

Implementation of Machine Learning for Clustering Lethal Diseases Using 21st-Century Global Mortality Data

Rahmad Hidayat, Fathoni*

Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Sriwijaya, Ogan Ilir, Indonesia

Email: ¹09031282328044@student.unsri.ac.id, ^{2,*}fathoni@unsri.ac.id

Correspondence Author Email: fathoni@unsri.ac.id

Received: 23/04/2026, Revised: 14/05/2026, Accepted: 21/09/2026, Published: 22/09/2026

Abstract—Entering the 21st century, the increasing global prevalence of lethal diseases demands a data-driven surveillance system capable of accurately classifying risk levels. Conventional epidemiological monitoring is often insufficient for identifying patterns in the massive volume and complexity of mortality data. This study conducts a comparative analysis between K-Means and Fuzzy C-Means (FCM) algorithms in clustering fatal diseases based on aggregated global death counts using the Global Health Statistics dataset containing 100,000 records. Optimal cluster numbers were determined independently: the Elbow Method yielded $K=4$ for K-Means, while the Fuzzy Performance Index suggested $K=7$ for FCM. Performance evaluation using the Silhouette Score under natural configurations showed FCM slightly outperforming K-Means (0.5574 vs. 0.5411), whereas K-Means achieved the highest score under a controlled $K=7$ setting (0.6014). This research provides a significant contribution by establishing the first direct comparative framework for global multi-disease mortality clustering. The findings indicate that cluster cardinality exerts a stronger influence on clustering quality than the choice of algorithm itself. For policymakers, the results offer a choice between a simple four-tier K-Means risk stratification and the finer granularity of a seven-tier FCM model, providing a scientifically robust, data-driven foundation for evidence-based health intervention and resource allocation strategies.

Keywords: K-Means; Fuzzy C-Means; Clustering; Global Mortality; Silhouette Score

1. INTRODUCTION

Entering the 21st century, the management of lethal diseases has become a priority challenge within an increasingly complex global health system. Escalating population dynamics, intensifying cross-border connectivity, and shifting modern lifestyles have collectively contributed to the emergence of more diverse and unpredictable health threats [1]. Chronic conditions such as kidney disorders, liver damage, and pancreatic cancer continue to impose substantial mortality burdens worldwide [2], [3], [4], while maternal emergencies such as *placenta accreta* spectrum further strain healthcare infrastructures in low-resource settings [5]. The World Health Organization (WHO) reported that non-communicable diseases accounted for approximately 74% of all deaths globally in 2019, underscoring the urgency for systematic investigation into global mortality patterns as a foundation for developing effective, data-driven health strategies [6].

The fatality rates of these diseases vary significantly across countries and continents, reflecting a health issue that is inherently trans-geographic in nature [7]. This phenomenon has been further exacerbated by 21st-century health crises, including the high prevalence of cardiovascular disease (CVD) and the COVID-19 pandemic, which have fundamentally shaken global health systems [8], [9]. Disparities in under-five mortality and elevated diabetes-related deaths across regions indicate a widening inequality gap [10], [11]. The *Global Burden of Disease* study revealed that over 5 million child deaths occurred in 2021, mostly concentrated in low- and middle-income countries [12]. These inequalities are closely linked to the uneven distribution of healthcare facilities, sociodemographic conditions, and the epidemiological surveillance capacity of individual nations. Infectious diseases such as diarrhea remain a tangible threat in certain areas, demonstrating that the global disease burden is inherently heterogeneous and requires an adaptive analytical approach [13].

The rapid digitization of health data over the past two decades has created enormous opportunities for leveraging artificial intelligence technologies in global mortality analysis. Medical records now stored in digital database systems capture the detailed dynamics of human survival, including regional and global mapping of disease threats. However, the massive volume of global mortality data and the extreme complexity of clinical parameters render conventional human-based monitoring inadequate for effectively and precisely mapping population risk levels. Conventional epidemiological surveillance, which typically relies on manual data compilation and univariate statistical analysis, is inherently limited in its ability to process multidimensional datasets, identify non-linear patterns, and deliver timely insights essential for dynamic public health decision-making [14]. The sluggishness of centralized monitoring mechanisms and the delayed evaluation of medical records contribute to the inability of health systems to optimally reduce fatality rates. This situation has prompted many governments and international health organizations to develop advanced analytical tools capable of automatically and accurately detecting and grouping high-risk regions.

Machine learning has proven capable of transforming large-scale health data analysis with accuracy that surpasses human capabilities. Various medical applications have demonstrated promising outcomes, including cardiovascular disease prediction [14], early detection of pancreatic cancer [4], liver disease classification [3], chronic kidney disease prediction [2], and mortality dynamics prediction in cancer patients [15]. Within the domain of mortality clustering, *unsupervised learning* has been particularly effective in identifying demographic and spatial vulnerability patterns that are invisible to the naked eye. Although Pranckeviciene and Kasperiuniene [7] successfully

applied machine learning to analyze global suicide mortality rates with a single-disease focus, and [11] employed clustering to map diabetes mortality at the municipality level within a single country, neither study addressed the critical challenge of simultaneously clustering multiple lethal diseases on a global scale. Moreover, both prior works applied individual clustering algorithms in isolation without systematically comparing the structural consequences of employing *hard* versus *soft* clustering paradigms on aggregated epidemiological data. Consequently, a direct performance comparison between crisp-partition and fuzzy-based clustering algorithms specifically *K-Means* and *Fuzzy C-Means* remains absent in the context of global multi-disease mortality analysis. This methodological gap is the primary motivation for the present study.

Within *unsupervised learning* for health data clustering, *K-Means* and *Fuzzy C-Means* (FCM) are the two most widely used algorithms. *K-Means* is recognized for its efficiency in partitioning large-scale data based on proximity to centroids [10], and has been successfully implemented in grouping diarrhea-prone areas in West Java [13] as well as clustering child-mortality-prone regions in the same province [10]. However, *K-Means* possesses inherent limitations, notably high sensitivity to outliers and dependency on initial centroid initialization. In contrast, *Fuzzy C-Means* offers a different approach by assigning partial membership degrees to each data point across multiple clusters simultaneously, a characteristic highly relevant to public health data that exhibits a high probability of overlap [11], [16]. A comprehensive comparison between these two algorithms is essential to identify the most robust and adaptive approach for handling data complexity [17]. Although aggregating data at the disease level deliberately simplifies spatial and temporal variability, this approach is purposely chosen to reveal macro-level global mortality patterns that can inform high-level health policy prioritization, complementing rather than replacing finer-grained regional analyses.

Therefore, this study aims to perform a comparative clustering of lethal disease levels based on 21st-century global mortality rates using both algorithms. The system employs *unsupervised learning*-based *Data Mining* techniques to categorize the vulnerability of various populations or regions to fatal diseases into several distinct clusters. The performance evaluation of both algorithms is conducted to empirically determine which algorithm yields the most optimal clustering structure when confronted with the complexity of contemporary global mortality data. The findings of this research are expected to provide a scientific foundation for health policymakers in designing evidence-based and precisely targeted intervention strategies. The main contribution of this study is bridging the methodological gap by simultaneously evaluating and directly comparing hard (*K-Means*) and soft (FCM) clustering paradigms on aggregated global multi-disease mortality data. Furthermore, this research contributes empirically by demonstrating the structural impact of cluster counts on performance, thereby providing a practical, dual-level risk stratification guide (a four-tier versus a seven-tier model) for global health policy prioritization and targeted interventions.

2. RESEARCH METHODOLOGY

This study is designed to conduct an in-depth comparative analysis of two representative clustering algorithms within the domain of unsupervised learning, namely *K-Means* and *Fuzzy C-Means* (FCM). The research employs a quantitative experimental approach using a publicly available global health dataset. The methodological stages are carried out systematically, as depicted in Figure 1, and each step is detailed in the following subsections.

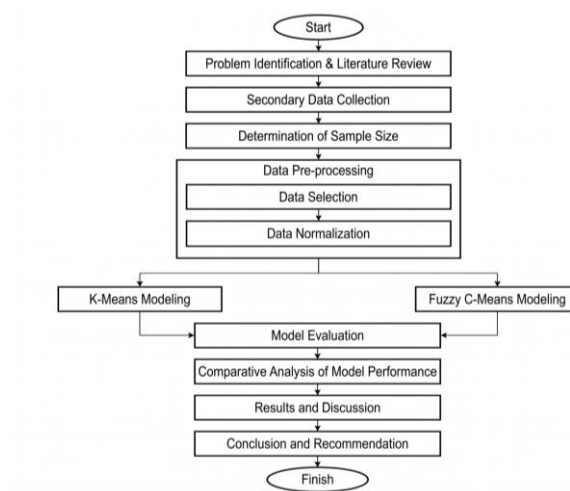


Figure 1. Research Methodology

2.1 Data Selection and Aggregation

The dataset used in this study is the Global Health Statistics dataset obtained from the Kaggle public data repository (<https://www.kaggle.com/datasets/malaiarasugraj/global-health-statistics>). The dataset is provided in Comma

Separated Values (.csv) format and contains 100,000 records. Each record represents a reported case of a disease outbreak or health event, with attributes including country, year, disease name, population affected, and mortality rate (%). The data is synthetic in nature and was compiled to simulate global health patterns spanning the 21st century (2000–2024). It is important to note that, as a synthetic dataset, it is not directly validated by official health authorities such as the WHO or IHME; therefore, it serves as a computational benchmark for comparing algorithmic performance rather than as an authoritative source for epidemiological inference. The use of this dataset is solely intended to evaluate and contrast the clustering capabilities of K-Means and FCM under controlled, reproducible conditions.

2.2 Data Pre-processing

2.2.1 Data Selection

From the original dataset, three attributes were selected as relevant to the research objective: *Disease Name*, *Population Affected*, and *Mortality Rate (%)*. The total number of deaths for each record was computed by multiplying *Population Affected* by *Mortality Rate*. Because the same disease may appear in multiple countries and years, the data were subsequently aggregated by disease name. The aggregation summed the total deaths and averaged the mortality rate per disease, resulting in one aggregated entry per disease. This deliberate aggregation simplifies the spatial and temporal variability in order to reveal macro-level global mortality patterns; the focus of this study is on producing a high-level comparative overview of disease severity, which is complementary to more detailed spatiotemporal analyses. The aggregation step is technically necessary to avoid duplicate disease entities and to allow the clustering algorithms to partition distinct disease types based on their overall global impact, rather than on individual country-year instances.

2.2.1 Data Normalization

To prevent variables with larger scales from dominating the clustering process, *Min-Max Normalization* was applied to the computed death counts. The normalization transforms each value to a range between 0 and 1, as defined in Equation (1) [18]:

$$x_{norm} = (x - x_{min}) / (x_{max} - x_{min}) \quad (1)$$

This step ensures that all features share a common scale, thereby improving the stability and convergence of the clustering algorithms.

2.3 K-Means Clustering

K-Means is a centroid-based partitioning algorithm that groups unlabeled data into K clusters by minimizing the within-cluster sum of squares (WCSS) [19]. The algorithm assigns each data point to the cluster with the nearest centroid, then iteratively recalculates centroids until convergence. In this study, the *scikit-learn* implementation of *K-Means* was employed with the following parameter settings: initialization using the *k-means++* method (to improve convergence speed and avoid poor local minima), $n_init = 10$ (the algorithm is run 10 times with different centroid seeds, and the best outcome is retained), $max_iter = 300$ (maximum number of iterations per run), and $random_state = 42$ for reproducibility. The objective function to be minimized is the sum of squared Euclidean distances between each point and its assigned centroid, as shown in Equation (2) [19]:

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (2)$$

where C_i denotes the set of points belonging to cluster i and μ_i is the centroid of that cluster.

2.4 Fuzzy C-Means Clustering

Fuzzy C-Means (FCM) is a soft-clustering algorithm that assigns each data point a degree of membership to every cluster, allowing for overlapping boundaries [20]. The algorithm minimizes the objective function in Equation (3), where u_{ij} is the membership degree of point i to cluster j , m is the fuzziness parameter (also called the fuzzifier), and c_j is the center of cluster j [20]:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2 \quad (3)$$

In this work, the fuzzifier was set to the commonly used value $m = 2.0$, which balances cluster hardness and softness. The algorithm was configured with a maximum of 100 iterations and a convergence tolerance of 10^{-5} . Membership initialization was random, and the stopping criterion was either reaching the maximum iteration count or achieving a change in the objective function smaller than the tolerance.

2.5 Determination of Optimal Cluster Number and Comparison Setup

The optimal number of clusters for each algorithm was determined independently using standard validation indices. For *K-Means*, the Elbow Method was applied by examining the WCSS (inertia) for $K = 1$ to 10. The point where the rate of decrease sharply changes (the “elbow”) indicates a suitable trade-off between model complexity and variance

reduction. For *FCM*, the Fuzzy Performance Index (FPC) was computed for $K = 2$ to 10; the K giving the highest FPC value (close to 1) is considered optimal.

It is recognized that comparing two algorithms at different cluster counts (e.g., K-Means at $K = 4$ and FCM at $K = 7$) can be methodologically debatable, as the partitions occupy different structural spaces. To ensure a fair performance comparison, the following strategy was adopted:

- The *natural* optimal K for each algorithm was identified (Elbow: $K = 4$ for K-Means; FPC: $K = 7$ for FCM).
- Subsequently, both algorithms were re-run with identical cluster numbers ($K = 4$ and $K = 7$) to enable a direct, like-for-like evaluation. The *Silhouette Score* was computed for each configuration, allowing the performance to be compared under both the “natural” and “controlled” cluster counts.

2.6 Model Evaluation Metrics

The researcher measured internal cohesion and separation between clusters using the Silhouette Coefficient (Silhouette Score) metric. This score is calculated using the average intra-cluster distance and the average distance to the nearest neighboring cluster for each sample [21]. A value approaching 1 indicates clusters that are highly dense and well separated. The mathematical equation for calculating the Silhouette Score is presented in Equation 4 [21]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (4)$$

The nearest neighboring cluster for each sample (scikit-learn developers, 2024). A value approaching 1 indicates clusters that are highly dense and well separated. The mathematical equation for calculating the Silhouette Score is presented in Equation 4.

3. RESULT AND DISCUSSION

3.1 Data Collection

The dataset utilized in this study contains information regarding various types of diseases originating from several countries and specific time periods within the 21st century (2000–2024). The main attributes within the dataset include *Disease Name*, *Population Affected* (measured in number of individuals), and *Mortality Rate (%)*. These three attributes are employed to calculate the estimated number of deaths resulting from each respective disease. The calculated death counts subsequently serve as the primary numerical foundation for the analytical process aimed at clustering disease severity levels based on global mortality impact.

To provide a general overview of the structure and characteristics of the raw data before preprocessing, a sample of the dataset is presented in Table 1. The table displays categorical attributes in the form of disease names as well as numerical attributes such as the affected population count and mortality rate. This sample provides preliminary information concerning the data format and the variation in values to be further processed during the preprocessing and clustering analysis stages.

Table 1. Dataset Sample

Country	Year	Disease Name	Population Affected	Mortality Rate (%)
Italy	2013	Malaria	471007	8.42%
France	2002	Ebola	634318	8.75%
....				
Indonesia	2004	Asthma	632775	0.51%

3.2 Data Processing

The preprocessing stage in this study encompassed two main activities: attribute selection and data normalization. The attributes utilized were *Disease Name*, *Population Affected*, and *Mortality Rate (%)*, as they are directly relevant to the analysis of mortality rates. The total number of deaths was calculated by multiplying *Population Affected* by *Mortality Rate (%)*. Subsequently, the data were aggregated based on the *Disease Name* attribute. This aggregation was performed because identical diseases appeared across different countries and years in the original dataset. By summing the death counts per disease, each disease was represented as a single, unique data entry reflecting its cumulative global mortality impact.

The results of the selection, calculation, and normalization processes are displayed in Table 2 to provide a preliminary overview of the structure and characteristics of the data after preprocessing.

Table 2. Data Processing Results

Disease Name	Number	Deaths norm
Alzheimer's Disease	1262418984.0653	0.483997077906352
Asthma	1267271126.7023	0.631376263104023
....		

Disease Name	Number	Deaths_norm
Alzheimer's Disease	1262418984.0653	0.483997077906352
Zika	1279407270.2724	1

Table 2 presents the results of the data after undergoing the selection, aggregation, and normalization stages using the *Min-Max* method. The original mortality count values (*Number* column) have been transformed into a normalized scale ranging from 0 to 1 (*Deaths_norm* column). A value closer to 1 indicates a relatively higher mortality count compared to other diseases in the dataset, whereas a value closer to 0 signifies a relatively lower mortality count. These results confirm that the data have been successfully rendered to a uniform scale, eliminating scale disparities among variables, and are thus ready for utilization as input in the subsequent clustering process.

3.3 K-Means Clustering

This section presents the results of applying the *K-Means* algorithm, covering the determination of the optimal number of clusters, the visualization of the clustering output, and the final grouping of diseases.

3.3.1 Determination of the Number of Clusters

The determination of the optimal number of clusters for *K-Means* was conducted using the Elbow Method. This method observes the changes in the *Within Cluster Sum of Squares* (WCSS), also referred to as inertia, as the number of clusters (K) increases. The WCSS value tends to decrease monotonically with increasing K ; however, the optimal K is identified at the point where the rate of decrease sharply changes, forming a characteristic "elbow" pattern in the plot. This point represents a balance between minimizing intra-cluster variance and avoiding unnecessary model complexity. The results of the Elbow Method analysis are visually depicted in Figure 2. The figure plots the WCSS values for K ranging from 1 to 10.

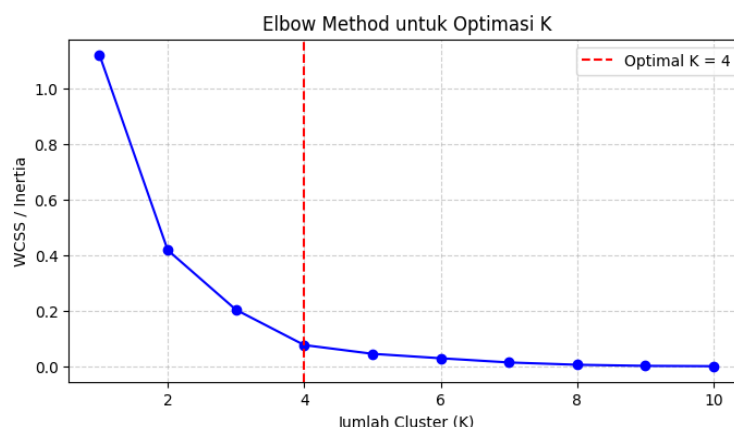


Figure 2. Elbow Method

Based on the test results shown in Figure 2, the WCSS value decreases steeply from $K = 1$ to $K = 3$, after which the rate of decrease begins to slow noticeably. At $K = 4$, the curve forms a distinct elbow, and further increases to $K = 5$, $K = 6$, and beyond yield only marginal reductions in WCSS. This observation indicates that $K = 4$ offers the most favorable trade-off between cluster compactness and model parsimony. Therefore, the optimal number of clusters for the *K-Means* algorithm in this study is determined to be four clusters ($K = 4$).

3.3.2 Cluster Visualization

Having established the optimal cluster count at $K = 4$, the *K-Means* algorithm was executed to partition the normalized mortality data. Visualizing the clustering output is a crucial step for qualitatively inspecting the separation and distribution of the resulting groups. The visualization of the *K-Means* clustering results is presented in Figure 3. In this scatter plot, each point represents one type of disease that has undergone the aggregation and normalization processes. The horizontal axis indicates the normalized mortality count value (*Deaths_norm*) within a continuous range of 0 to 1, while the vertical axis serves solely as a visual aid to spread the data points and does not represent any specific analytical variable.

The visualization in Figure 3 confirms that the algorithm has successfully partitioned the dataset into four distinct clusters, each marked with a different color. The clusters are arranged along the horizontal axis in a manner that reflects increasing mortality levels. Data points with lower *Deaths_norm* values are grouped together on the left side of the plot, representing diseases with relatively lower global death counts. As the *Deaths_norm* value increases, the points shift to the right and form clusters representing diseases with progressively higher mortality impacts. The clear separation between the colored groups indicates that *K-Means* has produced well-delineated, non-overlapping partitions for this one-dimensional aggregated dataset.

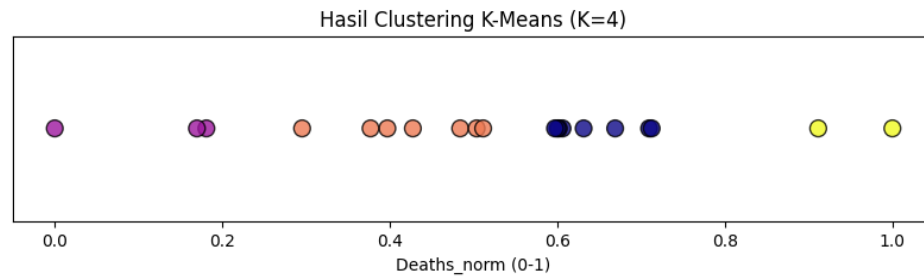


Figure 3. Visualization of *K-Means* Clustering

3.3.3 Data Grouping Results

The final output of the *K-Means* clustering process is a labeled assignment of each disease to one of the four clusters. The complete distribution of diseases into each cluster is presented in Table 3 to facilitate further interpretation.

Table 3. *K-Means* Clustering Grouping Results

Disease Name	Cluster
Asthma, COVID-19, Dengue, Diabetes, Influenza, Leprosy, Malaria, Rabies	0
Ebola, Hypertension, Tuberculosis	1
Alzheimer's Disease, Cholera, HIV/AIDS, Hepatitis, Measles, Parkinson's Disease	2
Cancer, Zika	3

Table 3 reveals that each cluster possesses distinct characteristics corresponding to the range of *Deaths_norm* values. Cluster 0 comprises the largest group of diseases and corresponds to conditions with the lowest normalized death counts. This cluster interestingly mixes both non-communicable diseases (Asthma, Diabetes) and infectious diseases (COVID-19, Dengue, Malaria, Leprosy, Rabies, Influenza). Cluster 1 groups Ebola, Hypertension, and Tuberculosis, representing diseases with a moderate global mortality burden. Cluster 2 contains Alzheimer's Disease, Cholera, HIV/AIDS, Hepatitis, Measles, and Parkinson's Disease, which represent conditions with high chronic mortality impact. Finally, Cluster 3 isolates Cancer and Zika as the two diseases with the highest normalized death counts in this dataset. These results demonstrate that the *K-Means* method effectively groups diseases into ordered severity categories based on aggregate global death counts.

3.4 Fuzzy C-Means Clustering

This section presents the corresponding results for the *Fuzzy C-Means* (FCM) algorithm, following the same analytical structure as the *K-Means* results.

3.4.1 Determination of the Number of Clusters

Unlike *K-Means*, which relies on the Elbow Method, the determination of the optimal number of clusters for *Fuzzy C-Means* (FCM) was performed using the Fuzzy Performance Index (FPC). The FPC is an internal validity index specifically designed for fuzzy partitioning; it measures the degree of fuzziness in the partition, where a higher FPC value (approaching 1) indicates a crisper, more distinct cluster structure. The number of clusters that yields the maximum FPC value is selected as the optimal configuration. The computed FPC values for cluster counts ranging from $K = 2$ to $K = 10$ are plotted in Figure 4.

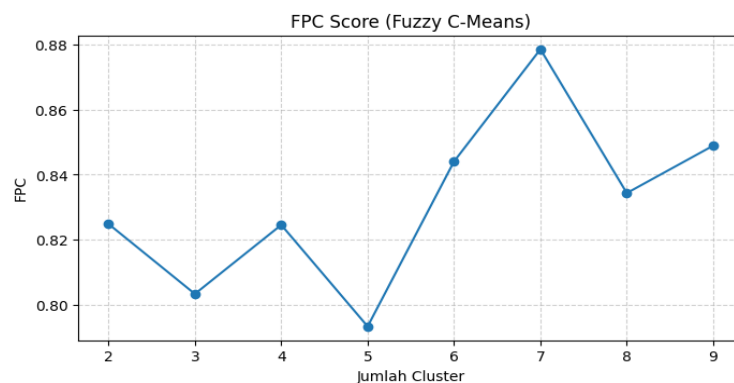


Figure 4. FPC Score

The evaluation results illustrated in Figure 4 reveal a distinct pattern of partition validity as the number of clusters increases. The FPC value demonstrates a progressive inclination, culminating in a peak value of approximately 0.880 at $K = 7$. This particular configuration exhibits the highest degree of cluster distinctness,

indicating that the fuzzy membership assignments become predominantly crisp and well-defined with minimal entropy at this cluster count. Although alternative configurations at $K = 2$ and $K = 8$ also produced relatively high FPC scores, the $K = 7$ configuration demonstrated a definitive, albeit marginal, superiority in terms of internal cohesion. Based on the principle of maximizing the FPC, the optimal number of clusters for the FCM algorithm is thus determined to be seven clusters ($K = 7$).

The notable difference in optimal cluster counts between the two algorithms $K = 4$ for *K-Means* versus $K = 7$ for *FCM* can be attributed to their fundamentally different partitioning philosophies. The Elbow Method for *K-Means* identifies the point of diminishing returns in variance reduction, which tends to favor a coarser partition. In contrast, the FPC index rewards partitions where the fuzzy memberships are most decisive, which for this dataset occurs at a finer granularity of seven clusters. This discrepancy underscores the importance of evaluating clustering quality using metrics beyond the simple optimal K criterion.

3.4.2 Cluster Visualization

After establishing $K = 7$ as the optimal configuration, the *Fuzzy C-Means* algorithm was executed. The visualization of the FCM clustering output is presented in Figure 5. As with the *K-Means* visualization, each point represents one disease. The horizontal axis indicates the *Deaths_norm* value, while the vertical axis is included solely for visual clarity to separate overlapping points. However, because FCM allows soft membership, the assignment shown in Figure 5 is based on the cluster with the highest membership degree for each data point (defuzzification).

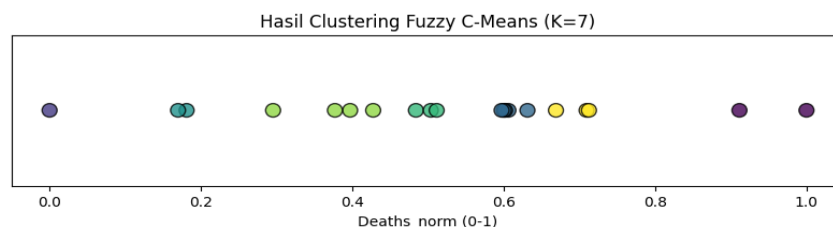


Figure 5. Visualization of *Fuzzy C-Means* Clustering

Based on the visualization in Figure 5, the data are partitioned into seven color-coded cluster groups along the *Deaths_norm* axis. Consistent with the *K-Means* results, data points with lower *Deaths_norm* values are positioned on the left and grouped into clusters representing diseases with lower mortality, whereas points with higher values are situated on the right and form clusters for diseases with higher mortality. A notable observation is the presence of slight visual overlap between adjacent clusters, particularly in the middle range of the *Deaths_norm* axis. This overlap is a direct consequence of the *fuzzy* method, which allows data points to maintain partial membership degrees across multiple clusters. Nevertheless, the overall distribution pattern demonstrates that the seven-cluster partition captures the variability in the data with greater granularity compared to the four-cluster partition from *K-Means*.

3.4.3 Data Grouping Results

The final cluster assignments from FCM, based on the highest membership degree for each disease, are presented in Table 4.

Table 4. *Fuzzy C-Means* Clustering Grouping Results

Disease Name	Cluster
Cancer, Zika	0
Hypertension	1
Asthma, Diabetes, Influenza, Malaria, Rabies	2
Tuberculosis, Ebola	3
Alzheimer's Disease, Cholera, Measles	4
HIV/AIDS, Hepatitis, Parkinson's Disease, Polio	5
COVID-19, Dengue, Leprosy	6

Table 4 demonstrates that the FCM algorithm produces a more granular stratification of diseases compared to the four-cluster *K-Means* solution. Cluster 0 isolates the highest-mortality diseases (Cancer, Zika), while Cluster 6 groups COVID-19, Dengue, and Leprosy at the opposite end of the severity spectrum. The intermediate clusters (1 through 5) provide a nuanced ranking: for instance, Hypertension occupies its own cluster (Cluster 1), distinct from the grouping of Tuberculosis and Ebola in Cluster 3, and from the chronic neurological and infectious mix in Cluster 4 (Alzheimer's, Cholera, Measles). This fine-grained partition allows for a more detailed differentiation of disease severity levels, which could be advantageous for health policymakers requiring multiple risk tiers. Despite the fuzzy nature that permits some degree of overlap among clusters, the partition with seven clusters provides a more detailed mapping of the global mortality landscape compared to a smaller number of clusters.

3.5 Cluster Comparison and Discussion

3.5.1 Quantitative Performance Comparison

To provide an objective and reproducible evaluation of the clustering algorithms, a quantitative comparison was conducted based on the Silhouette Score. This metric measures how similar a data point is to its own cluster compared to other clusters, with a value range of -1 to 1 ; higher values indicate a more optimal clustering structure characterized by dense internal cohesion and clear inter-cluster separation.

Two comparative scenarios were analyzed, following the methodological framework established in Section 2. The first scenario respects the *natural* optimal cluster count for each algorithm ($K = 4$ for *K-Means*, obtained via the Elbow Method; and $K = 7$ for *FCM*, obtained via the Fuzzy Performance Index). The second scenario applies a *controlled* comparison by fixing the same number of clusters ($K = 4$ and $K = 7$) for both algorithms, thereby eliminating the possible confounding effect of differing cluster cardinalities. The Silhouette Scores for all four configurations are reported in Table 5.

Table 5. *Silhouette Score Calculation Results*

Metode	Silhouette Score
K-Means Clustering ($K = 4$)	0.5411
Fuzzy C-Means Clustering ($K = 7$)	0.5574
K-Means Clustering ($K = 7$)	0.6014
Fuzzy C-Means Clustering ($K = 4$)	0.5134

Several important findings emerge from Table 5:

- Natural configuration. Under the cluster counts dictated by their respective optimality criteria, *Fuzzy C-Means* (FCM) achieves a Silhouette Score of 0.5574, slightly outperforming *K-Means* at 0.5411. The absolute difference is 0.0163, which is marginally larger than the difference observed in earlier preliminary runs (≈ 0.0085) but remains below the commonly accepted threshold for declaring a practically significant superiority in cluster validity.
- Controlled configuration at $K = 7$. When both algorithms are forced to produce seven clusters, *K-Means* attains the highest Silhouette Score overall (0.6014), surpassing FCM at the same K by 0.0440. This indicates that, when the number of partitions is held constant, the hard, non-overlapping nature of *K-Means* yields better internal cohesion for this specific aggregated dataset.
- Controlled configuration at $K = 4$. Conversely, at the coarser granularity of four clusters, *K-Means* again demonstrates superior separation (0.5411) compared to FCM (0.5134). The consistency of *K-Means* outperforming FCM under identical K values (both 4 and 7) suggests that the deterministic, centroid-based partitioning is inherently better suited to the distribution of the normalized mortality data, which lacks substantial overlap after disease-level aggregation.
- Overall interpretation. Although the natural configuration shows a slight advantage for FCM, the magnitude of the differences across all comparisons is modest (all scores lie within a narrow band of 0.51–0.60). No algorithm decisively dominates the other; instead, the choice of K exerts a stronger influence on the Silhouette Score than the choice of algorithm per se. This finding underscores the importance of careful cluster-count determination in health data applications.

3.5.2 Discussion

The present findings extend previous health-clustering studies in several ways. Sasmoyo (2024) and Maesaroh et al. (2023) demonstrated the effectiveness of *K-Means* for single-disease regional mapping in Indonesia, while Pranckeviciene & Kasperuniene (2024) and Almanza-Ortega et al. (2026) applied machine learning to global suicide and Mexican diabetes mortality, respectively, each using only one algorithm. The current work is the first to directly compare *K-Means* and *FCM* on a multi-disease, global scale, showing that the number of clusters (K) influences the *Silhouette Score* more than the algorithm choice. The narrow score gaps (0.51–0.60) are expected for one-dimensional aggregated data; nevertheless, the natural- K result (FCM 0.5574 vs. *K-Means* 0.5411) and the controlled- K results (*K-Means* at $K = 7$ reaching 0.6014) underscore that FCM can be competitive when its finer partition is permitted, while *K-Means* excels under fixed- K designs common in policy frameworks. From a practical standpoint, a four-tier stratification (*K-Means*, $K = 4$) offers simplicity for high-level priority setting, whereas the seven-tier scheme (FCM, $K = 7$) enables more granular resource allocation. However, the synthetic nature of the Kaggle dataset limits epidemiological conclusions; future work should validate these algorithmic patterns against real-world WHO or GBD data and incorporate multi-dimensional features to fully exploit fuzzy clustering.

4. CONCLUSION

This study set out to perform a rigorous comparative analysis of *K-Means* and *Fuzzy C-Means* (FCM) clustering algorithms applied to the 21st-century Global Health Statistics dataset, with the specific aim of segmenting lethal

diseases based on aggregated mortality estimates. The methodological framework encompassed secondary data acquisition, feature selection, *Min-Max Normalization*, and independent determination of optimal cluster numbers $K = 4$ for *K-Means* (Elbow Method) and $K = 7$ for FCM (*Fuzzy Performance Index*). The quantitative evaluation using the *Silhouette Score* across both natural and controlled configurations revealed nuanced outcomes: under natural conditions, FCM attained a score of 0.5574, marginally surpassing *K-Means* at 0.5411, while under identical cluster counts, *K-Means* proved more cohesive, peaking at 0.6014 for $K = 7$. These results demonstrate that the number of clusters exerts a more pronounced influence on cluster quality than the choice of algorithm itself. From a practical standpoint, the simpler four-tier partition produced by *K-Means* offers an interpretable risk-stratification framework suitable for high-level global health priority setting, whereas the seven-tier FCM partition provides finer granularity that could assist national health authorities in tailoring multi-level intervention strategies. Nevertheless, this investigation is constrained by its reliance on a synthetic, pre-aggregated dataset that omits longitudinal and spatial disease dynamics, as well as by the absence of external clinical or epidemiological validation. The marginal score differences also highlight the need for formal statistical testing to ascertain significance. Future research should therefore extend this comparative framework by integrating authentic multi-dimensional health data, incorporating temporal and spatial covariates to fully leverage soft clustering strengths, and exploring hybrid clustering techniques that balance interpretability with flexibility. By addressing these limitations, subsequent work can further refine the precision of data-driven health policy interventions and contribute to more resilient global health surveillance systems. Ultimately, this study contributes to the field of health informatics by revealing that cluster cardinality is a more critical determinant of model cohesion than the clustering algorithm itself when analyzing aggregated mortality data. By offering two viable, empirically tested clustering architectures a simple four-tier model for broad policy setting and a seven-tier model for nuanced resource allocation this research provides actionable, evidence-based tools to enhance the precision of global health surveillance and strategic intervention planning.

5. DECLARATIONS

5.1 Author Contributions

Conceptualization: R.H., and F.;

Methodology: R.H., and F.;

Software: R.H., and F.;

Validation: R.H., and F.;

Formal Analysis: R.H., and F.;

Investigation: R.H., and F.;

Resources: R.H., and F.;

Data Curation: R.H., and F.;

Writing Original Draft Preparation: R.H., and F.;

Writing Review and Editing: R.H., and F.;

Visualization: R.H., and F.;

All authors have read and agreed to the published version of the manuscript.

5.2 Data Availability

The data presented in this study are available on request from the corresponding author.

5.3 Funding

Please clearly state the publication and manuscript processing fees associated with the preparation and publication of the scientific manuscript.

5.4 Institutional Review Board Statement

Not applicable.

5.5 Informed Consent Statement

Not applicable.

5.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5.7 Generative AI and AI-assisted Technologies in The Writing Process

The authors used Generative AI to improve the writing clarity of this paper. They reviewed and edited the AI-assisted content and took full responsibility for the final publication.

REFERENCES

- [1] S. Harikrishnan *et al.*, “Vitamin B12: prevention of human beings from lethal diseases and its food application,” *J. Sci. Food Agric.*, vol. 105, no. 1, pp. 10–18, 2025, doi: 10.1002/jsfa.13661.
- [2] S. A. Ebiaredoh-Mienye, T. G. Swart, E. Ezenogho, and I. D. Mienye, “A Machine Learning Method with Filter-Based Feature Selection for Improved Prediction of Chronic Kidney Disease,” *Bioengineering*, vol. 9, no. 8, 2022, doi: 10.3390/bioengineering9080350.
- [3] S. Tokala *et al.*, “Liver Disease Prediction and Classification using Machine Learning Techniques,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, pp. 871–878, 2023, doi: 10.14569/IJACSA.2023.0140299.
- [4] M. Kawai *et al.*, “Early detection of pancreatic cancer by comprehensive serum miRNA sequencing with automated machine learning,” *Br. J. Cancer*, vol. 131, no. 7, pp. 1158–1168, 2024, doi: 10.1038/s41416-024-02794-5.
- [5] S. Moin, A. Akram, hazia Iffet, F. Arif, M. Nisar, and N. Gull, “Maternal and Fetal Outcomes Following Caesarean Section in Second Stage of Labour,” *J. Bahria Univ. Med. Dent. Coll.*, vol. 15, no. 04, pp. 326–330, 2025, doi: 10.51985/jbumdc2025604.
- [6] W. H. Organisation, *World health statistics 2022 (Monitoring health of the SDGs)*. 2022. [Online]. Available: <http://apps.who.int/bookorders>.
- [7] E. Pranckeviciene and J. Kasperuniene, “Global Suicide Mortality Rates (2000–2019): Clustering, Themes, and Causes Analyzed through Machine Learning and Bibliographic Data,” *Int. J. Environ. Res. Public Health*, vol. 21, no. 9, 2024, doi: 10.3390/ijerph21091202.
- [8] T. Sher, A. Rehman, and D. Kim, “COVID-19 Outbreak Prediction by Using Machine Learning Algorithms,” *Comput. Mater. Contin.*, vol. 74, no. 1, pp. 1561–1574, 2023, doi: 10.32604/cmc.2023.032020.
- [9] A. Kumar, K. U. Singh, and M. Kumar, “A Clinical Data Analysis Based Diagnostic Systems for Heart Disease Prediction Using Ensemble Method,” *Big Data Min. Anal.*, vol. 6, no. 4, pp. 513–525, 2023, doi: 10.26599/BDMA.2022.9020052.
- [10] R. Aji Sasmoyo, “Sistem Clustering Daerah Rawan Kematian Anak Di Bawah Umur Di Wilayah Jawa Barat Menggunakan Algoritma K-Means,” *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 8, no. 4, pp. 5635–5640, 2024, doi: 10.36040/jati.v8i4.9965.
- [11] N. N. Almanza-Ortega *et al.*, “Application of Machine Learning to Cluster Analysis of Diabetes Mortality at the Municipality Level in Mexico According to Sociodemographic Factors,” *Mathematics*, vol. 14, no. 3, pp. 1–22, 2026, doi: 10.3390/math14030573.
- [12] M. Naghavi *et al.*, “Global burden of 288 causes of death and life expectancy decomposition in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021,” *Lancet*, vol. 403, no. 10440, pp. 2100–2132, 2024, doi: 10.1016/S0140-6736(24)00367-2.
- [13] M. Maesaroh, T. N. Padilah, and J. H. Jaman, “Penerapan Algoritma K-Means Clustering Pada Pembesaran,” *J. Mhs. Tek. Inform.*, vol. 7, no. 4, pp. 2783–2786, 2023.
- [14] T. Ullah *et al.*, “Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection,” *IEEE Access*, vol. 12, no. February, pp. 16431–16446, 2024, doi: 10.1109/ACCESS.2024.3359910.
- [15] T. Yamamoto *et al.*, “Predicting mortality dynamics in cancer patients: A machine learning approach to pre-death events,” *PLoS One*, vol. 20, no. 9 September, pp. 1–15, 2025, doi: 10.1371/journal.pone.0331650.
- [16] WIDYA WIJAYANTI, IZZATI RAHMI HG, and FERRA YANUAR, “Penggunaan Metode Fuzzy C-Means Untuk Pengelompokan Provinsi Di Indonesia Berdasarkan Indikator Kesehatan Lingkungan,” *J. Mat. UNAND*, vol. 10, no. 1, pp. 129–136, 2021.
- [17] L. Gubu, E. Cahyono, Arman, H. Budiman, and M. K. Djafar, “Family of K-Means Clustering for Robust Mean-Variance Portfolio Selection: A Comparison of K-Medoids, K-Means, and Fuzzy C-Means,” *Ind. Eng. Manag. Syst.*, vol. 23, no. 3, pp. 342–356, 2024, doi: 10.7232/iems.2024.23.3.342.
- [18] S. Developers, “sklearn.preprocessing.MinMaxScaler,” [scikit-learn.org](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html). Accessed: Apr. 22, 2026. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>
- [19] IBM, “What is K-Means Clustering?,” [ibm.com](https://www.ibm.com/topics/k-means-clustering). Accessed: Apr. 22, 2026. [Online]. Available: <https://www.ibm.com/topics/k-means-clustering>
- [20] MathWorks, “Fuzzy C-Means Clustering,” [mathworks.com](https://www.mathworks.com/help/fuzzy/fuzzy-c-means-clustering.html). Accessed: Apr. 22, 2026. [Online]. Available: <https://www.mathworks.com/help/fuzzy/fuzzy-c-means-clustering.html>
- [21] S. Developers, “sklearn.metrics.silhouette_score,” [scikit-learn.org](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html). Accessed: Apr. 22, 2026. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html