Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 3, July 2026, pp: 1260-1276 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Clustering Analysis for Anemia Risk Profiling in Hospital Patients Using K-Prototypes Approach

Oki Setiono[1], Zaenal Sugiyanto[1], Dyah Ernawati[1], Arif Kurniadi[1], Ika Pantiawati[1] and Mohamad Nazri Husin[2]

[1] Faculty of Health Science, Universitas Dian Nuswantoro, Semarang, Indonesia
[2] Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, Kuala Nerus, Malaysia

Corresponding author: Oki Setiono (e-mail: okisetiono@dsn.dinus.ac.id), **Author(s) Email**: Zaenal Sugiyanto (e-mail: zaenal.sugiyanto@dsn.dinus.ac.id), Dyah Ernawati (e-mail: dyah_ernawati@dsn.dinus.ac.id), Arif Kurniadi (e-mail: arif.kurniadi@dsn.dinus.ac.id), Ika Pantiawati (e-mail: ikapantia13@dsn.dinus.ac.id), Mohamad Nazri Husin (e-mail: nazri.husin@umt.edu.my)

Abstract **Anemia in inpatients represent a complex clinical condition often associated with systemic responses such as inflammation and infection. However, conventional univariate approaches based solely on hemoglobin (Hb) levels frequently fail to capture the multidimensional nature of anemia risk, particularly in mixed-type clinical datasets. This study aims to develop an anemia risk stratification framework by integrating hematological and demographic variables using the K-Prototypes clustering algorithm. The dataset consisted of 587 patient records collected from multiple hospitals in Central Java, Indonesia, including hemoglobin, leukocytes, platelets, age, and sex. Data preprocessing involved cleaning, standardization, and mixed-data transformation prior to clustering analysis. Multiple cluster configurations (K = 3, K = 4, and K = 5) were evaluated using the Elbow Method and clustering validation metrics, including Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. The results identified K = 4 as the optimal clustering configuration, providing the best balance between cluster separation and interpretability. Subsequently, the identified clusters were clinically interpreted into three risk categories, including High Risk, Moderate Risk, and Normal Risk. The High-Risk group exhibited the lowest mean Hb level (9.27 g/dL), while the Moderate Risk group was characterized by elevated leukocyte counts (19.2k/µL), suggesting distinct hematological patterns. The Normal Risk group demonstrated relatively stable hematological profiles and higher mean age. Statistical testing confirmed significant differences among risk profiles for age, hemoglobin, leukocytes, platelets, and gender (p < 0.001). These findings demonstrate that anemia-related risk patterns are influenced by multidimensional interactions among hematological and demographic factors. The proposed framework provides clinically meaningful patient stratification and has potential applications in electronic medical record systems and clinical decision support environments.**

Keywords **Anemia; K-Prototypes; Mixed-Data Clustering; Risk Profile; Health Informatics**.

I. Introduction

Anemia is a major global health problem that frequently affects inpatients and vulnerable populations, including children, women of reproductive age, and pregnant women. It is commonly associated with nutritional deficiencies, chronic diseases, infections, and inflammatory conditions, leading to impaired cognitive function, reduced productivity, prolonged hospitalization, and increased mortality. In clinical settings, anemia often occurs alongside systemic disorders, highlighting the close relationship between anemia, infection, and inflammation [1], [2], [3], [4]. In clinical practice, anemia rarely occurs as an isolated condition and is often associated with alterations in other hematological parameters such as leukocyte and platelet counts. Elevated leukocyte levels may indicate infection or inflammation, while platelet abnormalities can reflect bone marrow dysfunction or inflammatory responses, providing additional clinical context for anemia risk assessment [5], [6], [7], [8]. Despite this complexity, anemia risk assessment is still predominantly based on hemoglobin thresholds or univariate analysis, which may oversimplify patient conditions and overlook interactions among clinical and demographic factors [9]. As a result, patients with similar hemoglobin levels may present different biological and clinical characteristics depending on variables such as age and sex [10]. Therefore, integrating hematological and demographic data is essential to improve risk stratification accuracy and support more effective clinical decision-making in healthcare settings [11].

Several studies have applied clustering techniques to support anemia detection and health risk

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

stratification. K-Means has been widely adopted because of its simplicity and efficiency in grouping patients based on laboratory parameters. However, most existing approaches rely primarily on numerical variables and fail to capture interactions between clinical and demographic characteristics. Furthermore, the inability of K-Means to directly process categorical variables may reduce the clinical interpretability of the resulting clusters [12], [13], [14], [15].

On the other hand, hospital datasets are inherently composed of mixed data, combining numerical laboratory variables such as hemoglobin, leukocytes, and platelets with categorical attributes such as sex [16]. Ignoring demographic variables may cause bias because anemia risk and normal hemoglobin thresholds differ between males and females [17]. Consequently, integrating both clinical and demographic information is essential for generating more accurate and clinically meaningful patient stratification.

Previous studies have reported that anemia is associated with adverse functional outcomes, including reduced work productivity, emphasizing the need for early identification of at-risk individuals through more comprehensive assessment approaches [18]. To address these limitations, this study employs the K-Prototypes algorithm, which is specifically designed to handle mixed datasets by combining Euclidean distance for numerical variables and simple matching dissimilarity for categorical variables [19], [20]. By integrating hemoglobin, leukocyte, platelet, age, and sex variables, the proposed approach aims to identify more comprehensive anemia risk profiles and reveal patient patterns that may not be detected through conventional diagnostic methods [21].

Although previous studies have applied clustering techniques for anemia detection and health risk stratification, several important limitations remain. First, most anemia-related clustering studies rely primarily on hemoglobin levels or numerical laboratory variables, while demographic characteristics that strongly influence anemia risk, such as sex and age, are often excluded or transformed in a way that reduces their clinical meaning. Second, conventional clustering approaches such as K-Means are limited in handling mixed-type datasets containing both numerical and categorical variables. Third, existing studies rarely evaluate multiple cluster configurations systematically to determine the most clinically meaningful grouping structure. Finally, the resulting clusters are generally presented as mathematical groupings without further interpretation into practical clinical risk categories that can support healthcare decision-making. Unlike previous studies that primarily focus on numerical laboratory parameters or employ conventional clustering methods, this research emphasizes the integration of mixed clinical and demographic data to generate clinically interpretable anemia risk profiles. The novelty of this study does not lie in the K-Prototypes algorithm itself, but in its application for comprehensive anemia risk stratification through the combination of hematological indicators, demographic characteristics, systematic cluster optimization, and clinical interpretation of resulting patient groups. Furthermore, the proposed framework was designed to support future integration into electronic medical record systems and clinical decision support environments, thereby bridging the gap between machine learning analysis and practical healthcare applications.

The main contributions of this study are summarized as follows:

1. **Integration of Mixed Clinical and Demographic Data.** This study developed an anaemia risk profiling framework by integrating haematological variables (haemoglobin, leukocytes, and platelets) with demographic variables (age and sex) within a mixed-data clustering environment. This approach enables a more comprehensive representation of patient characteristics compared with conventional single-parameter analyses.
2. **Systematic Evaluation of Optimal Cluster Structures.** Multiple clustering scenarios (K = 3, K = 4, and K = 5) were systematically evaluated using the cost function and Elbow method to determine the most representative cluster configuration for anaemia risk stratification.
3. **Development of Cluster-Based Anaemia Risk Profiles.** The resulting clusters were transformed into meaningful patient profiles that capture different haematological and demographic characteristics associated with anaemia-related conditions.
4. **Clinical Interpretation of Risk Categories.** The identified clusters were clinically interpreted into Normal, Moderate, and High-Risk categories, enabling more practical utilization of clustering outcomes in healthcare settings.
5. **Potential Integration with Electronic Medical Record Systems.** The proposed framework provides a foundation for future implementation within Electronic Medical Record (EMR) systems as a clinical decision support component for early anaemia risk assessment and patient monitoring.

The structure of this article is systematically organized to provide a comprehensive understanding of the research stages. Following the Introduction section, in which the background and urgency of the study are presented, the Related Works section is intended to review relevant literature on the pathophysiology of anemia and the development of

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

clustering algorithms in healthcare. The Research Methods section was designed to describe in detail the dataset sources, data preprocessing stages, and the working mechanism of the K-Prototypes algorithm in handling mixed-data. The Results section is presented to provide empirical findings, including the characteristics of each cluster and data distribution visualizations. Subsequently, the Discussion section was constructed to interpret these findings from a clinical perspective and to compare them with previous studies. Finally, the study was concluded in the Conclusion section, where the main findings, research limitations, and recommendations for future development are summarized.

II. Method

A. Data Description

The dataset used in this study consists of anonymized hematological examination records collected from five hospitals in Central Java, Indonesia within the period of 2022–2025 (Table 1). An initial dataset of 620 patient records was obtained through random sampling, and after data cleaning, duplicate removal, and exclusion of records that did not meet the predefined inclusion criteria, 587 patient records were retained for analysis. In this case, the selected variables included hemoglobin (Hb), leukocyte count, platelet count, age, and sex, representing both clinical and demographic factors relevant to anemia risk assessment. Hemoglobin values were interpreted according to WHO anemia criteria (<13 g/dL for males and <12 g/dL for females), while leukocyte and platelet counts were evaluated using standard clinical reference ranges of 4–11 ×10³/μL and 150–450 ×10³/μL, respectively, to support the clinical interpretation of the identified risk profiles.

The inclusion criteria applied in this study are as follows:

1. patient records originating from the participating hospitals during the period of 2022–2025;
2. records containing complete information for all selected variables, including hemoglobin level, leukocyte count, platelet count, age, and sex;
3. patients aged between 1 and 93 years; and
4. records with valid laboratory examination results obtained through routine clinical services.

Meanwhile, the exclusion criteria are as follows:

1. records with missing values in one or more selected variables;
2. duplicate patient records identified during the preprocessing stage;
3. records containing inconsistent demographic information, such as missing sex information, invalid age values (e.g., negative age), or discrepancies between demographic attributes recorded within the same patient entry; and
4. records with biologically implausible laboratory values based on clinical reference standards and data quality assessment, including abnormal values that are highly unlikely to occur in clinical practice, such as hemoglobin levels below 2 g/dL or above 25 g/dL, leukocyte counts exceeding 300,000/μL, or platelet counts exceeding 2,000,000/μL.

Several limitations should be acknowledged. The dataset did not include ferritin, C-reactive protein (CRP), ICD-based diagnoses, nutritional status, or pregnancy status. Therefore, the resulting clusters represent anemia-related risk patterns derived from the available hematological and demographic variables rather than confirmed clinical diagnoses. Future studies should incorporate these variables to improve clinical interpretability and risk stratification performance.

B. Data Collection

The data used in this study consisted of retrospective clinical records obtained from medical

Table 1. Characteristics of the Multi-Hospital Patient Dataset.

Item	Description
Final dataset size	587 patient records
Hospital source	• RSUD Brebes, Central Java (Type B) • RS Bhayangkara, Semarang (Type C) • RSUD DR. H Soewondo Kendal, Kendal (Type B) • RS Roemani Muhammadiyah, Semarang (Data collected in 2024) (Type C) • RS Roemani Muhammadiyah, Semarang (Data collected in 2023) (Type C)
Data period	2022 - 2025
Sampling technique	Random sampling
Age range	1 – 93 years
Numerical variables	Hemoglobin, Leukocyte, Platelet, Age
Categorical variable	Sex
Total variables	5

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

record examinations conducted in several hospitals in Central Java, Indonesia. Ethical approval was granted by the respective Health Research Ethics Committees (Komisi Etik Penelitian Kesehatan, KEPK) prior to data collection, and the study was conducted in accordance with the applicable ethical regulations and the Declaration of Helsinki [22]. All patient records were anonymized before analysis to ensure confidentiality and privacy. No personally identifiable information was included in the dataset. As this study utilized anonymized retrospective secondary data, the requirement for individual informed consent was waived by the respective ethics committees. The dataset was obtained from hospital laboratory databases covering the period from 2022 to 2025 across five collection phases. Random sampling was applied to select eligible records from the available dataset. The collected variables included demographic characteristics (age and sex) and key hematological parameters, namely hemoglobin (Hb), leukocyte count, and platelet count, which are clinically relevant for anemia risk analysis. A total of 587 patient records were retained for analysis and extracted from the digitized laboratory information systems of participating hospitals. As illustrated in Fig. 1, the study followed a systematic workflow beginning with multi-hospital data acquisition, followed by ethical approval, de-identification, and initial screening procedures to ensure data confidentiality and compliance with healthcare research standards. The dataset was subsequently filtered according to predefined inclusion and exclusion criteria [23]. Incomplete, duplicate, and inconsistent records were removed, resulting in a validated dataset comprising 587 observations. The final dataset was stored in a structured digital format to facilitate preprocessing activities, including data cleaning, normalization, feature preparation, and subsequent K-Prototypes clustering analysis.

C. Clinical Risk Interpretation

Although the K-Prototypes algorithm generated four statistical clusters under the optimal configuration (K = 4), these clusters were subsequently translated into three clinically interpretable risk categories based on their mean hemoglobin values. The cluster with the lowest mean hemoglobin level was classified as High Risk, the cluster with the highest mean hemoglobin level was classified as Normal Risk, while the remaining intermediate clusters were combined into the Moderate Risk category. This mapping strategy was implemented to improve clinical interpretability while maintaining the original clustering structure identified by the algorithm. Following the clustering process, clinical risk categories were determined according to the mean hemoglobin (Hb) level of each cluster. Clusters with the lowest, intermediate, and highest mean Hb values were subsequently labeled as High Risk, Moderate Risk, and Normal Risk, respectively.

D. Data Preprocessing and Normalization

The data preprocessing stage was initiated with the cleaning and validation of the dataset to ensure the integrity of the clinical information to be processed. As illustrated in Table 2 **(Proposed Method: Patient Hematological Risk Clustering Flow)**, this stage is a part of a structured pipeline that begins from raw data acquisition and continues through preprocessing, feature engineering, clustering, and risk interpretation.

Considering that the data were obtained from multiple hospitals, the first step involved synchronizing numerical formats and handling inconsistencies in measurement units and decimal separators.

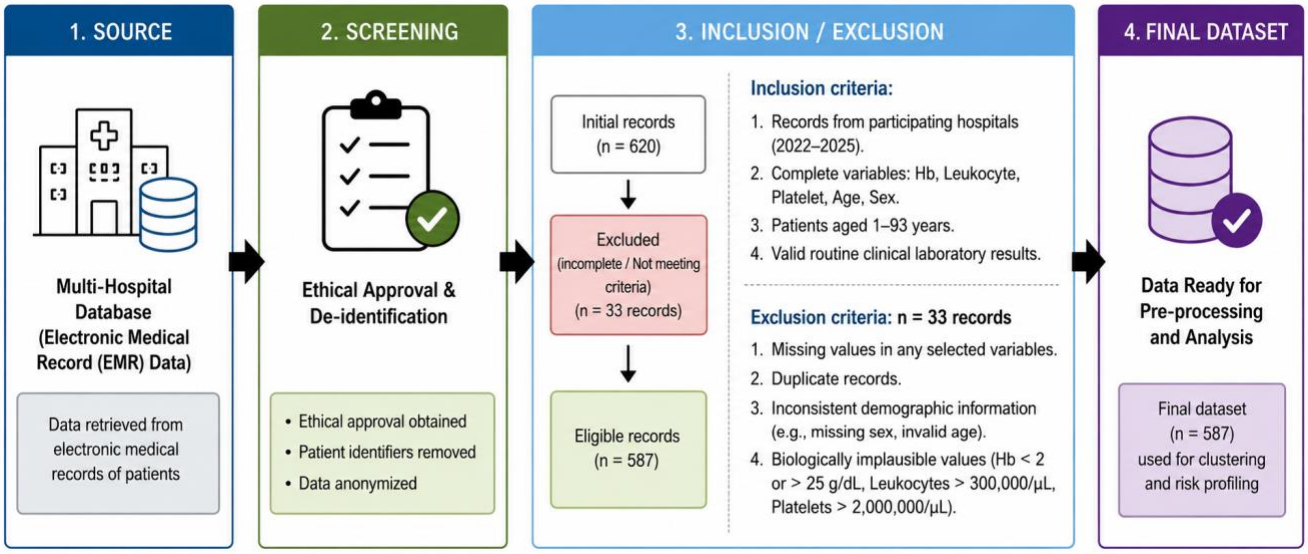

Fig. 1. Data collection, ethical screening, eligibility assessment, and final dataset selection for anemia risk profiling.

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

The entire data processing workflow was implemented in Python using the Google Colab environment. Data preprocessing and manipulation were performed using Pandas and NumPy, numerical variables were standardized using Scikit-learn, and clustering analysis was conducted using the K-Prototypes algorithm from the kmodes library. The clustering model was initialized using the Cao method ($init$='Cao') with five independent runs (n_init=5) to improve clustering stability and reduce sensitivity to centroid initialization. A fixed random seed ($random\_state$=42) was applied to ensure reproducibility of the results, while other parameters were maintained at their default library settings. Clustering experiments were performed using K values of 3, 4, and 5, and the optimal cluster configuration was determined using the Elbow Method. Data visualization was performed using Matplotlib and Seaborn [24].

Data containing missing values in key variables such as hemoglobin, leukocytes, and platelets were systematically removed to avoid computational bias, while the categorical variable of sex was standardized into binary labels, where the value 0 represents females and the value 1 represents males. Furthermore, as depicted in Fig. 2, the clustering stage employed the K-Prototypes algorithm [25] by evaluating multiple cluster scenarios, specifically K = 3, K = 4, and K = 5, to explore variations in data partitioning and to ensure a robust assessment of cluster structure [26]. These variations were considered to identify the most representative grouping pattern for mixed-type clinical data. The resulting clusters were then mapped into three interpretable risk categories: Normal, Moderate Risk, and High Risk, to facilitate clinically meaningful interpretation and support decision-making processes [27]. After ensuring the dataset was complete and free of missing values, numerical features were standardized using Z-score normalization to achieve scale equivalence among variables [11]. This step is necessary due to significant

Table 2. The Proposed Method Patient Hematological Risk Clustering Flow.

STEP	INPUT	PROCESS	OUTPUT
1. Data Collection	• Hospital systems • Laboratory records	Collecting patient data and extracting relevant variables	Raw Data (e.g., Age, Gender, Hb, Leukocytes, Platelets)
2. Inclusion / Exclusion Criteria	• Hematological parameters • Age and Gender	• Applying inclusion criteria • Applying exclusion criteria	Eligible Dataset (after applying criteria)
3. Data Cleaning	• Eligible dataset	• Handling missing values • Checking data consistency • Removing outliers / implausible values	Cleaned Dataset (high-quality data)
4. Normalization	• Cleaned dataset	• Standardizing numerical variables (Z-Score) • Categorical variable (Gender) remains categorical	Normalized Dataset (numerical variables standardized)
5. K-Prototypes Clustering	• Normalized dataset • Parameter settings (K = 3, 4, 5)	• Applying K-Prototypes algorithm • Testing multiple K values (K = 3, 4, 5)	Cluster Solutions (for multiple K values)
6. Cluster Validation	• Cluster solutions	• Evaluating clustering quality using internal metrics • Computing DBI, Silhouette, Calinski-Harabasz Index	Optimal K Selection (best cluster configuration)
7. Clinical Interpretation	• Optimal cluster configuration	• Interpreting clusters based on hematological & demographic characteristics • Analyzing patterns & clinical relevance	Clinically Interpretable Clusters (meaningful patient groupings)
8. Risk Profiling	• Interpretable clusters	• Assigning risk levels (High, Moderate, Normal) • Summarizing hematological & demographic profiles	Anemia Risk Profiles (risk categories with profile characteristics)
9. EMR / CDSS Integration	• Anemia risk profiles	• Integrating risk profiles into EMR • Enabling CDSS for risk stratification and follow-up	EMR/CDSS Integration (real-time clinical application)

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

differences in hematological value ranges, where hemoglobin was measured on a small scale, while leukocyte and platelet counts can reach much higher magnitudes. As shown in Fig. 2, these scale disparities could bias the clustering process by causing variables with larger values to disproportionately influence cluster formation. Therefore, standardization was applied to ensure that each clinical parameter contributed proportionally to the clustering model. The normalization process begins by calculating the mean μ of each numerical feature, as expressed in Eq. (1), followed by the computation of the standard deviation σ in Eq. (2). The mean represents the central tendency of the data, whereas the standard deviation measures the dispersion of observations around the mean. Subsequently, each observation was transformed into a standardized score using the Z-score formulation shown in Eq. (3), resulting in variables with a mean of zero and a standard deviation of one [28].

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

In this equation, μ, x_i, n denote the mean of the variable, the observation at index i, and the total number of observations, respectively. The mean was calculated to determine the central tendency of each numerical feature before normalization [29].

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} \quad (2)$$

In this equation, σ, x_i, μ, and n, represent the standard deviation, observation value, mean of the variable, and total number of observations, respectively.

Standard deviation measures data dispersion and is

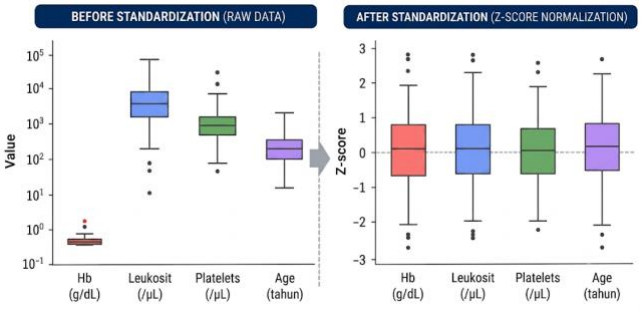

Fig. 2. Distributions of numerical features before and after Z-score standardization.

required for Z-score normalization [30].

$$Z = \frac{x - \mu}{\sigma} \quad (3)$$

In this equation, Z denotes the standardized (Z-score) value, x is the original observed clinical value, μ represents the mean of the corresponding numerical feature, and σ is the standard deviation, indicating the variability of the data.

E. Clustering-Based Risk Profiling and Statistical Analysis

To identify latent patterns within mixed clinical and demographic data, this study employed the K-Prototypes clustering algorithm. The method was selected because it can simultaneously process numerical and categorical attributes, making it suitable for medical datasets containing hematological measurements and demographic characteristics. Unlike K-Means, which is limited to numerical variables, K-Prototypes integrates distance measures for both data types into a unified clustering framework [31]. To identify latent patient profiles from mixed clinical and demographic data, the K-Prototypes algorithm was employed. Numerical variables were measured using Squared Euclidean Distance Eq. (4) [32].

$$d_n(x,y) = \sum_{j=1}^{m}(x_i - y_j)^2 \quad (4)$$

In this equation, $d_n(x,y)$ denotes the numerical distance between objects x and y, while x_i and y_j represent the corresponding numerical attributes of objects x and y, respectively, and m is the total number of numerical attributes. This distance measures similarity among numerical clinical variables. While categorical variables were evaluated using Simple Matching Dissimilarity Eq. (5)

$$\delta(x_i, y_j) = \begin{cases} 0, x_i = y_i \\ 1, x_j \neq y_j \end{cases} \quad (5)$$

In this equation, $\delta(x_i, y_j)$ denotes the categorical dissimilarity, where a value of 0 indicates identical categories and a value of 1 indicates different categories. This measure was applied to the gender attribute. The K-Prototypes formula in Eq. (6) is as follows [31]:

$$d(x,y) = \sum_{j=1}^{p}(x_{j-} y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j) \quad (6)$$

In this equation, $d(x,y)$ denotes the total dissimilarity between objects x and y. $\sum_{j=1}^{p}(x_{j-} y_j)^2$ represents the squared Euclidean distance for the p numerical features, γ is the weighting factor balancing the contributions of numerical and categorical variables, $\sum_{j=p+1}^{m}\delta(x_j, y_j)$ denotes the simple matching dissimilarity for the categorical features, and $\delta(x_j, y_j)$ is the Kronecker delta function, which equals 0 when $x_j = y_j$ and 1 otherwise.

$$C_i = \arg min_k \left(\sum_{j \in num} (x_{ij} - \mu_{kj})^2 + \gamma \sum_{j \in cat} \sigma\ (x_{ij} - \mu_{kj}) \right) \quad (7)$$

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Eq. (7) [33], Assignment rule was used to assign each data point x_i into cluster C_k that produces the minimum dissimilarity value. In this equation, x_{ij} represents the value of the j feature of data point i, while μ_{kj} denotes the centroid of cluster k. The term γ is used as a weighting factor for categorical attributes, and $\delta(.)$ indicates the simple matching function between categorical values. The clusters were determined by minimizing the combined numerical and categorical distance. The weighting coefficient γ balances the contribution of numerical and categorical attributes. The clustering process minimizes the objective function in Eq. (8) [34]:

$$E = \sum_{i=1}^{n}\sum_{k=1}^{K} u_{ik} d(x_i, c_k) \quad (8)$$

In this equation, E denotes the clustering cost, u_{ik} is the membership indicator, x_i represents observation i, c_k is the centroid of cluster k, K is the total number of clusters with numerical and categorical centroids updated according to Eq. (9) and Eq. (10) [35], respectively.

$$c_K = \frac{1}{n_k} \sum_{x_i \in C_k} x_i \quad (9)$$

In this equation, c_K , n_k, and C_k represent the centroid, the number of observations, and cluster k, respectively.

Eq. (10):

$$c_k^{cat} = Mode(x_i) \quad (10)$$

In this equation, c_k^{cat} and $Mode(x_i)$ represent the categorical centroid and the most frequent category in cluster k.

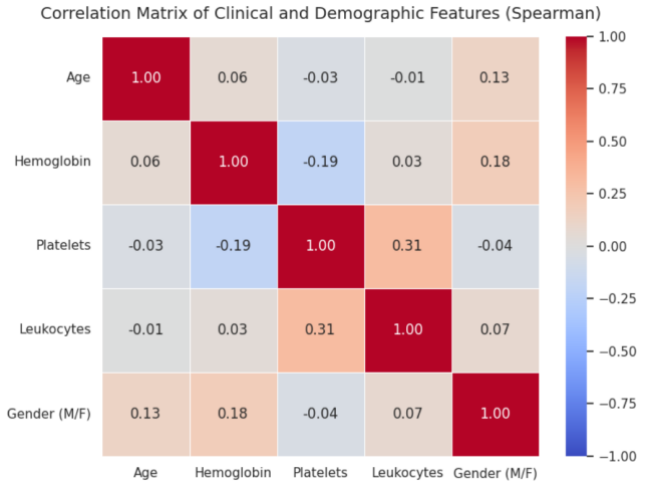

Fig. 3. Spearman correlation matrix of the clinical and demographic variables used for clustering.

In this study, γ was automatically estimated using the default mechanism of the kmodes library; therefore, no additional tuning was performed. Fig. 3 presents the feature correlation matrix among hemoglobin (Hb), leukocytes, platelets, age, and gender. Gender was encoded as a binary variable (Female = 0, Male = 1) prior to the correlation analysis and machine learning modeling. Consequently, the sign of the Spearman correlation coefficient involving sex should be interpreted according to this coding scheme, where positive coefficients indicate greater association with male participants and negative coefficients indicate greater association with female participants. The results show generally low to moderate correlations, indicating that each variable contributes complementary information to the dataset. Since no strong multicollinearity ($r >$ 0.80) was observed, all variables were retained for clustering analysis [36]. Cluster quality was evaluated using the Silhouette Coefficient Eq. (11) [37]:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (11)$$

In this equation, $s(i)$ denotes the silhouette value of observation i, $a(i)$ represents the average distance between observation i and all points within the same cluster, and $b(i)$ is the average distance between observation i and the nearest neighboring cluster. The Silhouette Coefficient evaluates how well an observation fits within its assigned cluster relative to neighboring clusters. Values close to 1 indicate strong cluster membership, whereas values close to −1 indicate possible misclassification. Then, Eq. (12) shows the Average Silhouette Score [37]:

$$S = \frac{1}{n}\sum_{i=1}^{n} s(i) \quad (12)$$

In this equation, S denotes the average silhouette score, n is the total number of observations, and $s(i)$ represents the silhouette coefficient of observation i.

SSE (Sum of Squared Errors), Eq. (13) [38]:

$$SSE = \sum_{k=1}^{K}\sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad (13)$$

SSE was used to measure the compactness of clusters by summing the squared distance between each data point and its cluster centroid. In this equation, x_i represents the i-th data point, μ_k denotes the centroid of cluster k, and K is the total number of clusters. A lower SSE indicates that data points are more closely grouped within their clusters. TSS (Total Sum of Squares), Eq. (14) [38]:

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 3. Clinical Characteristics of Risk Profiles for the Selected K = 4 Configuration.

Variable	High Risk (n = 155)	Middle Risk (n = 217)	Normal (n = 187)
Age (years)	41.72 ± 16.97	26.53 ± 18.36	60.00 ± 11.07
Hemoglobin, Hb (g/dL)	9.27 ± 2.52	11.54 ± 1.91	12.13 ± 1.99
Leukocytes (×10³/µL)	14.94 ± 7.92	19.19 ± 16.95	10.67 ± 5.48
Platelets (×10³/µL)	498.35 ± 121.49	293.21 ± 96.34	279.78 ± 101.73
Sex, Male/Female	64 / 91	130 / 87	122 / 65

$$TSS = \sum_{i=1}^{n} \| x_i - \bar{x} \|^2 \quad (14)$$

TSS represents the total variance of the dataset before clustering. Here, x_i denotes the i-th data point, $\bar{x}$ is the overall mean of the dataset, and n is the total number of data points. This metric was used as a baseline to evaluate how much variance was reduced after clustering. Davies–Bouldin Index, Eq. (15) [37]:

$$DB = \frac{1}{K} \sum_{i=1}^{K} Max_{j \neq 1} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (15)$$

In this equation, DB denotes the Davies–Bouldin Index, S_i and S_j represent the intra-cluster dispersion of clusters i and j, respectively, M_{ij} is the distance between the centroids of clusters i and j, and K is the total number of clusters.

Lower Davies–Bouldin values indicate better clustering quality because they represent smaller within-cluster dispersion and greater separation between clusters. Calinski–Harabasz Index, Eq. (16) [37]:

$$CH = \frac{Tr(B_k)/(K-1)}{Tr(W_k)/(n-K)} \quad (16)$$

In this equation, CH denotes the Calinski–Harabasz Index, $Tr(B_k)$ and $Tr(W_k)$ represent the between-cluster and within-cluster dispersions, respectively, between-cluster dispersion, K is the total number of clusters, and n is the total number of observations. Higher Calinski–Harabasz values indicate better-defined clusters characterized by high between-cluster separation and low within-cluster variance. These metrics were jointly used to determine the optimal cluster configuration by assessing cluster compactness and separation.

Finally, inferential statistical analysis was conducted to examine differences among the identified risk profiles. Data normality was assessed using the Shapiro–Wilk test, followed by the Kruskal–Wallis H-test Eq. (17) [39]:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \quad (17)$$

In this equation, H denotes the Kruskal–Wallis test statistic, N is the total number of observations across all groups, k represents the number of groups (clusters), R_i is the sum of ranks for group i, and n_i is the number of observations in group i. For non-parametric comparisons. Significant results were further analyzed using Dunn's post-hoc test with Bonferroni correction Eq. (18) [40]:

$$Z = \frac{R_i - R_j}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \quad (18)$$

In this equation, Z denotes Dunn's test statistic, R_i and R_j represent the mean ranks of groups i and j, respectively, n_i and n_j are the numbers of observations in groups i and j, respectively, and N = is the total number of observations. The association between gender distribution and risk profiles was evaluated using the Chi-Square test Eq. (19) [39]:

$$X^2 = \sum \frac{(O-E)^2}{E} \quad (19)$$

In this equation, X^2 denotes the Chi-Square test statistic, O represents the observed frequency, and E represents the expected frequency with expected frequencies calculated using Eq. (20):

$$E = \frac{(Row\ Total)(Column\ Total)}{Grand\ Total} \quad (20)$$

In this equation, E represents the expected frequency, $Row\ Total$ and $Column\ Total$ denote the total observations in the corresponding row and column categories, respectively, and $Grand\ Total$ is the total number of observations in the contingency table. Statistical significance was defined as $p < 0.05$ [39], [41].

F. Proposed Clustering Algorithm

Algorithm 1 summarizes the proposed workflow for anemia risk profiling using the K-Prototypes clustering

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

algorithm. The process begins with loading the patient dataset, followed by the application of inclusion criteria, duplicate removal, and missing value handling. Numerical variables are standardized using Z-score normalization, while categorical variables are retained for mixed-data clustering. The K-Prototypes algorithm is then executed with K = 3, K = 4, and K = 5. Clustering performance is evaluated using the Cost function, Silhouette Coefficient, Davies–Bouldin Index, and Calinski–Harabasz Index to determine the optimal configuration. Finally, the selected clusters are clinically interpreted and assigned to anemia risk profiles, producing the final risk categories.

Algorithm 1. Anemia Risk Profiling Using K-Prototypes

Input:
Patient dataset
Output:
Risk profile
1. Loading the dataset
2. Applying the inclusion criteria
3. Removing the duplicates
4. Handling the missing values
5. Normalizing the numerical features
6. Defining the categorical variables
7. Running the K-Prototypes for K=3,4,5
8. Computing the Cost
9. Evaluating the Silhouette, DBI, CHI
10. Selecting optimal K
11. Generating clusters
12. Applying clinical interpretation
13. Assigning risk profile
14. Producing final anemia risk categories

III. Result

A. Clustering Convergence and Optimal K-Value

Before profile interpretation, the convergence and stability of the clustering algorithm were evaluated using the cost function (Inertia), Silhouette Score, Davies–Bouldin Index (DBI), and Calinski–Harabasz Index (CHI) across multiple cluster configurations (K = 3–5) [26]. As K increased, the cost function decreased (1595.08 to 1120.65), while the Silhouette Score (0.200–0.226) and CHI (128.96–148.93) increased, and the DBI decreased (1.619–1.229), indicating improved clustering quality. Although K = 5 achieved slightly better validation metrics, K = 4 was selected as the optimal configuration based on the balance between clustering performance and clinical interpretability. To further characterize the input variables prior to cluster interpretation, pairwise associations were examined using Spearman's rank correlation Fig. 3. Spearman's method was selected because the dataset included binary and non-normally distributed variables. Sex was encoded as a binary variable (Female = 0, Male = 1), and correlations involving sex should therefore be interpreted according to this coding scheme. All absolute Spearman correlation coefficients were below 0.70, indicating no evidence of strong multicollinearity among the variables used for clustering.

Furthermore, the clustering validation results were summarized for K = 3–5. The Elbow method identified K = 4 as the optimal configuration, supported by a Silhouette Score of 0.221, a DBI of 1.339, and a CHI of 148.020, indicating a balanced clustering solution with good compactness, separation, and interpretability.

Table 3 presents the clinical characteristics of the identified risk profiles. The High-Risk group exhibited the lowest mean hemoglobin level (9.27 g/dL) and the highest mean platelet count (498 ×10³/µL), indicating a greater likelihood of anemia-related abnormalities. The Moderate Risk group showed the highest mean leukocyte count (19.19 ×10³/µL), suggesting the presence of inflammatory or infection-related hematological conditions. In contrast, the Normal Risk group demonstrated the highest mean hemoglobin level (12.13 g/dL) and relatively stable hematological parameters, representing the most favorable clinical profile among the identified groups.

Furthermore, Gender distribution varied across the identified risk profiles. The High Risk group was predominantly female, accounting for 58.7% of patients,

Table 4. Kruskal–Wallis, Dunn's Post-Hoc, and Chi-Square Test Results Across Risk Profiles.

Variable	Test Statistic (H/χ^2)	Overall p-value	HR–MR	HR–NR	MR–NR
Age	249.7	<0.001	<0.001*	<0.001*	<0.001*
Hemoglobin	120.7	<0.001	<0.001*	<0.001*	0.011*
Leukocytes	36.7	<0.001	1.000	<0.001*	<0.001*
Platelets	240.1	<0.001	<0.001*	<0.001*	1.000
Sex	21.4	<0.001	—	—	—

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

while males comprised 41.3%. In contrast, the Moderate Risk group was dominated by males (59.9%), with females representing 40.1%. A similar pattern was observed in the Normal Risk group, where males constituted 65.2% and females 34.8%. These differences indicate that sex may contribute to the variation in anemia risk profiles, with females appearing more frequently in the High Risk category and males predominating in the Moderate and Normal Risk groups.

Abbreviations: HR = High Risk; MR = Middle Risk; NR = Normal Risk; Hb = Hemoglobin. H denotes the Kruskal–Wallis test statistic for continuous variables, whereas χ^2 denotes the Chi-square test statistic for the categorical variable (Sex). Indicates statistical significance ($p < 0.05$). Dunn's post-hoc test was performed only for continuous variables that showed significant overall differences. Table 4 summarizes the inferential statistical analysis and post-hoc comparisons among the identified risk profiles. The Kruskal–Wallis test revealed significant differences for Age, Hemoglobin, Leukocytes, and Platelets (all $p < 0.001$). Post-hoc Dunn's tests with Bonferroni correction showed that Age and Hemoglobin differed significantly across all pairwise comparisons. For Leukocytes, significant differences were observed between the Normal Risk group and both the High Risk and Moderate Risk groups, whereas no significant difference was found between High Risk and Moderate Risk. For Platelets, significant differences were detected between the High-Risk group and the other two groups, while Moderate Risk and Normal Risk showed comparable platelet distributions. In addition, the Chi-Square test indicated a significant association between gender and risk profile ($\chi^2 = 21.431$, $p < 0.001$), suggesting that sex contributes to the heterogeneity of the identified anemia risk categories.

B. Clinical and Demographic Profiling Analysis

The following table presents the clinical and demographic characteristics of each cluster, providing an overview of the distinct patterns identified in the data.

Table 5 shows that different cluster configurations (K = 3, K = 4, and K = 5) produce varying levels of clinical interpretability. The K = 3 configuration tended to group heterogeneous patient profiles into broader categories, whereas K = 5 resulted in cluster fragmentation and reduced interpretability. The K = 4 configuration provided the most balanced separation of hematological patterns, with the High-Risk group exhibiting the lowest mean Hb level (9.27 g/dL), the Moderate-Risk group showing elevated leukocyte counts (19.2 ×10³/μL), and the Normal-Risk group maintaining relatively stable hematological parameters (Hb = 12.13 g/dL). Although K = 4 generated four statistical clusters, these were subsequently mapped into three clinically interpretable risk categories based on mean hemoglobin levels. Therefore, K = 4 was selected as the optimal configuration due to its balance between cluster differentiation, data distribution, and clinical relevance for anemia risk stratification.

Table 5. Comparison of Clinical and Demographic Risk Profiles Across K = 3, K = 4, and K = 5 Configurations.

Risk Profile	Variable	K = 3	K = 4	K = 5
High Risk	Age (years)	27.71	41.72	45.82
	Hb (g/dL)	9.77	9.27	7.89
	Platelets (×10³/μL)	358	498	226
	Leukocytes (×10³/μL)	11.5	14.9	8.2
	Gender (Male/Female)	103 / 142	64 / 91	30 / 42
Moderate Risk	Age (years)	45.57	26.53	32.26
	Hb (g/dL)	10.87	11.54	11.06
	Platelets (×10³/μL)	381	293	397
	Leukocytes (×10³/μL)	43.2	19.2	18.7
	Gender (Male/Female)	43 / 20	130 / 87	182 / 147
Normal Risk	Age (years)	54.92	60.00	60.33
	Hb (g/dL)	12.48	12.13	12.68
	Platelets (×10³/μL)	324	280	293
	Leukocytes (×10³/μL)	11.7	10.7	11.0
	Gender (Male/Female)	170 / 81	122 / 65	104 / 54

C. Cluster Distribution and Feature Interaction

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

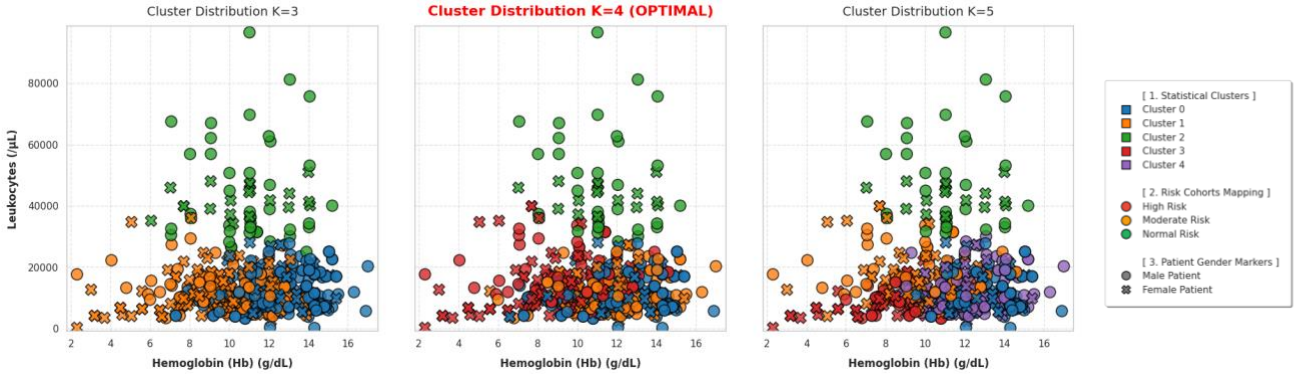

Fig. 4. Distribution of hemoglobin and leukocyte values across K = 3, K = 4, and K = 5 clustering configurations, with clinically interpreted risk categories and sex-specific markers.

The visualization of data distribution using a scatter plot FIG. 4 is shown to demonstrate a strong interaction between Hemoglobin and Leukocyte variables in distinguishing risk profiles. The cluster distribution plots illustrate the interaction between hemoglobin and leukocyte levels across different clustering configurations K = 3, K = 4, and K = 5 highlighting how cluster structure evolves with increasing granularity. In the K = 3 configuration, the clusters appear more overlapped, particularly in the moderate range of hemoglobin values, where cases with extreme leukocyte elevations were grouped together, limiting clear clinical differentiation. The K = 4 configuration shows a more distinct separation, where clusters become more compact and better differentiated, especially in identifying patients with elevated leukocyte levels that indicate inflammatory or infectious conditions, while maintaining a consistent grouping for low hemoglobin cases. In contrast, the K = 5 configuration introduces additional fragmentation, where clusters become more dispersed and less distinct, reducing interpretability. Overall, the visualization demonstrates that K = 4 provides the most balanced representation of feature interaction, effectively separating anemia-related patterns from inflammation-driven profiles, while preserving cluster clarity and cohesion. From a medical informatics perspective, the proposed framework has the potential to be integrated into Electronic Medical Record (EMR) systems and Clinical Decision Support Systems (CDSS). Routine hematological and demographic data, including hemoglobin, leukocyte count, platelet count, age, and sex, can be automatically processed by the K-Prototypes model to generate patient-specific risk labels. These outputs may be visualized through clinical dashboards to assist healthcare professionals in identifying high-risk patients, prioritizing interventions, and improving patient monitoring workflows. Furthermore, the framework can operate as a secondary risk-screening layer within existing EMR infrastructures without requiring additional laboratory examinations, since all input variables were routinely collected during standard hematological assessments.

IV. Discussion

A. Interpretation of Clustering Results in Clinical Context

The evaluation of multiple cluster configurations (K = 3, K = 4, and K = 5) demonstrates that determining an optimal number of clusters is essential for achieving meaningful patient segmentation. Although the cost function consistently decreases as the number of clusters increases, the improvement becomes marginal beyond K = 4, indicating an optimal balance between cluster quality and model complexity. In the K = 3 configuration, extreme leukocyte values tend to be grouped into a single cluster, limiting the differentiation of clinical conditions, whereas the K = 5 configuration introduces excessive fragmentation that reduces representativeness. In contrast, the K = 4 configuration provides clearer separation and more compact cluster structures, enabling the distinction of anemia-related and inflammation-associated hematological patterns.

Based on this configuration, the resulting clusters were clinically interpreted using established hematological reference ranges, where the High Risk category was assigned to clusters with the lowest hemoglobin levels and the greatest deviation from normal values, the Moderate Risk category represented intermediate hematological profiles with moderate hemoglobin levels or elevated leukocyte and platelet counts, and the Normal Risk category included clusters whose hematological parameters were generally closer to standard clinical reference ranges. This approach ensures that the resulting risk profiles are not only statistically robust but also clinically meaningful and interpretable.

B. Comparison with Prior Studies

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

As summarized in Table 6, the proposed framework shares several methodological similarities with previous studies while offering important clinical and analytical contributions. Similar to K-Prototypes-based studies [19], [31], the proposed approach clusters mixed numerical and categorical clinical data to identify meaningful patient groups. Likewise, the MSKP framework [20] demonstrates the importance of integrating heterogeneous medical data for healthcare analysis, supporting the suitability of mixed-data clustering methods in clinical applications. In contrast, studies using K-Means [13], [14] primarily relied on numerical data for disease stratification or population vulnerability analysis, whereas the present study integrates both hematological parameters and demographic characteristics to better capture patient heterogeneity. Furthermore, unlike the work of Aschenbruck et al. [19], which focused on improving K-Prototypes initialization strategies, this study emphasizes the development of clinically interpretable anemia risk profiles supported by comprehensive cluster validation using the Silhouette Coefficient, Davies–Bouldin Index, and Calinski–Harabasz Index, followed by statistical significance testing using the Kruskal–Wallis, Dunn's post-hoc, and Chi-square tests. Compared with the hypertension risk profiling study in [31], the proposed framework specifically targets anemia risk stratification using routine hematological indicators collected from multiple hospitals, providing practical support for early risk assessment and potential integration into electronic medical record (EMR) and clinical decision support system (CDSS) environments.

C. Clinical Implications for Healthcare Providers

The K-Prototypes clustering model provides practical value for hospital information systems by supporting automated clinical decision-making through early risk detection directly from electronic medical records, enabling faster intervention for high-risk patient profiles. It also improves operational efficiency by enhancing triage and bed management, allowing prioritization of patients based on identified risk phenotypes and reducing bottlenecks in emergency and intensive care services. Additionally, the study underscores the importance of data quality governance, as initial inconsistencies in laboratory

Table 6. Methodological Comparison Between the Proposed Anemia Risk-Profiling Framework and Previous Studies

Ref	Method	Similarity	Difference
[13]	K-Means	Applied unsupervised clustering to stratify patients using clinical numerical data.	K-Means on pediatric asthma data; this study used K-Prototypes on mixed hematological data for anemia risk profiling.
[14]	K-Means	Employed clustering to classify population risk based on observed characteristics.	Early-marriage vulnerability mapping; this study focuses on hospital-based anemia risk profiling.
[19]	K-Prototypes	Applied K-Prototypes to cluster mixed numerical and categorical healthcare data.	Focuses on initialization strategies; this study develops clinically interpretable anemia risk profiles and validates cluster quality.
[20]	MSKP	Utilized mixed-source medical data for clustering-based healthcare analysis.	Proposes an enhanced K-Prototypes algorithm; this study applies standard K-Prototypes for multi-hospital anemia risk profiling.
[26]	Q-learning	Adopted a data-driven approach to support intelligent decision-making.	Reinforcement learning for Federated Edge Cloud optimization; this study performs unsupervised patient clustering for anemia risk assessment.
[31]	K-Prototypes	Used K-Prototypes to identify disease-related risk groups from mixed clinical data.	Hypertension risk detection; this study targets anemia risk profiling using hematological indicators.
This Study	K-Prototypes	Clustered mixed hematological and demographic variables from multiple hospitals.	Multi-hospital anemia risk profiling with cluster validation and statistical significance analysis for potential EMR/CDSS integration.

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

units highlight the need for standardized data entry and automated validation to ensure reliable secondary data use for clinical, research, and interoperability purposes.

D. Limitations of the Study

Despite its strong validation and clinically meaningful risk stratification, this study has several limitations. The analysis was based on retrospective clinical data and relied primarily on routinely available hematological and demographic variables. Important clinical factors associated with anemia etiology, including ferritin, C-reactive protein (CRP), ICD-based diagnoses, nutritional status, pregnancy status, medication history, and clinical outcome data, were not available in the dataset. In addition, potential data quality issues inherent to electronic health record (EHR) systems may affect data completeness and consistency. Although the dataset was collected from multiple hospitals in Central Java, external validation using independent datasets from other healthcare institutions has not yet been conducted, which may limit the generalizability of the identified risk profiles across different healthcare settings. Future studies should incorporate additional clinical biomarkers and perform multi-center external validation to improve the robustness and clinical applicability of the proposed framework.

V. Conclusion

This study aimed to develop an anemia risk profiling model by integrating mixed clinical data, including numerical hematological parameters (hemoglobin, leukocytes, and platelets) and the demographic variable sex, using the K-Prototypes clustering approach. The results identified K = 4 as the optimal clustering configuration, providing a balance between cluster quality and clinical interpretability. Under this configuration, the High-Risk group exhibited the lowest mean hemoglobin level (9.27 g/dL) and was predominantly female, the Moderate-Risk group showed elevated leukocyte counts (19.2 × 10^3/μL) with intermediate hematological characteristics, and the Normal-Risk group demonstrated relatively stable hematological profiles (Hb = 12.13 g/dL). These findings indicate that anemia risk is influenced by the interaction of hematological and demographic factors rather than hemoglobin levels alone, highlighting the potential of mixed-data clustering for more comprehensive risk stratification and future integration into electronic medical record–based clinical decision support systems. Nevertheless, this study was limited by the absence of additional clinical variables such as ferritin, C-reactive protein (CRP), ICD diagnosis, nutritional status, and pregnancy status, which may influence anemia interpretation. Future research should incorporate richer clinical data, such as NLP-derived notes, ICD-10 codes, and patient outcomes, to improve prognostic risk profiling. Multi-center validation and HL7 FHIR-compliant implementation are also needed to enhance model generalizability, interoperability, and real-time clinical application across diverse healthcare systems.

Acknowledgment

The authors gratefully acknowledge Universitas Dian Nuswantoro, the Institute for Research and Community Services of Universitas Dian Nuswantoro, the One Family Health research group, the participating hospitals, and the respective ethics committees for their administrative, technical, and institutional support during this study.

Funding

This research was supported by the Internal Research Grant of Universitas Dian Nuswantoro through the Institute for Research and Community Services of Universitas Dian Nuswantoro. The funder had no role in the study design, data collection, data analysis, interpretation of results, preparation of the manuscript, or decision to submit the article for publication.

Data Availability

The dataset used in this study contains anonymized hospital laboratory records and is subject to institutional privacy and ethical restrictions. The data may be obtained from the corresponding author upon reasonable request, subject to approval from the participating hospitals and relevant ethics committees.

Code Availability

The source code used for data preprocessing, K-Prototypes clustering, statistical analysis, and visualization is available from the corresponding author upon reasonable request. A permanent public repository should be provided after publication whenever institutional and data-governance requirements permit.

Author Contribution

Oki Setiono contributed to the conceptualization, methodology, data analysis, implementation of the K-Prototypes algorithm, visualization, and preparation of the original manuscript. Zaenal Sugiyanto and Dyah Ernawati contributed to clinical validation, interpretation of hematological findings, and manuscript review. Arif Kurniadi contributed to data processing, methodology development, and technical validation. Ika Pantiawati contributed to public health interpretation and manuscript review. Mohamad Nazri Husin provided supervision, methodological advice, and critical revision of the manuscript. All authors reviewed and approved the final version of the manuscript.

Declarations

Ethical Approval

This study was conducted in accordance with the Declaration of Helsinki and applicable institutional ethical standards. Ethical approval was obtained from

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

the relevant health research ethics committees of the participating institutions under approval numbers: 001354/Universitas Dian Nuswantoro/2024, B/119/III/DIK.2.6/2024/Rumkit, B-3.3/1080/RSR/IV/2024, 070/1491.a/RSUD/2024, and EA-064/KEPK-RSR/VII/2025. All records were de-identified before analysis to protect patient confidentiality.

Consent to Participate.
The requirement for individual informed consent was waived by the relevant ethics committees because this retrospective study used anonymized secondary clinical data and involved no direct contact with participants.

Consent for Publication.
Not applicable. This manuscript does not contain identifiable personal data, images, or clinical details that could reveal the identity of individual participants.

Competing Interests
The authors declare that they have no competing financial, professional, or personal interests that could have influenced the work reported in this manuscript.

References

[1] W. Zheng, B. Peng, Y. Wu, L. Gauan, S. Wang, and H. Ning, "Global, regional, and national anemia burden among women of reproductive age (15–49 years) from 1990 to 2021: an analysis of the Global Burden of Disease Study 2021," *Front. Nutr.*, vol. 12, 2025, doi: https://doi.org/10.3389/fnut.2025.1588496.

[2] Z. Sugiyanto *et al.*, "Clinical Data and Laboratory Tests on Covid-19 Patients in Six (6) Hospitals in 2021," *Int. J. Heal. Lit. Sci.*, vol. 2, no. 1, pp. 41–46, 2024, doi: https://doi.org/10.60074/ihelis.v2i1.57.

[3] S. Aiwale *et al.*, "Non-invasive Anemia Detection and Prediagnosis," *J. Pharmacol. Pharmacother.*, vol. 15, pp. 408–416, 2024, doi: https://doi.org/10.1177/0976500X241276307.

[4] M. Anai *et al.*, "Decrease in hemoglobin level predicts increased risk for severe respiratory failure in COVID-19 patients with pneumonia," *Respir. Investig.*, vol. 59, pp. 187–193, 2020, doi: https://doi.org/10.1016/j.resinv.2020.10.009.

[5] S. Rahmaniar, B. Kusuma, and Irfanuddin, "Analysis of the Role of Age Factors on Hemoglobin, Leukocyte and Platelet Levels in Breast Cancer Patients Post Chemotherapy 3 Cycles at Dr. Mohammad Hoesin General Hospital, Palembang, Indonesia," *Sriwij. J. Surg.*, vol. 7, no. 2, pp. 629–637, 2024, doi: https://doi.org/10.37275/sjs.v7i2.103.

[6] O. Karnabeda, T. Yehorova, and K. Khurdepa, "Anemia of Chronic Disease (Literature review and a clinical case)," *Fam. Med. Eur. Pract.*, no. 4, pp. 154–161, 2025, doi: https://doi.org/10.30841/2786-720X.4.2025.349484.

[7] Q. Zhao *et al.*, "Association between platelet/high-density lipoprotein cholesterol ratio and blood eosinophil counts in American adults with asthma: a population-based study," *Lipids Health Dis.*, vol. 24, 2025, doi: https://doi.org/10.1186/s12944-025-02479-9.

[8] O. Golomb, I. Schushan-Eisen, A. Maayan-metzger, N. Elisha, T. Strauss, and R. Mazkereth, "The Clinical Significance of Extreme Leukocytosis among Newborns: A Retrospective Cohort Study," *Am. J. Perinatol.*, vol. 41, pp. e470–e476, 2022, doi: https://doi.org/10.1055/a-1906-9048.

[9] M. A. Haft *et al.*, "Identification of Data-Driven Preoperative Hemoglobin Strata That Predict the Likelihood of Blood Transfusion and the Risk of Major Complications and Prosthetic Joint Infection After Total Hip Arthroplasty," *J. Am. Acad. Orthop. Surg.*, vol. 33, pp. 127–134, 2024, doi: https://doi.org/10.5435/jaaos-d-24-00435.

[10] P. Yang *et al.*, "Clinical significance of hemoglobin level and blood transfusion therapy in elderly sepsis patients: A retrospective analysis.," *Am. J. Emerg. Med.*, vol. 73, pp. 27–33, 2023, doi: https://doi.org/10.1016/j.ajem.2023.08.005.

[11] A. Pranolo *et al.*, "Enhanced Multivariate Time Series Analysis Using LSTM: A Comparative Study of Min-Max and Z-Score Normalization Techniques," *Ilk. J. Ilm.*, vol. 16, no. 2, pp. 210–219, 2024, doi: https://doi.org/10.33096/ilkom.v16i2.2333.210-220.

[12] A. A. A. R. Hassan *et al.*, "Perioperative Anemia, Gender, Blood Transfusion and Urine Output during Cardiopulmonary Bypass: Risk Factors for Acute K. Injury," *Pakistan J. Med. Heal. Sci.*, vol. 16, no. 4, pp. 270–272, 2022, doi: https://doi.org/10.53350/pjmhs22164270.

[13] M. Rezaeiahari *et al.*, "Pediatric asthma population risk stratification using k-means clustering," *J. Asthma*, vol. 62, pp. 2060–2069, 2025, doi: https://doi.org/10.1080/02770903.2025.2552745.

[14] D. N. A. Putri, M. Jannah, S. D. Puspita, and Y. Yuanta, "Mapping Adolescents Vulnerable to Early Marriage Using K-Means Clustering: Strategies for Nutrition Prevention," *Int. J. Stud. Soc. Sci. Humanit.*, vol. 2, no. 2, pp. 124–133,

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

2025, doi: https://doi.org/10.25047/ijossh.v2i2.6722.

[15] R. Buaton and S. Solikhun, "The Application of Numerical Measure Variations in K-Means Clustering for Grouping Data," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 23, no. 1, pp. 103–112, 2023, doi: https://doi.org/10.30812/matrik.v23i1.3269.

[16] A. D. Ratulangi, H. Djohan, and L. Kamilla, "The Relationship Between NS1 Examination and the Examination of Hemoglobin, Hematocrit, Leukocytes, Platelets, and Erythrocytes in Dengue Fever (DBD) Patients in The Pediatric Ward," *MEDICA (International Med. Sci. Journal)*, vol. 7, no. 1, pp. 32–39, 2025, doi: https://doi.org/10.53770/medica.v7i1.486.

[17] I. Blydt-Hansen *et al.*, "Evaluating Hemoglobin Thresholds on the Basis of Sex: Preliminary Insights from a Systematic Review," *Blood*, vol. 142, no. 1, pp. 7200–7201, 2023, doi: https://doi.org/10.1182/blood-2023-190173.

[18] K. Ito, Y. Mitobe, R. Inoue, and M. Momoeda, "The quality of life and work productivity are affected by the presence of nausea/vomiting in patients taking iron preparations for heavy menstrual bleeding or anemia: a population-based cross-sectional survey in Japan," *BMC Womens. Health*, vol. 24, 2024, doi: https://doi.org/10.1186/s12905-024-03104-0.

[19] R. Aschenbruck, G. Szepannek, and A. F. X. Wilhelm, "Initialization strategies for clustering mixed-type data with the k-prototypes algorithm," *Adv. Data Anal. Classif.*, 2025, doi: https://doi.org/10.1007/s11634-025-00639-4.

[20] R. Tao, H. Li, and X. Yu, "MSKP: A Proposed K-Prototypes Clustering Algorithm for Multi-Source Medical Data," *2025 44th Chinese Control Conf.*, pp. 3933–3938, 2025, doi: https://doi.org/10.23919/CCC64809.2025.11178972.

[21] J. M. G. Vilar and L. Saiz, "Reliably quantifying the evolving worldwide dynamic state of the COVID-19 outbreak from death records, clinical parametrization, and demographic data," *Sci. Rep.*, vol. 11, 2021, doi: https://doi.org/10.1038/s41598-021-99273-1.

[22] A. D. Chestnikhina, D. O. Efremenko, N. A. Kabina, M. O. Reviakina, I. A. Snimshchikova, and A. V Konshina, "Developing Legal Competences in Medical Students as Applied to Biomedical Research: Pedagogical Experience Based on the Standards of the Declaration of Helsinki," *Rudn J. Psychol. Pedagog.*, vol. 22, no. 3, pp. 552–573, 2026, doi: https://doi.org/10.22363/2313-1683-2025-22-3-552-573.

[23] D. Costal, C. Farr'e, X. Franch, and C. Quer, "Inclusion and Exclusion Criteria in Software Engineering Tertiary Studies: A Systematic Mapping and Emerging Framework," *Proc. 15th ACM / IEEE Int. Symp. Empir. Softw. Eng. Meas.*, 2021, doi: https://doi.org/10.1145/3475716.3484190.

[24] A. B. Wiratman and W. Wella, "Personalized Learning Models Using Decision Tree and Random Forest Algorithms in Telecommunication Company," *JOIV Int. J. Informatics Vis.*, vol. 8, no. 1, pp. 318–325, 2024, doi: https://dx.doi.org/10.62527/joiv.8.1.1905.

[25] M. L. Tauryawati and A. F. Zainuddin, "Mapping Disaster-Prone Areas on Java Island Using The K-PROTOTYPES Algorithm," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 20, no. 1, pp. 179–196, 2025, doi: https://doi.org/10.30598/barekengvol20iss1pp0179-0196.

[26] G. F. Shidik, O. Setiono, E. J. Kusuma, L. B. Handoko, P. N. Andono, and M. F. Abdollah, "Novel Unsupervised Cluster Reinforcement Q-Learning in Minimizing Energy Consumption of Federated Edge Cloud," *IEEE Access*, vol. 13, pp. 92577–92595, 2025, doi: 10.1109/ACCESS.2025.3572084.

[27] E. Mantadakis, "Iron Deficiency Anemia in Children Residing in High and Low-Income Countries: Risk Factors, Prevention, Diagnosis and Therapy," *Mediterr. J. Hematol. Infect. Dis.*, vol. 12, 2020, doi: https://doi.org/10.4084/mjhid.2020.041.

[28] A. S. Yaro, F. Maly, P. Prazak, and K. Malý, "Outlier Detection Performance of a Modified Z-Score Method in Time-Series RSS Observation With Hybrid Scale Estimators," *IEEE Access*, vol. 12, pp. 12785–12796, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3356731.

[29] B. Peng and J. Zhu, "Self-Calibration Method for Microfacet Slope Standard Deviation σ in Leaf BRDF Microfacet Model Fitting," *Opt. Contin.*, vol. 4, no. 12, pp. 2813–2824, 2025, doi: https://doi.org/10.1364/OPTCON.564304.

[30] B. D. Roth, M. G. Saunders, C. M. Bachmann, and J. van Aardt, "On Leaf BRDF Estimates and Their Fit to Microfacet Models," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 1761–1771, 2020, doi: https://doi.org/10.1109/JSTARS.2020.2988428.

[31] H. Hafid and S. Annisa, "Implementation of K-MEDOIDS And K-PROTOTYPES Clustering For Early Detection of Hypertension Disease," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 19, no. 1, pp. 465–476, 2025, doi: https://doi.org/10.30598/barekengvol19iss1pp046

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

5-0476.

[32] A. V Eremeev, A. V Kel'manov, M. Y. Kovalyov, and A. V Pyatkin, "Selecting a subset of diverse points based on the squared euclidean distance," *Ann. Math. Artif. Intell.*, vol. 90, pp. 965–977, 2021, doi: https://doi.org/10.1007/s10472-021-09773-z.

[33] M. Molina and D. P. Payá, "Data-Driven Inventory Policy Assignment in ETO Environments Using Fuzzy K-Prototypes Clustering," *Mathematics*, vol. 14, no. 12, pp. 1–23, 2026, doi: https://doi.org/10.3390/math14122206.

[34] N. I. Hussain, K. Boruah, and A. Akhtar, "Predicting Evolutionary Importance of Amino Acids through Mutation of Codons Using K- means Clustering," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 1, pp. 13–26, 2025, doi: https://doi.org/10.35882/jeeemi.v6i4.457.

[35] K. Siddavatam and S. Shinde, "of Imbalance Medical Data Using Nature Inspired Percolation Clustering," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 3, pp. 740–762, 2025, doi: https://doi.org/10.35882/jeeemi.v7i3.835.

[36] Z. Jiao, S. Chen, H. Shi, and J. Xu, "Multi-Modal Feature Selection with Feature Correlation and Feature Structure Fusion for MCI and AD Classification," *Brain Sci.*, vol. 12, 2022, doi: https://doi.org/10.3390/brainsci12010080.

[37] D. Chicco, A. Campagner, A. Spagnolo, D. Ciucci, and G. Jurman, "The Silhouette coefficient and the Davies-Bouldin index are more informative than Dunn index, Calinski-Harabasz index, Shannon entropy, and Gap statistic for unsupervised clustering internal evaluation of two convex clusters," *PeerJ Comput. Sci.*, vol. 11, p. e3309, 2025, doi: https://doi.org/10.7717/peerj-cs.3309.

[38] H. Benhammou, K. Anoune, and A. Tajmouati, "Determining LFP Battery Parameters through Sum of Squared Errors Optimization," *2025 11th Int. Conf. Optim. Appl.*, pp. 1–4, 2025, doi: https://doi.org/10.1109/ICOA66896.2025.11236837.

[39] D. Chicco, A. Sichenze, and G. Jurman, "A simple guide to the use of Student's t-test, Mann-Whitney U test, Chi-squared test, and Kruskal-Wallis test in biostatistics," *BioData Min.*, vol. 18, 2025, doi: https://doi.org/10.1186/s13040-025-00465-6.

[40] L. Stěpánek, F. Habarta, I. Malá, and L. Marek, "A short note on post-hoc testing using random forests algorithm: Principles, asymptotic time complexity analysis, and beyond," *2022 17th Conf. Comput. Sci. Intell. Syst.*, vol. 30, pp. 489–497, 2022, doi: http://dx.doi.org/10.15439/2022F265.

[41] H. Mukherjee and P. Bhonge, "Assessing Skew Normality in Marks Distribution, a Comparative Analysis of Shapiro Wilk Tests," in *arXivLabs: experimental projects with community collaborators*, 2025, pp. 1–8. doi: https://doi.org/10.48550/arXiv.2501.14845.

Author Biography

Oki Setiono, M.Kom, He is a lecturer in Medical Records and Health Information, specializing in electronic medical record systems, healthcare databases, and clinical data analytics. He completed his Bachelor's degree in Informatics Engineering in 2014 and his Master's degree in Informatics Engineering in 2017. With over 10 years of experience in teaching, research, and web-based system development, his work has been focusing on the integration and analysis of clinical data. In addition, he is engaged in research and community service related to the security and efficiency of health information systems.

dr. Zaenal Sugiyanto, M.Kes, is a medical professional with extensive experience in public health and clinical practice. He earned his degree as a General Practitioner from the Faculty of Medicine, Diponegoro University (UNDIP) in 1991. He later pursued postgraduate studies and obtained a Master of Public Health from the same university in 2006. His academic and professional interests focus on disease coding and the management of infectious diseases. He has a particular specialization in Dengue Hemorrhagic Fever, contributing to both clinical understanding and public health strategies. His work supports improved healthcare data accuracy and disease control efforts.

Dyah Ernawati, S.Kep, Ns, M.Kes, completed her Diploma in Nursing in 1999 and later pursued a Bachelor of Nursing degree (2023–2025), followed by the Ners professional program (2025–2026). She earned a Master of Health with a concentration in Hospital Administration from Diponegoro University, Semarang (2010–2012). Her expertise lies in disease classification to support health information management within hospital administration. She has been an active lecturer at Universitas Dian Nuswantoro since 2004, teaching various courses. She is also an experienced speaker on disease classification and medical procedures using ICD-10 and ICD-9-CM since 2010.

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Arif Kurniadi, M.Kom, earned his Bachelor's degree in Computer Science in Information Systems from Universitas Dian Nuswantoro, Semarang, in 1998. He later completed his Master's degree in Computer Science (M.Kom) in Informatics Engineering at the same university in 2008. He has extensive experience in the development and implementation of information systems, particularly in the healthcare sector. His current research interests include Health Information Systems, machine learning, and the management of health-related variables and metadata. His work focuses on improving data quality, interoperability, and decision-making processes in healthcare environments through innovative technological approaches.

Dr. Ika Pantiawati, S.Si.T, M.Kes obtained her degree in Health Sciences from Universitas Diponegoro, Semarang, Indonesia. She is currently a faculty member at the Faculty of Health Sciences, Universitas Dian Nuswantoro, and is actively involved in teaching, research, and community service. Her research interests include public health, particularly maternal and child health, as well as community-based health interventions. She focuses on improving health outcomes through preventive strategies and health education. Additionally, her work emphasizes the application of research findings into practical programs to support sustainable healthcare development and improve the quality of life in communities.

Dr. Mohamad Nazri Husin, is currently affiliated with the Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, Malaysia. He is actively involved in academic and research activities and is a member of the Special Interest Group on Modeling and Data Analytics. His research interests include data analytics, predictive modeling, and the application of computational approaches to solve real-world problems, with a focus on enhancing data-driven decision-making and innovation across various domains.

Manuscript received 30 April 2026; Revised 30 June 2026; Accepted 10 July 2026; Available online 24 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1723
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).