Diyala, Diyala, Iraq.

[4]Engineering School of Digital Technologies, Yugra State University, Khanty-Mansiysk, Russia.

Abstract

The increasing availability of structured primary care data creates opportunities to identify latent patient subgroups that could support exploratory analysis, improve patient stratification, and inform clinical decision support. This paper presents an interpretable unsupervised learning model based on K-Means clustering for identifying patient health patterns from demographic, physiological, and clinical features. The analyzed dataset includes age, gender, blood pressure, pulse, cholesterol level, body mass index, diagnosis, and treatment plan. Missing numerical values were imputed, categorical variables were encoded, non-essential identifiers were removed, and retained features were normalized to prevent scale bias before clustering. The Elbow Method identified the selected number of clusters, while silhouette analysis assessed internal cohesion and separation. Model interpretation was augmented with Principal Component Analysis, a correlation heatmap, feature-distribution graphs, box plots, a radar chart, and cluster-size visualizations. Three balanced patient populations were identified with different demographic, cardiovascular, metabolic, diagnostic, and therapeutic characteristics. Cluster 1 showed the highest mean age, blood pressure, heart rate, and body mass index, indicating a more adverse risk profile. Cluster 0 showed high average cholesterol and diagnostic coding, while Cluster 2 was younger, with a lower heart rate, moderate body mass index, and an alternative treatment pattern. The mean silhouette score of 0.0929 indicates substantial overlap among patient records, so the clusters should be regarded as exploratory phenotypes rather than clinical categories. Despite this limitation, the combined clustering and visualization workflow reveals multivariate patterns that may remain unnoticed in univariate descriptive statistics. The framework provides a systematic starting point for healthcare-data evaluation and requires validation using larger, longitudinal, and externally sourced clinical datasets.

Keywords. K-Means clustering, Healthcare analytics, Unsupervised learning, Patient stratification, Clinical decision support, Interpretable machine learning.

2010 Mathematics Subject Classification. 62H30, 62P10, 68T05.

1. Introduction

Health systems are increasingly generating structured and semi-structured data through electronic health records, laboratory systems, monitoring devices, insurance systems and hospital information systems. The interpretation identifies patients from multiple complementary angles, including demographic data, physical examination, biochemical and metabolic information, as well as clinical diagnosis and therapy choices. When all the right data is examined, it will reveal trends that are much harder to see using an individual variable or relatively standard summary measures. As a result, machine-learning approaches have emerged as an important analytical tool for health-risk assessment and disease surveillance and clinical decision-making because they are able to analyze the multidimensional data and extract relationships that may help practitioners and healthcare professionals [1, 16].

Received: 04 July 2026; Accepted: 03 August 2026.
∗ Corresponding author. Email: alhumaimaali@uodiyala.edu.iq.

One major challenge, though, is that many healthcare databases do not contain reliable target labels. As current models of care are increasingly reliant on data and channels of health, diagnostic classifications may be too limited (the result of a single test is unlikely to correlate perfectly with what you see in the clinic), therapeutic codes may not produce a one-to-one correlation with your result. Supervised models are limited in these scenarios as they require a known response variable ahead of time. Another perspective comes from the area of unsupervised learning that investigates the underlying structure in feature space. Whereas clustering identifies patients into groups based on similarity, without prediction or predetermined outcome, and may also discover subpopulations for further clinical investigation. This capability is critical to precision medicine, wherein patients with the same nominal diagnosis share heterogeneous risk profiles, physiology, treatment dependencies, or predicted burden of healthcare resource utilization [8, 19].

At its core, patient classification is a multi-faceted problem. On the other hand, while cardiovascular and metabolic health are characterized by age, blood pressure, heart rate, cholesterol levels and BMI diagnosis & therapeutic factors provide clinical context. A patient with moderate characteristics on a single variable may show an even higher risk profile when many are evaluated together. By clustering, these features could be combined together, and records could be grouped according to their similarity. However, innovative healthcare research has demonstrated that unsupervised analysis can uncover clinically distinguishable risk categories, medication treatment and resource use patterns [6, 15, 21]. These findings support the use of clustering for exploratory analysis of population health metrics.

K-Means remains a widely used among centroid-centric techniques because it is computationally efficient, scalable, reproducible by seeding initializations, and relatively interpretable. This method assigns each observation to the nearest centroid and iteratively improves the cluster centers, minimizing intra-cluster variability. These features make it a suitable transparent base for tabular data in the healthcare space. However, its simplicity does not free us from important methodological requirements; K-Means is sensitive to feature scales, to outliers, to initialization, to the number of clusters you select and also the numerical encoding of respective categorical variables. Therefore, an ideal healthcare application needs to properly log the preprocessing decisions taken and evaluate their impact critically rather than directly claim the cluster labels being equivalent as clinical diagnosis.

This is quite critical for data preparation, particularly when demographic, physiological and clinical factors are considered simultaneously [5]. Handle missing values without inflicting drastic bias, remove non-informative identifiers, and pass by normalizing the variables that are measured in different units. Without scaling, features with larger numerical ranges may dominate the Euclidean distance and determine the split that is created. Similarly, categorical encoding has to be expressed explicitly as numbers can lead to a false order. Modern pretreatment guidelines highlight that data transformation and feature engineering are aspects of the analytical model, rather than mere auxiliary technical processes [11, 13].

Another challenge is the selection and validation of clusters. It can be hard to evaluate the quality of the partition in unsupervised learning since there are no ground-truth labels, making classification accuracy a poor metric. The Elbow Method provides a intuitive guide for where diminishing reductions in within-cluster sum of squares take place as K increases. In this respect, silhouette analysis improves the assessment through a comparison of intra-cluster coherence and the distance to the nearest neighboring cluster [17]. Countless studies in real-world health datasets would earn a low or modest silhouette score because the biological and clinical conditions are in constant flux - not simply defined categories. As such, validation results should be reported honestly and interpreted in the context of feature profiles, cluster size and relevance to domain.

Interpretability is central to the real-world utility of clustering. An algorithm yielding several labels for a division is not useful unless the resulting classes can be characterized by variables that are interpretable to clinicians and health care analysts. Various visualization methods such as PCA projections, box plots, cluster-profile heatmaps, distribution plots and correlation heatmaps demonstrate differences in the cluster centroids, variability of clusters or overlap between clusters or segments including outlier detection. Combining quantitative summaries with various visual perspectives reduces key metric dependency and facilitates the exploration of underlying patient trends.

Despite a growing literature surrounding artificial intelligence in healthcare, much of the existing work still centers around supervised prediction, image classification or disease-specific outcome assessment. Few studies present a

holistic, interpretable framework for making sense of broad structured healthcare data containing demographic, physiological, diagnostic and therapy information. Additionally, some clustering studies focus on the performance of the algorithm with minimal clinical elucidation of the observed clusters. This creates a methodological gap for conducting transparent exploratory research which builds clear linkages between preprocessing, cluster-number determination, internal validation, multivariable visualization and cautious clinical inference.

In order to address this deficiency, the present study employed K-Means clustering on a structured dataset including age, gender, blood pressure, heart rate, cholesterol level, BMI as well as diagnosis and treatment regimen. This contribution is not in generating a new clustering method, but rather about the formulation and evaluation of a joint coherent reusable and interpretable analytical framework to discovering patient profiles across heterogeneous healthcare data. The research: (i) prepares several variables for distance-based analysis with imputation, encoding, feature selection, and standardization; (ii) determines a three-cluster solution using the Elbow Method; (iii) evaluates the partition by silhouette analysis; and (iv) interprets patient groups through tabular summaries additional visualizations. The resulting clusters should be considered exploratory health profiles, rather than diagnostic labels to maintain a clear distinction between identification of patterns driven by data and clinical verification.

2. Related work

In current literature on healthcare artificial intelligence, it emphasizes the criticality of data-centric approaches towards prevention, diagnosis and patient surveillance. Hameed Mousa et al. [10] designed an early screening and intervention mobile health framework for diabetes, while Alazzawi et al. [3] presented a novel deep-learning-based multimodal health data-enhanced framework for heart disease early-stage detection. These investigations illustrate the increasing ability of computational models to integrate many aspects of health. However, the main focus for them is supervised prediction. The present work addresses an additional but related question: Can we identify important patient structures without a definitive target label?

Healthcare analytics is strongly influenced by data preprocessing and feature representation. Kang et al. [12] combined dynamic clustering with particle-swarm optimization in order to perform feature selection from high-dimensional healthcare datasets, while Dai et al. [7] present a method combining neighborhood-entropy for heterogeneous feature selection. Izonin et al. Normalization techniques can greatly impact the effectiveness of medical models as demonstrated by [11]. Although their contributions mostly focus on improving prediction models, their main finding also has implications for clustering: the usefulness of the inferred structure is determined by how well comparable are variables. This study explicitly records how missing values are dealt with, the encodings applied for different categories, features omitted and normalization.

Agyemang et al. Class imbalance in the context of high dimension low sample size (HDLSS) health-data categorization including preprocessing influence remains also unexplored [2]. Shoaib Al-Khamees et al. [4] proposed an adaptive k-nearest-neighbor algorithm for diabetes and breast cancer prediction. In both investigations, the general trend toward algorithmic decision assistance is strengthened, but they are guided by results with annotations. Exploratory clustering does not require knowledge of clinically meaningful categories a priori. Thus, it may be useful during the exploratory part of research when the goal is to develop hypotheses, explore variation or discover subset populations for later supervised confirmation.

Machine learning has progressed rapidly in medical imaging. Amin et al. [5] presented a joint approach to classify COVID-19, tuberculosis and pneumonia using chest radiographs. Vong et al. Zhu et al. [20] This review focused on deep-learning techniques for glioma histopathology Photoacoustic-orientated Feature Based Framing for Lung Cone-Beam CTv [22]. We explore these image-centric inquiries which operate with high-dimensional visual datasets with diagnostic or reconstruction objectives. Although they capture the variety of medical AI, they differ from the present study in that those tasks are not all employed based on structured data. This gap emphasizes the need for transparent yet practical methods when only tabular clinical data is available.

New studies provide direct evidence for clustering inside the healthcare sector. Momahhed et al. [15] investigated K-Means clustering on outpatient prescription claims to identify groups characterized by distinctive medication-exposure trajectories. Yang et al. A smart-healthcare framework in which cluster methodologies were applied to medical (imaging) information [21]. Eckardt et al. In the field of cancer research, [8] used unsupervised meta-clustering

on clinical and genomic data to identify risk profiles in acute myeloid leukemia. Taroi (Yassin Cataniciu) recently characterized a cohort of patients with chronic ulcers [6] and K-Means clusters to healthcare resource utilization. Together, these investigations demonstrate that clustering can expose clinically or operationally important subtypes of patients in a broad range of health care settings.

Methodology continues to improve upon classic partitioning methods. Zubair et al. On the other hand, [23] proposed an improved K-Means method with better performance of data-driven modeling, whereas Lai et al. In [14], silhouette-based weighting was applied to improve the effectiveness of K-Means. These developments address well-known limitations of traditional centroid-based clustering, such as sensitivity to initialization and geometric cluster shape. The current work intentionally uses the standard K-Means algorithm as an understandable baseline. This alternative enables a clear assessment of insights derived from disciplined preparation, internal validation and visualization first before the more complex alternatives are presented.

Since, unsupervised methods cannot be monitored by existing labels internal validation is essential. Silhouette analysis is used to evaluate clustering and a comprehensive review of this topic can be found in the work by Shutaywi and Kachouie [17], while Ekemeyong Awong and Zielinska [9] analyzed clustering quality in self-organizing maps. These studies illustrate that for a cluster solution; no single metric fully explains its value. But a low silhouette value suggests true overlap, especially with health populations where risks are changing from time to time. Thus, the present work merges together the Elbow Method and silhouette coefficient via cluster equilibrium proportions, centroid profiles, PCA based ac- curacy visualization, as well as feature level distributions.

3. Data and methodology

3.1. **Data description.** This study uses a healthcare dataset provided by Kaggle [18]. It contains 500 patients' data which includes demographic, physiological and clinical datasets. Preserved variables are Age, Gender, Blood Pressure, Heart Rate, Cholesterol Level, BMI and Diagnosis and Treatment Plan. Together, these features provide a multidimensional characterization of patients and allow an exploratory assessment of whether records are homogeneous subsets without specifying cluster-based outcome classes in advance.

The dataset from an analytical point of view can be represented as a multivariate matrix:

$$X = [x_{ij}] \in R^{n \times p}, \tag{3.1}$$

where n is the total number of patient records while p is the number of selected healthcare features. The dataset can be clarified using a feature vector: Each patient i has a feature vector:

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}), \tag{3.2}$$

where x_{ij} is the value of j-th feature for i-th patient. The healthcare dataset can be understood as patient vectors in multivariate feature space.

Three conceptual categories were identified in the preserved characteristics. Age and gender were collected as demographic data. Physiological and metabolic factors included blood pressure, heart rate, cholesterol concentration and body mass index. The clinical details comprised diagnosis and management. This integration enables the clustering algorithm to account for background characteristics, current health indicators/metrics and management data at the same time.

Figure 1 is a schematic of relationships between the quantitative and encoded features. The diagonal elements are equal to one because each variable is fully correlated with itself. Most off-diagonal coefficients are weak, which indicates loose linear dependence among the variables selected. Weak correlations, do not mean clinical irrelevance as they suggest that certain characteristics provide additional insight, and the final clustering outcome is highly unlikely to be determined by a single significant collinear pair.

3.2. **Data preprocessing.** Prior to clustering, a systematic preprocessing stage was performed to reduce the possibility of missing values, inconsistent data types, identifiers or different scaled measurements generating false splits. Data preparation involved the following steps: imputation of missing values, encoding of categorical variables, non-informative identifiers removal, features selection and Z-score normalization. These changes were done before the assessment of the final K-Means model.

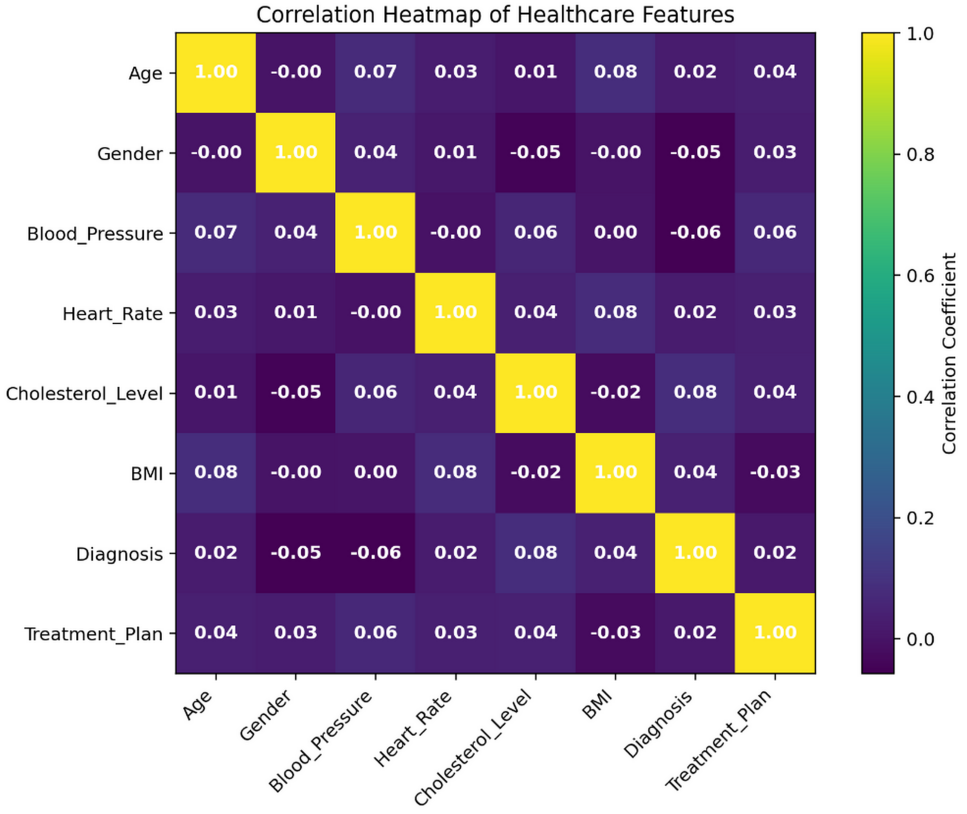

FIGURE 1. Correlation heatmap of healthcare features.

Missing numerical observations were replaced with the given attribute median. Median imputation was preferred across other methods because it is much more robust and yields less bias for clinical measurements with skewed distributions or outliers when compared to mean imputation. This technique retains all patient records but diminishes the effect of statistical outliers.

For a numerical feature x_j, the imputed value is expressed as:

$$x_j^{imp} = median\,(x_j)\,. \tag{3.3}$$

This strategy preserves the central tendency of the variable while minimizing distortion caused by extreme observations.

Gender, Diagnosis, and Treatment Plan are categorical variables and so cannot be easily incorporated into normal Euclidean K-Means. These variables were converted into numerical codes prior to model fitting. This depiction facilitates computation, although the ensuing codes have to be regarded as analytical encodings instead of clinically significant continuous scales.

If a categorical variable contains k distinct categories, the transformation can be represented as:

$$f\,(c_i) = z_i,\ \ i = 1, 2, \ldots, k, \tag{3.4}$$

where c_i denotes the original category and z_i its corresponding numerical value.

Fundamental label encoding has a limitation that by mapping Categories to numbers, it may indicate there is an ordering or distance between categories. The clusters were cautiously evaluated as the encoded clinical variables were

not used to guide segmentation. Future validation should assess one-hot encoding or clustering methods targeted in heterogeneous data types.

Variables not clinically related to patient health, such as a Patient ID (identifiers offer no clinical information and may cause unnecessary variation), were filtered out. Retained those factors which were related to demographics, physiology, metabolism, diagnosis or therapy.

The final feature subset used in this study is given by:

$$X^{'} = \{\text{Age},\ \text{Gender},\ \text{Blood Pressure},\ \text{Heart Rate}, \text{Cholesterol Level},\ \text{BMI},\ \text{Diagnosis},\ \text{Treatment Plan}\}. \tag{3.5}$$

This results in a more interpretable feature space that ensures similarity is derived from patient characteristics rather than administrative identifiers.

Several numerical scales were then used to evaluate the preserved variables. For example, age, blood pressure, cholesterol and BMI values are between certain ranges, while our categories that we encode using few integers. If we compute Euclidean distance on unnormalized values the feature ranges will dominate the partition.

To alleviate this issue, we centered and scaled all of the features with z-score normalization. The transformation is defined as:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{3.6}$$

where

- x_{ij} is the original value of the j-th feature for the i-th patient,
- μ_j is the mean of feature j,
- σ_j is the standard deviation of feature j,
- z_{ij} is the normalized value.

Z-score normalization normalizes each variable to a mean of zero and expresses the observations as how many standard deviations they are from the mean. This adjustment makes the scaled sizes have a fairly similar effect on the K-Means objective, still nothing that removes the conceptual limits of encoding categorical features completely.

The analytical process is summarized in Figure 2. It starts with the well-structured healthcare data set, passes through several steps of data cleaning, imputation, categorical encoding, feature selection and standardization then evaluates the possible number of clusters using the Elbow Method. Thereafter, the definitive K-Means essentially labels clusters, which is assessed through silhouette analysis and clarified with numeric summaries and graphical representations.

3.3. **K-Means clustering.** The K-Means function was used in the second step, as the main algorithm for unsupervised learning. When given a new patient, the algorithm will separate patient records into K different disjoint clusters such that variation within each group is kept to a minimum. Members assigned to the same cluster are hypothesized to have more similar standardized feature profiles, while members assigned to different clusters are hypothesized to be more dissimilar. These methods were chosen for being a common standard for exploratory analysis of tabular healthcare data, clear and computationally efficient.

Let the dataset be partitioned into K clusters:

$$C = \{C_1, C_2, \ldots, C_K\}, \tag{3.7}$$

such that

$$C_i \cap C_j = \varnothing \quad \text{for } i \neq j, \tag{3.8}$$

and

$$\bigcup_{i=1}^{K} C_i = X^{'}. \tag{3.9}$$

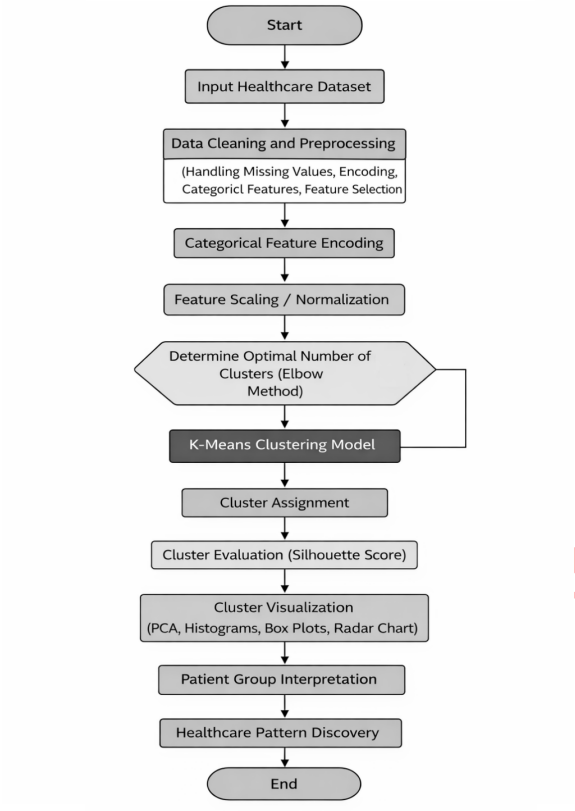

FIGURE 2. Flowchart of the proposed K-Means clustering framework for healthcare data analysis.

3.4. **K-Means objective.** The objective of K-Means is to minimize the WCSS (within-cluster sum of squares), also known as the clustering objective function:

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2, \tag{3.10}$$

where

- J is the total clustering cost,
- C_i denotes the i-th cluster,
- μ_i is the centroid (mean vector) of cluster i,
- $\|x - \mu_i\|^2$ is the distance between a patient record and the cluster centroid.

Each cluster centroid is calculated by:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x, \tag{3.11}$$

where $|C_i|$ denotes the number of observations assigned to cluster i.

3.5. **Distance Metric.** Similarity between standardized patient profiles was quantified using the Euclidean distance. This measure is the same as the K-Means objective function and calculates the direct distance in feature space for two observations after data transformation. The clustering is also less biased, since no individual continuous feature might

be allowed to dominate the distance calculation only because it is measured in a different unit (for example meters) or numeric range.

$$d(x_p, x_q) = \sqrt{\sum_{j=1}^{p} (x_{pj} - x_{qj})^2}. \tag{3.12}$$

This value indicates the proximity or distance between two patient profiles in the standardized feature space. Because all features had been normalized before applying clustering, Euclidean distance serves as an equally weighted similarity metric across all dimensions.

3.6. **Algorithmic procedure.** K-Means clustering is an iterative optimization process that can be distilled into the following steps:

1. Select the number of clusters K.
2. Initialize K centroids randomly in the feature space.
3. Assign each patient record to the nearest centroid according to Euclidean distance.
4. Recalculate each centroid as the mean of all records assigned to that cluster.
5. Repeat the assignment and update steps until convergence is achieved.

Convergence occurs when either:

- cluster assignments no longer change, or
- centroid positions become stable between iterations.

3.7. **Optimal number of clusters.** The value chosen for K is a crucial aspect of the model. Too few clusters may merge clinically different profiles, while too many clusters may lead to unstable or over-fragmented groups. Elbow Method was used here, testing across the different possible K values through the movement of within-cluster sum of squares The solution is concordant to the way beyond which more clusters will produce rather minor improvements in compactness.

The Elbow Method was used to decide the most suitable number of clusters. This method examines WCSS for a range of K and takes note of where the decrease in clustering error appears to plateau.

The WCSS for a specific K is computed as follows:

$$WCSS(K) = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2 . \tag{3.13}$$

The optimal value of K is selected at the "elbow point", where adding more clusters does not provide a significant gain in compactness.

3.8. **Cluster Validation.** Since there are no reference labels, internal validation was used to evaluate the partition. The silhouette coefficient compares a patient to every other patient in the same cluster average distance to the patient in the nearest neighboring cluster smallest average distance. Positive values imply that a record is closer to its given cluster, values near zero indicate border overlap and negative values suggest an alternative cluster would be more appropriate for the assignment.

Let:

- $a(i)$ = average distance between record i and all other records in the same cluster,
- $b(i)$ = minimum average distance between record i and all records in the nearest neighboring cluster.

Then the silhouette score for observation i is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}. \tag{3.14}$$

The overall silhouette score for the dataset is computed as:

$$S = \frac{1}{n} \sum_{i=1}^{n} s(i). \tag{3.15}$$

The interpretation is as follows:

- $S \approx 1$: strong cluster separation,
- $S \approx 0$: overlapping clusters,
- $S < 0$: potential misclassification.

This measure provides an indication of whether the discovered patient groups are compact and distinct.

3.9. **Reduction and visualization.** PCA was used for visualization purpose 2D projection of standardized feature space only and clustering itself was performed in the full set of retained standardized variables. PCA is a dimensionality reduction technique that succinctly capture the primary axis of variation in the data and then visually expose how well clusters separate or overlap. To characterize group-level differences, we used additional box plots, distribution plots, a radar chart and cluster-frequency plots.

The transformation induced by PCA can be expressed as:

$$Z = X^{'}W, \tag{3.16}$$

where

- $X^{'}$ is the standardized feature matrix,
- W is the matrix of eigenvectors,
- Z is the transformed lower-dimensional representation.

The covariance matrix is computed as:

$$\Sigma = \frac{1}{n-1}\left(X^{'}\right)^{T} X^{'}, \tag{3.17}$$

and the principal components are obtained by solving:

$$\Sigma v = \lambda v, \tag{3.18}$$

where v represents an eigenvector and λ is the corresponding eigenvalue.

3.10. **Software implementation and reproducibility.** implemented of the analysis using common open-source scientific-computing libraries in Python. The scikit-learn library was used for preprocessing, K-Means clustering, silhouette analysis and PCA, the pandas and NumPy libraries were used to handle data and perform numerical operations. Visuals were created using Matplotlib and Seaborn. To improve reproducibility, a random state should be set for centroid initialization. Further, any public release of the code should also report the complete preprocessing configuration, candidate K range, initialisation settings and library versions.

3.11. **Clinical interpretation and practical applications.** The proposed workflow could enable preliminary identification of groups in need of further evaluation in addition to population health assessment and exploratory patient stratification. A higher-risk cluster (e.g. Cluster 1) may support targeted screening or resource planning, whereas differentiated prevention or follow-up strategies may be indicated for other clusters. These applications continue to function as decision-support tools, rather than automated clinical decisions. Cluster membership must not be a substitute for physician judgment, established diagnostic criteria or prospective validation. The low silhouette score stresses the need for external outcome validation, an analysis of stability, replication in a larger dataset and clinically relevant encoding strategies before any clinical application.

4. Results

The standardized healthcare dataset was split into three clusters using K-Means. The three groups exhibited specific patterns with respect to age, blood pressure, heart rate, cholesterol level, BMI, diagnosis encoding and treatment-plan encoding. These differences suggest the algorithm was able to capture multivariable patient profiles instead of classifying records based on a single measurement. Cluster interpretation: As the analysis was unsupervised and the average silhouette value was low, clusters are interpreted as exploratory phenotypes needing external clinical validation.

Table 1 shows the distribution of the 500 patient records. There are 170 patients (34.0%) in cluster 0, 157 patients (31.4%) in cluster 1 and there are 173 patients (34.6%) in cluster 2. The sizes are more or less equal, thus assuaging

fears the solution is biased by a tiny fringe cohort. However, balance in itself does not indicate clinical validity; it should be viewed alongside the cluster profiles and internal-validation findings.

TABLE 1. Cluster distribution of patient records.

Cluster	Number of patients	Percentage (%)
Cluster 0	170	34.0
Cluster 1	157	31.4
Cluster 2	173	34.6
Total	500	100.0

Table 2 provides the clearest basis on which to articulate rationale for the three groups. Cluster 1 was the cluster with the highest mean age (61.02 years), blood pressure (139.89), heart rate (85.66) and BMI (34.10). This combination has a relatively high cardiovascular and metabolic risk profile. Cluster 0, is younger on average and has the highest mean cholesterol level (248.54), as well as the highest diagnosis encoding determining a more peculiar clinical pattern not explained by age or BMI alone. Cluster 2 is also the youngest, has the lowest mean heart rate (72.54), moderate blood pressure, and the highest encoding of treatment-plan. The results illustrated that the groups differ in their combinations of manifest variables characteristic of them.

TABLE 2. Mean feature values across the identified clusters.

Feature	Cluster 0	Cluster 1	Cluster 2
Age	41.62	61.02	42.55
Gender	0.73	0.99	1.20
Blood pressure	119.60	139.89	138.05
Heart rate	81.10	85.66	72.54
Cholesterol level	248.54	212.82	218.76
BMI	26.95	34.10	26.72
Diagnosis	3.12	2.04	0.97
Treatment plan	1.66	2.06	2.29

The selected pairwise associations are presented in Table 3. Most connections fall in the weak or weak-to-moderate correlation range, indicating that the attributes do not directly serve as substitutes for one another. This is suitable for exploratory clustering, since one side of the split could be a healthy attribute, while the other indicates symptoms of disease. We only use correlation matrix to measure linear dependence without accounting for nonlinear dependences or interactions.

TABLE 3. Correlation matrix summary of the selected healthcare features.

Feature pair	Correlation interpretation
Age–BMI	Weak positive relationship
Blood pressure–heart rate	Weak to moderate relationship
Cholesterol level–diagnosis	Weak positive relationship
Gender–cholesterol level	Weak negative relationship
BMI–heart rate	Weak positive relationship

Table 4 presents the results of the internal assessment. Elbow Method revealed that K = 3, while the average profile of silhouette score was at 0.0929. A value close to 0 indicated that a lot of observations are located on the edges of clusters, meaning the groups are quite overlapping with each other. The result should be seen as either differential separation or gentle separation. Indicates a nonserious but perceptible exploratory schema among a mixed age-class.

These clusters may lay the foundation for a hypothesis-generating descriptive classification but should not operate as independent diagnostic entities.

TABLE 4. Clustering quality evaluation.

Evaluation metric	Value
Optimal number of clusters (K)	3
Average silhouette score	0.0929

For k = 1 to 10, the decrease in the within-cluster sum of squares is plotted in Figure 3. In the first few solutions, the curve drops off rapidly and begins to plateau. The strongest change in slope suggests a partition of three clusters. Maybe it is not the right k but, for sure, the selected K represents a balanced compromise between interpretability and throughput even thougth this rule of thumb is partly subjective and should be applied together with silhouette analysis.

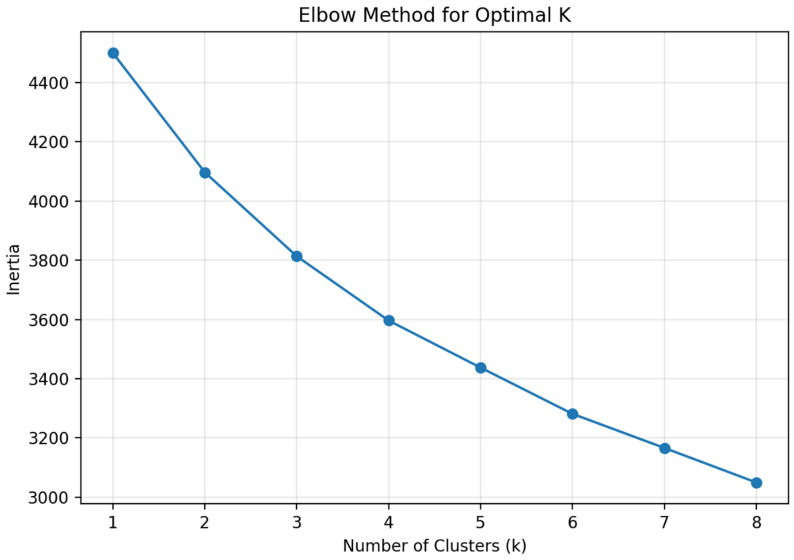

FIGURE 3. Elbow Method for selecting the number of clusters.

The clustered records are visualized on the first two principal components (PC1 and PC2) in Figure 4. There is a substantial amount of overlap, but each color occupies partially different regions. This agrees with our earlier results showing a low silhouette score and supports the notion that patient profiles fall along continua rather than form isolated classes. All the following experiments use only PCA for visualization, not as a substitute for evaluation in full standardized feature space.

Figure 5 confirmed that the numbers of patients are almost evenly distributed among the three groups. Clusters with comparable sizes improve the stability of feature summaries comparisons and help eliminate cases when a cluster only represents a small group of unrepresentative records. Clinical relevance is determined by the characteristics of the groups rather than the size.

Cluster-level feature profiles can be compared in a manner similar to that shown in Figure 6, where group-differentiating dimensions are most apparent. Cluster 1, which is older and visibly shows increases in blood pressure, heart rate, and BMI; cluster 0 show higher cholesterol and diagnosis encodings; and cluster 2 slightly lower HR alongside a different treatment pattern. Thus, the profile plot provides further evidence for a multivariable interpretation of patient heterogeneity.

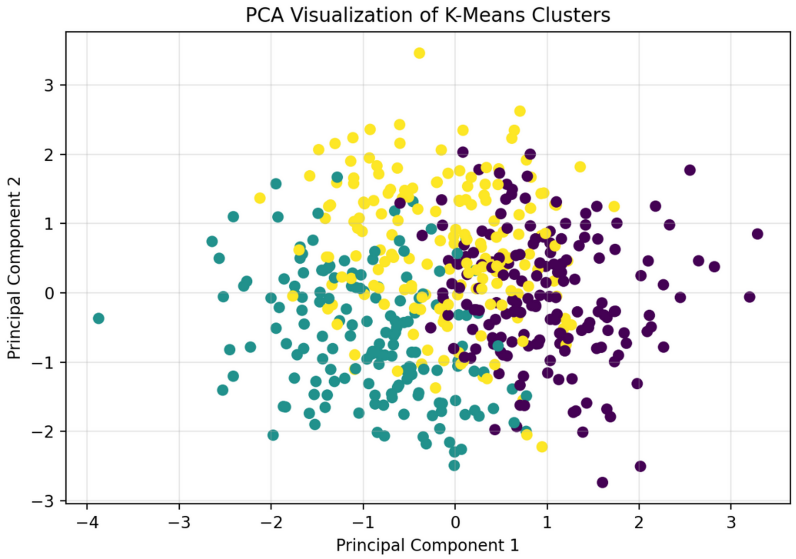

FIGURE 4. PCA visualization of K-Means clusters.

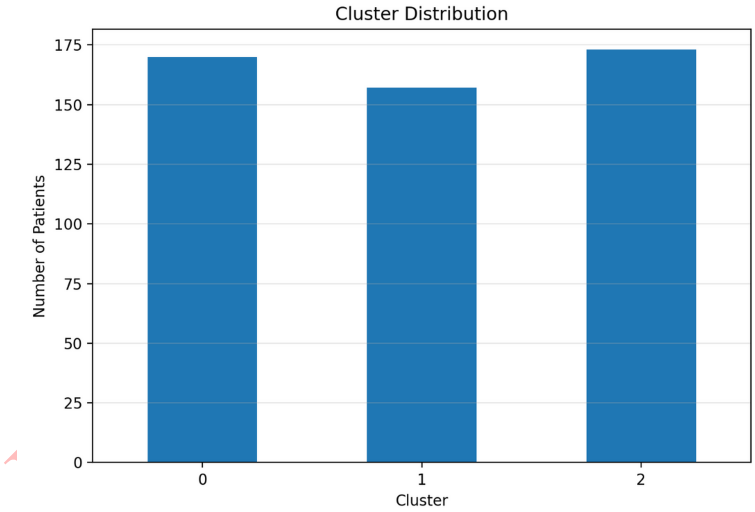

FIGURE 5. Distribution of patients across clusters.

Cluster-wise distribution of selected variables is shown in Figure 7. We can see differences in location and spread for age, blood pressure, heart rate, BMI and diagnosis encoding. Simultaneously, overlapping distributions justify low silhouette coefficient. The panel in the figure shows that clusters tell us about relative tendencies across many features, not deterministic thresholds.

A silhouette value for individual observations is shown in Figure 8. Values around zero and negative values imply ambiguous boundaries and alternative candidates within certain records. The average value ˜ 0.09 confirms weak separation. Cluster 2 looks comparatively more compact compared to portions of the plot, but no cluster is completely free from overlap. This finding underlines the 114 prudent uses of the model as an exploratory segmentation tool.

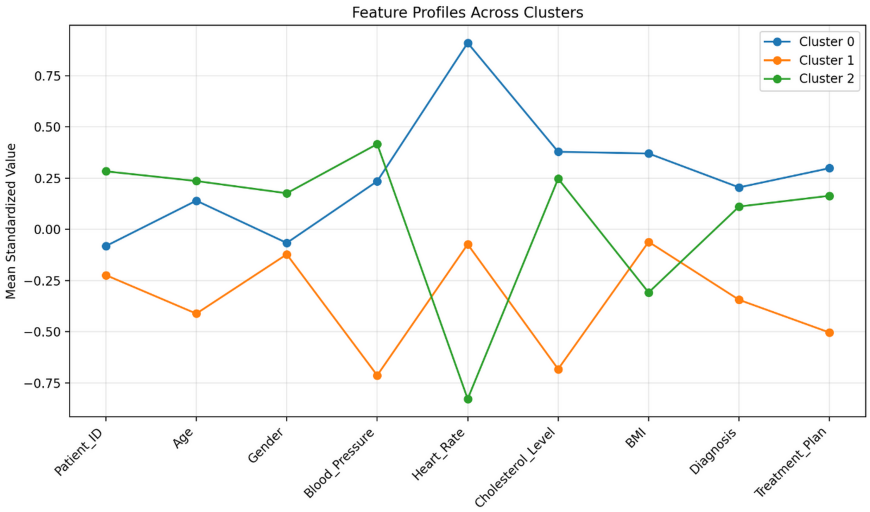

FIGURE 6. Feature profiles across clusters.

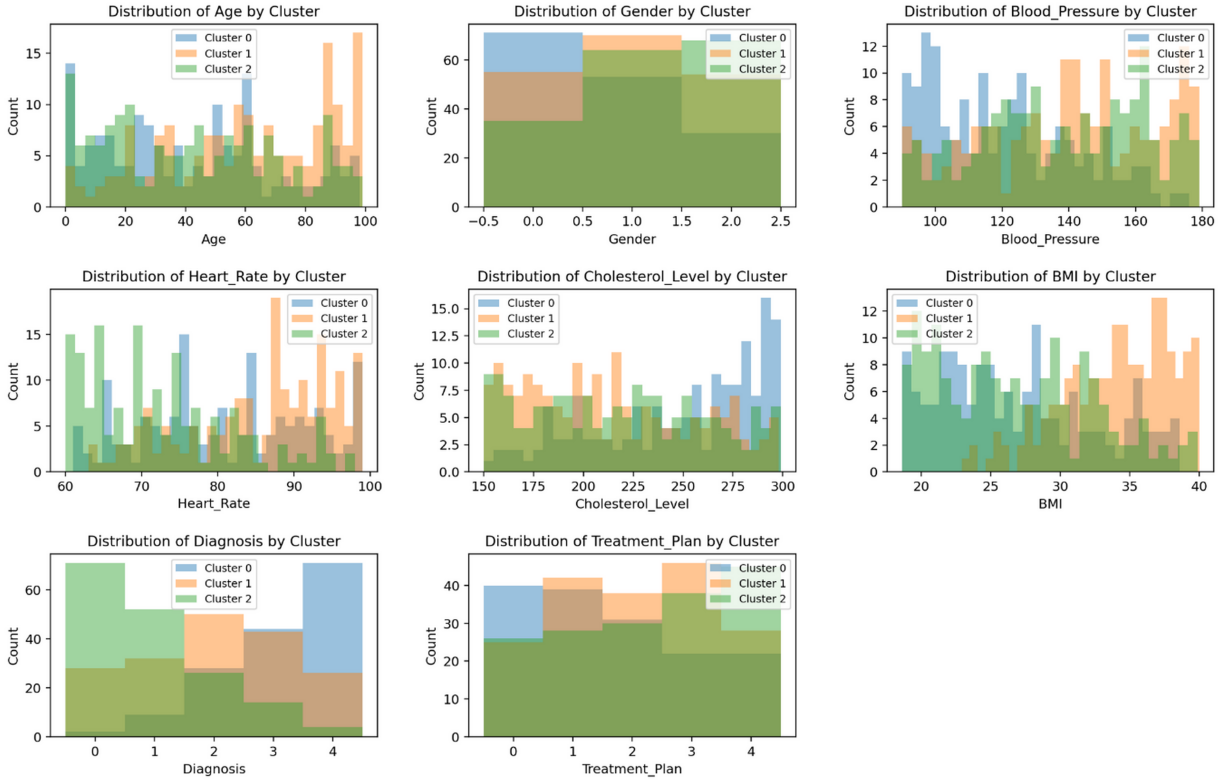

Figure: Cluster-wise distributions of the healthcare dataset features across the identified K-Means groups.

FIGURE 7. Cluster-wise distributions of healthcare features.

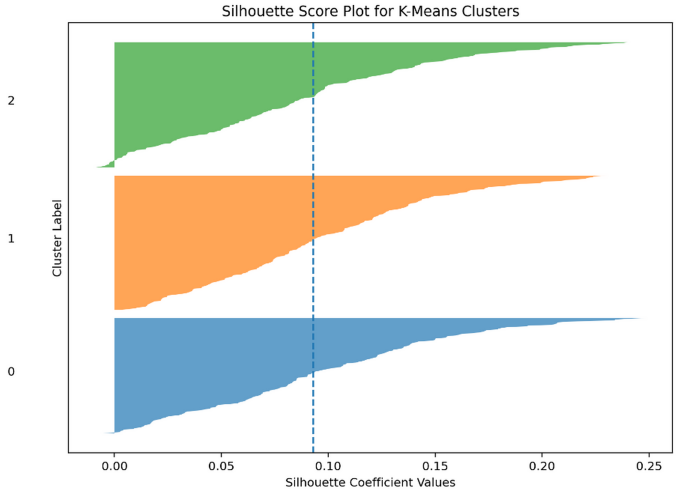

FIGURE 8. Silhouette scores for the K-Means clusters.

Medians, interquartile ranges and outliers for the key continuous variables presents in Figure 9. These verified Clusters from the first paragraph correspond to the fact that Cluster 1 typically consists of elderly patients with elevated blood pressure, heart rate and BMI. They additionally demonstrate sizeable within-cluster variation and overlap, in accordance with heterogeneous evidence from healthcare data. The plots thus enhance the clinical description of the groups and re-confirm that the partition does not delineate patient classes sharply.

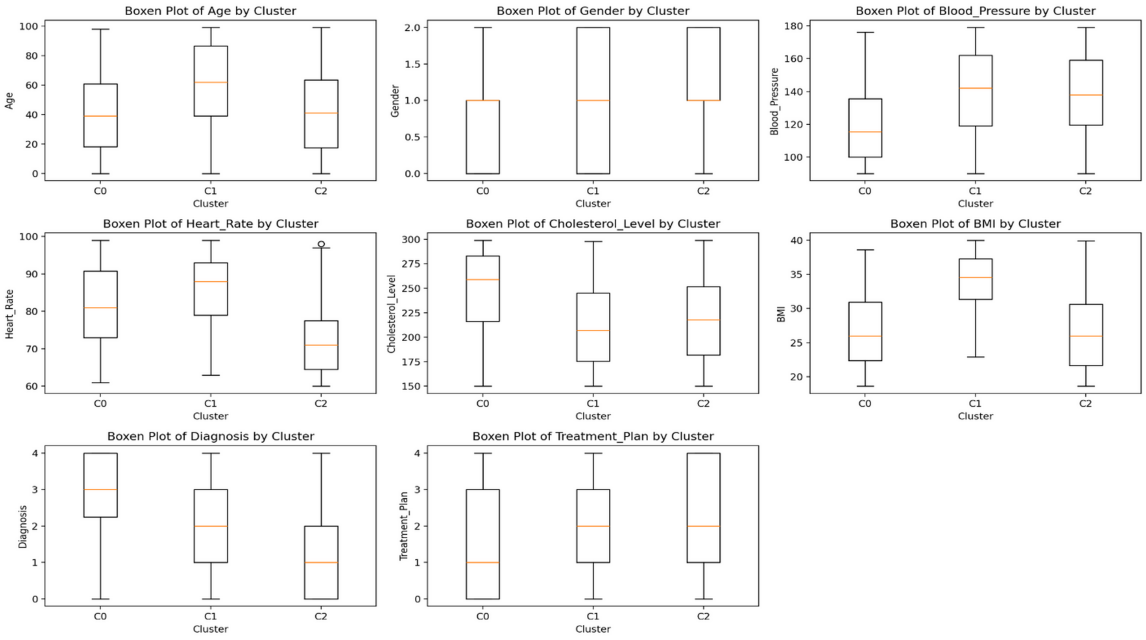

FIGURE 9. Cluster-wise box plots of healthcare features.

5. Conclusion

This study developed a K-Means pipeline that is interpretable for use in discovering latent patient patterns using structured healthcare data. Data were comprehensively analyzed to incorporate demographic, physiological, metabolic, diagnostic and treatment features from the dataset with imputation performed for missing-values; followed by categorical encoding of variables, removal of non-informative features and finally z-score standardization prior to clustering. To communicate the numerical partition, we characterized clusters based on visualizations and used the Elbow Method to justify a three-cluster solution.

Of the 1,637 patients screened from February 2013 to May 2016, the three groups were relatively comparable but clinically different. Cluster 1 with the features of elderly patients and the highest average blood pressure, heart rate and BMI contained the most pronounced cardiovascular and metabolic profile. Cluster 0 was characterized by high cholesterol and diagnosis-encoding mean, while Cluster 2 presented a specific treatment-plan pattern, average BMI and a lower heart rate. These findings further demonstrate that multivariable clustering can bring to light hidden patient heterogeneity that is otherwise masked by isolated summary statistics.

A silhouette score of 0.0929 is a important limitation. This implies that many patient records overlap and are not distinctly separate. Consequently, the identified groups should be considered exploratory profiles instead of categorical clinical phenotypes or diagnostic classes. Thus, while the study does provide some evidence that the population composition consists of three exploratory patient subgroups, its main contribution is offering a transparent and integrated analytical workflow.

The framework may facilitate hypothesis generation, descriptive patient stratification, targeted assessment of higher-risk profiles and planning of health services. Interpretability was supported by integrating cluster centers with PCA, silhouette analysis, cluster-profile heatmaps, distribution plots and box plot. Nonetheless, any clinical use would require physician supervision and outcome-based validation, as well as an assessment of the fairness of inferences across demographic groups.

Future work should investigate the stability of the clusters under repeated initialization and resampling. In addition, K-Prototypes compares standard K-Means with hierarchical clustering, Gaussian mixture models and density-based approaches and should have other strategies for one-hot encoding categorical variables as well. Larger multi-institutional and longitudinal datasets are required to assess generalizability and temporal change. Exploratory profiles must then be evaluated for their predictive or clinical value by assessing external outcomes, such as resource utilization, disease progression, treatment response or hospitalization. These represent potential pathways towards the responsible deployment of healthcare, enhanced by privacy-preserving federated analysis and explainable cluster-assignment methods.

Conflict of interest

The authors declare no conflict of interest.

References

[1] S. E. Abhadiomhen, E. O. Nzeakor, and K. Oyibo, *Health risk assessment using machine learning: A systematic review*, Electronics, *13*(22) (2024), Art. 4405.

[2] E. F. Agyemang et al., *Addressing class imbalance problem in health data classification: Practical application from an oversampling viewpoint*, Appl. Comput. Intell. Soft Comput., *2025* (2025), Art. 1013769.

[3] A. K. Alazzawi, H. Alharbi, H. A. A. Al-Khamees, and M. M. Abdul Zahra, *An intelligent and scalable framework for early heart disease detection using multimodal health data and optimized deep learning strategies*, SN Comput. Sci., *6*(7) (2025), Art. 765.

[4] H. A. A. Al-Khamees, N. S. Sani, A. S. Gifal, L. X. W. Liu, and M. I. Esa, *A dynamic model using k-NN algorithm for predicting diabetes and breast cancer*, Comput. Biol. Med., *192* (2025), Art. 110276.

[5] S. U. Amin, S. Taj, A. Hussain, and S. Seo, *An automated chest X-ray analysis for COVID-19, tuberculosis, and pneumonia employing ensemble learning approach*, Biomed. Signal Process. Control, *87* (2024), Art. 105408.

[6] M. Taroi Yassin Cataniciu, I. Gligorea, R. Fleacă, L. Vecerzan Novac, A. Prihoi, and C.-D. Domnariu, *Profiling patients with chronic ulcers using K-means clustering and analysis of the impact on the consumption of medical resources: Retrospective study on hospitalized patients in Romania*, J. Clin. Med., *14*(17) (2025), Art. 6252.
[7] J. Dai, W. Chen, and L. Xia, *Feature selection based on neighborhood complementary entropy for heterogeneous data*, Inf. Sci., *682* (2024), Art. 121261.
[8] J.-N. Eckardt et al., *Unsupervised meta-clustering identifies risk clusters in acute myeloid leukemia based on clinical and genetic profiles*, Commun. Med., *3*(1) (2023), Art. 68.
[9] L. E. Ekemeyong Awong and T. Zielinska, *Comparative analysis of the clustering quality in self-organizing maps for human posture classification*, Sensors, *23*(18) (2023), Art. 7925.
[10] A. Hameed Mousa et al., *Diabetes at a glance: Assessing AI strategies for early diabetes detection and intervention via a mobile app*, Mesopotamian J. Comput. Sci., (2025), 288–301.
[11] I. Izonin, R. Tkachenko, N. Shakhovska, B. Ilchyshyn, and K. K. Singh, *A two-step data normalization approach for improving classification accuracy in the medical diagnosis domain*, Mathematics, *10*(11) (2022), Art. 1942.
[12] Y. Kang et al., *A fast hybrid feature selection method based on dynamic clustering and improved particle swarm optimization for high-dimensional health care data*, IEEE Trans. Consum. Electron., *70*(1) (2024), 2447–2459.
[13] P. Koukaras and C. Tjortjis, *Data preprocessing and feature engineering for data mining: Techniques, tools, and best practices*, AI, *6*(10) (2025), Art. 257.
[14] H. Lai, T. Huang, B. Lu, S. Zhang, and R. Xiao, *Silhouette coefficient-based weighting K-means algorithm*, Neural Comput. Appl., *37*(5) (2025), 3061–3075.
[15] S. S. Momahhed, S. Emamgholipour Sefiddashti, B. Minaei, and Z. Shahali, *K-means clustering of outpatient prescription claims for health insureds in Iran*, BMC Public Health, *23*(1) (2023), Art. 788.
[16] S. Rani et al., *Machine learning-powered smart healthcare systems in the era of big data: Applications, diagnostic insights, challenges, and ethical implications*, Diagnostics, *15*(15) (2025), Art. 1914.
[17] M. Shutaywi and N. N. Kachouie, *Silhouette analysis for performance evaluation in machine learning with applications to clustering*, Entropy, *23*(6) (2021), Art. 759.
[18] S. Singh, *Healthcare Dataset*, Kaggle, 2024. Available at: https://www.kaggle.com/datasets/sanlavisingh/healthcare-dataset-csv-32-13-kb (accessed 1 April 2026).
[19] A. Trezza, A. Visibelli, B. Roncaglia, O. Spiga, and A. Santucci, *Unsupervised learning in precision medicine: Unlocking personalized healthcare through AI*, Appl. Sci., *14*(20) (2024), Art. 9305.
[20] C. K. Vong, A. Wang, M. Dragunow, T. I.-H. Park, and V. Shim, *Brain tumour histopathology through the lens of deep learning: A systematic review*, Comput. Biol. Med., *186* (2025), Art. 109642.
[21] W.-C. Yang, J.-P. Lai, Y.-H. Liu, Y.-L. Lin, H.-P. Hou, and P.-F. Pai, *Using medical data and clustering techniques for a smart healthcare system*, Electronics, *13*(1) (2024), Art. 140.
[22] J. Zhu et al., *Feature-targeted deep learning framework for pulmonary tumorous cone-beam CT enhancement with multi-task customized perceptual loss and feature-guided CycleGAN*, Comput. Med. Imaging Graph., *121* (2025), Art. 102487.
[23] M. Zubair, M. A. Iqbal, A. Shil, M. J. M. Chowdhury, M. A. Moni, and I. H. Sarker, *An improved K-means clustering algorithm towards an efficient data-driven modeling*, Ann. Data Sci., *11*(5) (2024), 1525–1544.